Prévalences du VIH en Afrique : validité d'une mesure

Joseph Larmarange

Thèse de doctorat en démographie
Université Paris 5 René Descartes

2007

N° attribué par la bibliothèque : |__|__|__|__|__|__|__|__|__|__|

Prévalences du VIH en Afrique : validité d'une mesure

Université Paris 5 René Descartes • UFR de Sciences Sociales • Sorbonne

Thèse pour l'obtention du grade de Docteur de l'Université Paris Descartes, discipline Démographie, présentée et soutenue publiquement le 27 novembre 2007 par

Joseph LARMARANGE

sous la direction de

Benoît FERRY (IRD)

Jury

Éva LELIÈVRE (INED)

Philippe MSELLATI (IRD)

Yves CHARBIT (Université Paris Descartes)

Arnaud FONTANET (Institut Pasteur)

Nicolas MÉDA (Centre Muraz)

Prévalences du VIH en Afrique : validité d'une mesure

thèse présentée par

Joseph LARMARANGE

sous la direction de

Benoît FERRY

Paris, 2007

à Frida,
la fille du Nord des chansons de Brel

dix ans déjà
mais dans mon cœur à jamais

Remerciements

De nombreuses personnes ont permis à ce travail de voir le jour et je ne peux toutes les nommer ici. Qu'elles reçoivent néanmoins ma profonde gratitude.

En premier lieu, il me faut citer Benoît FERRY qui a accepté d'encadrer mon travail tout au long de ces trois dernières années. Mes nombreuses discussions avec lui, tant formelles qu'informelles, ont été une importante source d'enrichissement.

Je tiens à remercier Éva LELIÈVRE et Philippe MSELLATI qui ont accepté d'être rapporteurs de cette thèse ainsi qu'Arnaud FONTANET, Nicolas MÉDA et Yves CHARBIT pour leur participation au jury de soutenance.

Je n'aurais pu consacrer trois années à cette thématique sans le soutien de l'ANRS au travers du financement d'une allocation de recherche et du projet IRD/Centre Muraz « *Mesure et Ajustement des prévalences nationales du VIH en Afrique subsaharienne* » (ANRS 12114) sans lequel je n'aurais pu me rendre sur le terrain ou participer à des conférences internationales.

Avoir à travailler au sein d'une équipe pluridisciplinaire oblige à devoir clarifier en permanence les concepts et les méthodes employés et à explorer d'autres champs disciplinaires. Un grand merci donc aux différents membres du projet, à savoir Benoît FERRY et Nicolas MÉDA (coordination scientifique), Roselyne VALLO, Seydou YARO et Philippe MSELLATI.

Ce projet ne s'est pas limité aux seules équipes de l'IRD et du Centre Muraz. Il n'aurait pu avoir lieu sans le soutien des coordinateurs Nord et Sud de l'ANRS et la participation active des divers partenaires au Burkina Faso et au Cameroun, dont les Conseils Nationaux de Lutte contre le Sida et les Instituts Nationaux de la Statistique de ces deux pays.

Merci aux représentations locales de l'IRD pour leur appui technique et logistique lors de mes différentes missions.

J'ai eu l'opportunité de participer à plusieurs dynamiques d'équipes qui se sont avérées complémentaires et n'ont cessé d'alimenter, parfois indirectement, mes réflexions :

- le laboratoire POPINTER dans lequel s'est effectué ce travail qui fut le lieu de ma première socialisation à la recherche et qui m'a incité à conserver un caractère académique dans mes réflexions ;
- le Groupement d'Intérêt Scientifique CEPED qui m'a permis de rencontrer des chercheurs de différentes institutions et m'a induit à garder à l'esprit la nécessité de produire des résultats opérationnels ;
- et enfin le séminaire des doctorants de l'INED.

La relecture d'une thèse est longue et fastidieuse. Je dois beaucoup à Carole, Daniel, Marie, Nathanaël, Pascaline, Pili et surtout Yvonne pour y avoir consacré une partie de leur temps, ainsi qu'à Gabrielle pour avoir vérifier le résumé anglais.

D'une certaine manière, c'est à Michèle DION que je dois d'en être là car ce fut elle qui m'incita à approfondir la démographie lors de mes premières années universitaires en sociologie à Dijon.

Enfin, mais non des moindres, je n'aurais jamais réussi à terminer ce travail sans le soutien de ma famille, de mes amis et des mes collègues qui ont su rester patients, même après avoir subi mes interminables monologues sur l'état d'avancement de mon travail.

Sommaire

La thèse est accompagnée d'un CD-Rom.

Remarques préliminaires

Les références bibliographiques sont indiquées dans le corps de texte soit en note de bas de page soit entre parenthèses par le nom du premier auteur suivi de l'année d'édition. Pour les ouvrages ayant connu plusieurs éditions et/ou traductions, l'année de la première édition originale a été retenue, cette dernière ayant plus de sens que l'année de la réédition. S'il s'agit d'une édition revue et augmentée, c'est l'année de cette réédition corrigée qui a été utilisée.

Les numéros de page mentionnés à la suite d'une citation font quant à eux référence à l'édition qui a été consultée et dont les détails sont précisés dans la bibliographie finale de la thèse. Pour certains ouvrages, afin de faciliter la recherche du texte original entre plusieurs éditions, les citations sont suivies du numéro de chapitre et/ou de section d'où elles sont tirées.

Les citations d'ouvrages anglophones sont retranscrites dans le corps de texte en anglais (lorsque nous avons consulté l'ouvrage original). Nous proposons pour chacune une traduction en français par nos soins en note de bas de page.

Nous invitons le lecteur soucieux de plus de précisions sur les références bibliographiques à lire les notes relatives à ces dernières en page 335.

Sauf mention explicite de notre part, les passages soulignés dans les citations sont le fait de l'auteur de celle-ci. Nos aménagements sont indiqués pour leur part entre crochets.

Les notes de bas de page se continuant sur plusieurs pages sont indiquées par le symbole *(…)*.

Nous avons eu recours, pour une majorité de graphiques, à la palette de couleurs *Spectral* de la série ColorBrewer (HARROWER 2003). Ces schémas colorimétriques peuvent être consultés sur http://www.colorbrewer.org.

Avant-propos

Le projet de cette thèse est né suite à la réalisation, fin 2003, dans le cadre d'une vacation effectuée au CEPED (Centre Population et Développement), d'une synthèse bibliographique sur les estimations des prévalences nationales du VIH en Afrique subsaharienne. Les résultats des premières enquêtes nationales en population générale venaient juste d'être publiés et la surveillance sentinelle des femmes enceintes constituait la source de référence pour les estimations.

L'objectif initial de la thèse visait à élaborer une approche méthodologique d'ajustement des données de surveillance sentinelle à partir des Enquêtes Démographiques et de Santé (EDS) à l'image des techniques utilisées pour corriger la sous-fécondité des femmes enceintes infectées par le VIH (voir Annexe 1). Par ailleurs, il était envisagé de procéder à une analyse comparative entre les résultats produits par différents modèles épidémiologiques.

Plusieurs éléments ont profondément changé cette problématique telle qu'elle se posait dans le contexte scientifique international. Tout d'abord, les résultats provenant d'enquêtes nationales en population générale (notamment des EDS) sont devenus de plus en plus nombreux au fur et à mesure des cinq dernières années. Pour un nombre croissant de pays, les résultats obtenus divergeaient sensiblement des estimations réalisées jusqu'alors (voir Figure 1.8 page 52), interrogeant à la fois la validité de la surveillance sentinelle et celle de ce nouveau type de mesures.

Pour les premières EDS avec dépistage du VIH (réalisées en 2001/2002), il n'était pas possible de lier les résultats des tests avec les données des questionnaires individuels. Courant 2005, lorsque les bases de données des EDS réalisées en 2003 ont été mises en ligne, il s'est avéré qu'il était dorénavant possible de lier les résultats des questionnaires et du dépistage. En juillet 2005, lors du XXV^{e} Congrès International sur la Population, qui s'est tenu à Tours, ORC Macro, qui coordonne les différentes EDS, a annoncé avoir élaboré une procédure d'anonymisation des

données pour permettre de lier également les résultats des tests VIH aux coordonnées géographiques (longitude et latitude) des zones enquêtées. Ces dernières ont été mises en ligne à partir de la fin de l'année 2005.

Dès lors, il nous a semblé qu'il était nécessaire de tirer parti de ces informations pour mieux appréhender, à un niveau local, les différences observées entre les EDS et la surveillance sentinelle. Cependant, il a fallu pour cela développer une approche cartographique spécifique et recourir à la modélisation et simulation d'enquêtes.

Parallèlement, l'ONUSIDA a fait évoluer le logiciel EPP (Estimation and Projection Package) qu'elle utilise pour ses estimations biennales des prévalences nationales du VIH pays par pays. Il s'agit d'un modèle épidémiologique simple qui ajuste une courbe des prévalences aux données de surveillance sentinelle en cliniques prénatales. Deux nouveautés majeures ont été introduites dans la version 2.0 utilisée pour l'élaboration du rapport 2006 : d'une part, une technique d'ajustement des niveaux qui permet de tenir compte de l'extension progressive de la surveillance sentinelle dans certains pays et, d'autre part, un calibrage de la courbe estimée sur des données d'enquêtes en population générale. En résumé, EPP estime les tendances à partir des femmes enceintes et le niveau à partir d'une enquête en population générale. Cependant, la littérature internationale invite à rester prudent sur les estimations en population générale en raison de taux de personnes non testées (refus ou absence) non négligeables.

De nouvelles évolutions viennent encore d'être apportées à EPP dans sa version 2007 publiée en milieu d'année. Par ailleurs, certains pays devraient bientôt conduire une seconde enquête en population générale avec dépistage du VIH.

Dans ce contexte mouvant, plusieurs pistes de recherche envisagées se sont révélées caduques avant même d'avoir été explorées. La thèse s'est finalement orientée sur une analyse méthodologique critique des différentes sources plutôt que sur l'élaboration d'une nouvelle méthode d'ajustement. Notre objectif vise donc ici à déterminer la portée, les limites et la signification objective de la surveillance sentinelle des femmes enceintes, d'une part, et des enquêtes nationales en population générale, d'autre part, en centrant notre propos sur l'estimation des prévalences nationales du VIH en Afrique subsaharienne.

Le lecteur pourra être éventuellement dérouté par le fait que, en différents endroits de la thèse, le nombre de pays analysés peut varier. Cela vient notamment de la publication progressive de nouvelles EDS avec dépistage du VIH. Lorsque cela nous a été possible, nous avons essayé d'inclure un maximum de pays dans nos analyses. Cependant, les données de trois pays, Burkina Faso, Cameroun et Kenya, serviront de fil directeur dans les chapitres 3 à 5. Il s'agit du Burkina Faso, du Cameroun et du Kenya.

Le Burkina Faso et le Cameroun constituent les deux sites ANRS sur lesquels porte le projet IRD/Centre Muraz intitulé « *Mesure et Ajustement des prévalences nationales du VIH en Afrique subsaharienne* » (ANRS 12114) et dans lequel s'inscrit cette thèse. Pour ces deux pays, une collaboration avec les autorités

nationales a pu être mise en place. Elle m'a permis, d'une part, de rencontrer les experts nationaux en charge de la surveillance et des estimations dans leur pays et, d'autre part, d'affiner et de préciser certains résultats lors de plusieurs ateliers de travail locaux.

Le Kenya, pour sa part, a été sélectionné pour la disponibilité de ses données et sa complémentarité par rapport au Burkina Faso et au Cameroun. Outre sa positon géographique, il présente une prévalence nationale plus élevée ainsi qu'un système de surveillance sentinelle plus ancien et plus développé. Par ailleurs, c'est le seul pays, à notre connaissance, où une analyse cartographique de la prévalence du VIH a été réalisée à partir d'une EDS.

Chapitre 1
Petite histoire épidémiologique de la surveillance du VIH/SIDA

Les définitions de l'épidémiologie sont multiples. L'une des plus largement admises correspond à celle retenue lors d'un symposium de l'Organisation Mondiale de la Santé[1] (OMS) en 1968 : « *étude de la distribution des maladies et des invalidités dans les populations humaines, ainsi que des influences qui déterminent cette distribution* ». L'épidémiologie s'est historiquement axée sur la fréquence d'une maladie et plus précisément sur le risque (ou encore la probabilité) de contracter une pathologie.

Une majorité d'auteurs distingue trois volets correspondant à des élargissements successifs du champ d'action de cette discipline et définis ainsi par Daniel SCHWARTZ[2] :

> *« L'épidémiologie descriptive vise à estimer ce risque [de contracter une maladie] dans une population,*

[1] OMS EURO, *L'Enseignement de l'épidémiologie en médecine et santé publique : rapport sur un symposium*, Copenhague (DK), OMS, 1968.

[2] SCHWARTZ D., *L'explication en épidémiologie*, Chaire Quetelet 1987, Louvain-la-Neuve (BE), 13-16 octobre 1987, Ciaco, Institut de Démographie - Université Catholique de Louvain, 1989, p. 127.

> *l'épidémiologie analytique vise à déterminer les facteurs qui le gouvernent ou "facteurs de risque", à quantifier leur rôle et, si possible, à l'interpréter ;*
>
> *l'épidémiologie expérimentale vise à évaluer l'effet de mesures destinées, soit à diminuer le risque, ce qui est le cas en prévention, soit, par extension, à modifier le cours de la maladie. »*

1.1 Indicateurs de base en épidémiologie

Dans le domaine de l'épidémiologie descriptive, bien que la définition de l'état de santé comme « *état complet de bien-être physique, mental et social*[3] » ait amené les épidémiologistes à construire des indicateurs de santé complexes, les indicateurs fondamentaux de la discipline restent les *indices de mortalité et de morbidité*.

Ces indicateurs permettent de décrire la dynamique d'une épidémie dans ces différentes dimensions : flux entrants (*incidence*), flux sortants (*mortalité*), population atteinte (*prévalence*). À cela viennent s'ajouter les fluctuations liées aux migrations. Présentée ainsi, l'épidémiologie descriptive est très proche de la démographie et de sa manière de décrire les populations. Il importe néanmoins de prendre en compte un autre type d'évènements constituant également un flux sortant, à savoir la *guérison*.

L'incidence et la prévalence sont définies en ces termes par l'OMS[4] :

> *« [La prévalence correspond au] nombre des cas de maladies ou des personnes malades, ou de tout autre évènement tel qu'un accident, existant ou survenant dans une population déterminée, sans distinction entre les cas nouveaux et les cas anciens. »*
>
> *« L'incidence est le nombre des cas de maladies qui ont commencé, ou des personnes qui sont tombées malades pendant une période donnée et pour une population déterminée. »*

L'incidence et la prévalence sont usuellement exprimées sous forme de taux, le nombre de cas étant alors rapporté à la population totale à la date correspondante

3 OMS, *Constitution de l'Organisation Mondiale de la Santé*, Genève (CH), 2006, p. 1.

4 HOGARTH J., *Vocabulaire de la santé publique*, Copenhague (DK), OMS Bureau régional de l'Europe, 1977.

pour la prévalence et à la population moyenne sur la période de temps concernée pour l'incidence.

Une relation existe entre incidence et prévalence, la prévalence étant proportionnelle au produit de l'incidence et de la durée moyenne de la maladie considérée. Si la dynamique épidémique d'une pathologie est en équilibre, c'est-à-dire si l'incidence et la prévalence restent constantes dans le temps, alors nous obtenons l'équation suivante (RUMEAU-ROUQUETTE 1981, p. 169) :

Équation 1.1
Lien entre incidence et prévalence

$$p_{t+1} = i_{t,\,t+1} \times d$$

p_{t+1} : prévalence à la date t
$i_{t,t+1}$: incidence sur la période $(t, t+1)$
d : durée moyenne de la maladie

Cette équation montre notamment qu'à incidence constante, si la durée moyenne de la maladie diminue, alors la prévalence diminue proportionnellement.

Les notions d'incidence et de prévalence, indicateurs de base en épidémiologie, servent également à la définition des différents types de prévention :

> *« Prévention primaire*
>
> *Tous actes destinés à diminuer l'incidence d'une maladie dans une population en réduisant le risque d'apparition de cas nouveaux. Cette définition correspond à la définition traditionnelle de la prévention.*
>
> *Prévention secondaire*
>
> *Tous actes destinés à diminuer la prévalence d'une maladie dans une population en en réduisant l'évolution et la durée. Cette définition prend en compte certains aspects du traitement.*
>
> *Prévention tertiaire*
>
> *Tous actes destinés à diminuer la prévalence des incapacités chroniques dans une population en réduisant au minimum les invalidités fonctionnelles consécutives à la maladie. Cette définition étend le concept de la prévention au domaine de la réadaptation. »*[5]

Si la prévention primaire se situe en amont d'une éventuelle infection en visant justement à empêcher cette dernière, la prévention secondaire porte sur les

5 HOGARTH, *Vocabulaire de la santé publique*, p. 1.

personnes déjà infectées et repose en particulier sur le dépistage et le traitement précoces des infections.

Il y a une asymétrie apparente entre la définition de la prévention primaire et celle de la prévention secondaire. La prévention primaire cherche à limiter un flux, à savoir les nouvelles infections tandis que la prévention secondaire vise à réduire le nombre total de personnes infectées (stock). Cette réduction ne peut être induite par la réduction des flux entrants puisque cela correspondrait alors à de la prévention primaire. Il s'agit donc de réduire la prévalence par la diminution de la durée moyenne de l'infection (paramètre d de l'Équation 1.1), réduction obtenue implicitement par l'augmentation du nombre de personnes guéries (flux sortant). Ainsi la prévention secondaire est-elle le plus souvent assimilée à « *la fourniture d'un traitement et de soins pour les personnes infectées et malades* »[6]. Cependant, dans le cadre d'infections incurables comme l'infection à VIH, seul le décès des personnes infectées constitue un flux sortant. Dans cette situation, les traitements et actes de soins dispensés auprès des personnes infectées visent à empêcher ou au moins à retarder le décès. Ils induisent alors une augmentation de la durée moyenne de l'infection et, par là-même, une augmentation de la prévalence[7]. Ainsi, dans le cadre de ce type de pathologie, l'objectif de la prévention secondaire sera non pas la diminution de la prévalence mais celle de la mortalité. La diminution de la prévalence sera un objectif global de l'action conjointe des préventions primaires et secondaires. En effet, une baisse de la prévalence pourra être obtenue à la fois en développant des actions permettant de réduire la transmission de l'infection (prévention primaire) mais également en réduisant l'infectiosité des personnes infectées à l'aide de traitements adaptés, la prévention secondaire contribuant alors aux objectifs de la prévention primaire.

1.2 Émergence de l'épidémie de SIDA

En 1980, à Los Angeles (USA), le Docteur Michael GOTTLIEB soigne un homosexuel présentant des signes cliniques d'amaigrissement, de mycose, de fièvre, de candidose buccale et de pneumonie[8]. Ce patient présente un taux sanguin anormalement bas de lymphocytes T4. Deux autres hommes homosexuels

[6] ONUSIDA/OMS, *Les maladies sexuellement transmissibles : politiques et principes de prévention et de soins*, Genève (CH), 1997a, p. 12.

[7] Si l'incidence reste constante.

[8] Pour une vision d'ensemble des évènements présentés, voir Encadré 1.6 page 58.

présentent des symptômes proches. À Atlanta (USA), le Center for Disease prevention and Control (CDC) ouvre une enquête.[9]

En 1981, d'autres cas analogues sont décrits aux États-Unis et en Europe. À la fin de l'année, des médecins américains décident de nommer cette affection *Acquired Immune Deficiency Syndrome (AIDS)* qui sera traduite en français par *Syndrome d'Immunodéficience Acquise (SIDA)*. Ce terme apparut pour la première fois dans le *Morbidity and Mortality Weekly Report (MMWR)* du CDC en 1982[10]. Entre le 1er juin 1981 et le 15 septembre 1982, le CDC a reçu 593 déclarations de cas de SIDA. Parmi ces patients, 243 (41 %) sont décédés.

En 1983, le terme SIDA apparaît pour la première fois dans le *Relevé Épidémiologique Hebdomadaire (REH)* de l'OMS[11] et dans le *Bulletin Épidémiologique Hebdomadaire (BEH)* de la Direction Générale de la Santé (DGS) française. L'OMS publie une première série de chiffres concernant l'Europe : 153 cas ont été diagnostiqués au 30 juin 1983 dans 14 pays (OMS 1983c).

La majorité des cas décrits aux USA et en Europe concerne des homosexuels masculins (environ les trois quarts), des toxicomanes utilisant des produits injectables en intraveineuse ou des hémophiles. D'autres cas sont observés chez des patients originaires d'Haïti ou d'Afrique centrale. Une petite minorité de ces derniers appartient à des groupes dits à risques et la proportion de femmes y est relativement élevée. Si dès 1983 des cas ont été reportés à Haïti (202 cas), l'identification de cas de SIDA chez des patients originaires d'Afrique n'a pas été effectuée en Afrique elle-même et les données retrouvées en Europe chez des patients africains doivent encore être confirmées directement sur ce continent. (DGS 1983, OMS 1983a)

Le premier atelier de l'OMS sur le SIDA en Afrique a eu lieu à Bangui (République Centrafricaine) du 22 au 25 octobre 1985. Si les participants des neuf pays représentés[12] ont tous admis que le SIDA constituait un problème de santé publique en Afrique Centrale, ils se heurtaient à des difficultés pour collecter des données dans la région, notamment en raison de l'absence de diagnostique de laboratoire. Bien que la nécessité de mettre en place des systèmes de surveillance

[9] HIV-SIDA.COM, *L'Épidémie de 1980 à 2000*, 2001, page web consultée le 30 mai 2007. (http://www.hiv-sida.com/historique2.shtml)

[10] CDC, « Update on Acquired Immune Deficiency Syndrome (AIDS) - United States », *Morbidity and Mortality Weekly Report*, n°31(37), 1982.

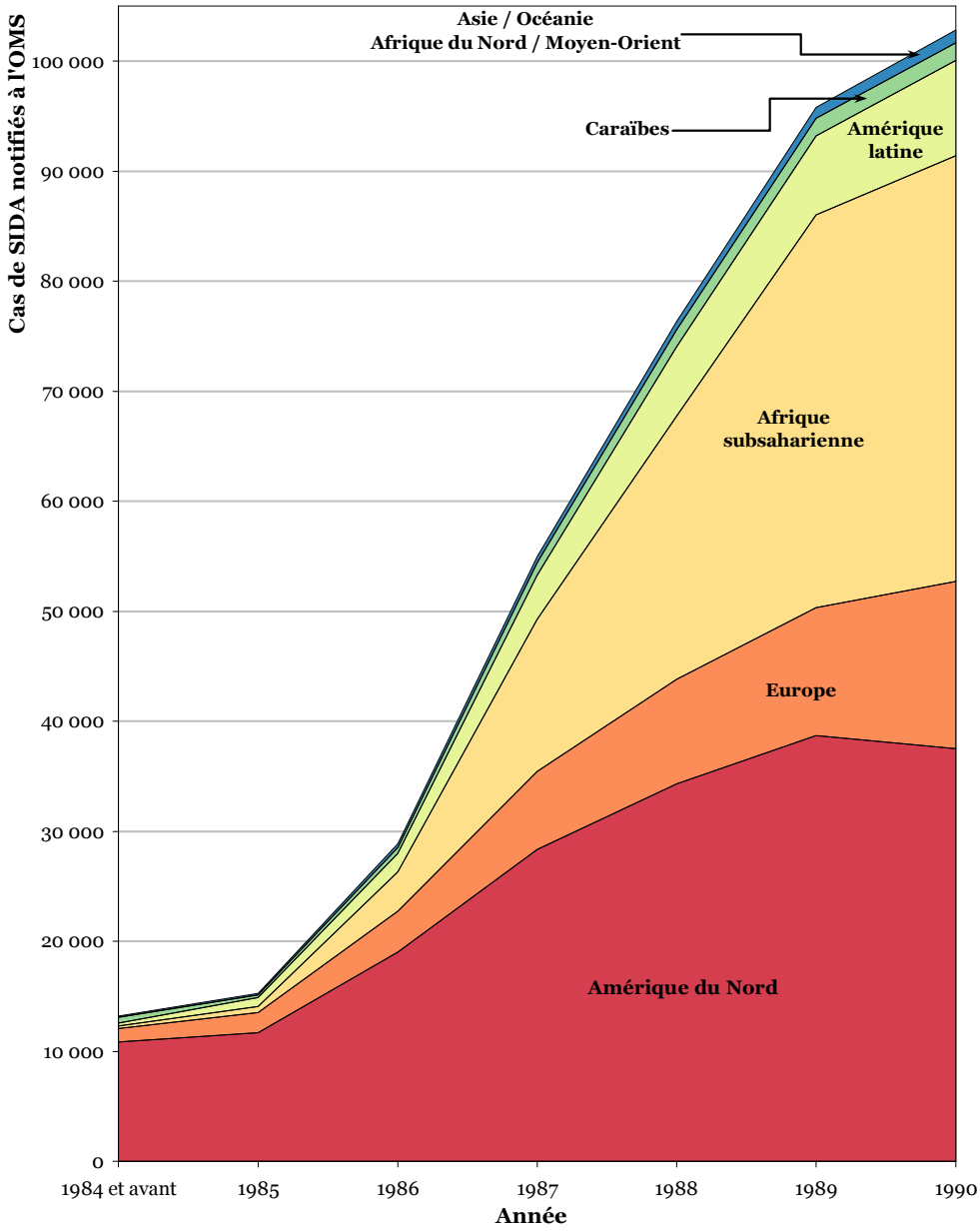
[11] OMS, « Syndrome Immunodéficitaire Acquis (SIDA) », *Relevé Épidémiologique Hebdomadaire*, n°58(14), 1983b.

[12] Burundi, Cameroun, Congo, Gabon, Ouganda, République Centrafricaine, République Unie de Tanzanie, Rwanda et Zaïre.

de l'épidémie a été affirmée, aucune donnée chiffrée n'est fournie dans le rapport de l'atelier publié dans le REH (OMS 1985).

Figure 1.1

Nombre de cas de SIDA notifiés à l'OMS au 1er janvier 1992, par année et aire géographique d'affinité



Sources : *AIDS in the World* (MANN 1992), à partir du tableau 3.4 page 116 et appendice 3.1 page 893 et suivantes. Ces tableaux sont eux-même construits à partir de données de l'OMS, de la Pan American Health Organization et du European Center for Epidemiological Monitoring of AIDS.
Pour le découpage du monde en grandes régions, voir Encadré 1.2 page 38.

Quelques mois plus tard, 41 états membres de la région africaine de l'OMS ont tenu une réunion sous-régionale à Brazzaville (Congo) du 3 au 7 mars 1986. Ils ont notamment élaboré un plan d'action pour la lutte contre le SIDA.

> « *Au 6 mars 1986, un rapport de situation sur les activités liées au SIDA avait été reçu de 21 pays de la Région. Sept pays signalaient des cas de SIDA, 8 gouvernements ont officiellement reconnu que le SIDA constituait un problème de santé publique, 16 ont diffusé des communications sur le SIDA à l'aide des médias locaux et 11 ont déjà formé des comités nationaux anti-SIDA. Cinq pays ont institué une surveillance officielle du SIDA et 11 autres ont l'intention de le faire dans un proche avenir. Neuf pays sont en mesure de procéder sur place à l'épreuve ELISA*[13] *pour le dépistage des anticorps anti-LAV/HTLV-III*[14] *et 17 pays ont dit vouloir mettre sur pied ou renforcer des installations de laboratoire pour l'épreuve ELISA*. » (OMS 1986a)

Tableau 1.1

Cas de SIDA signalés à l'OMS, par continent et date de notification/diagnostic, au 14 novembre 1986

Continent	date inconnue	1979	1980	1981	1982	1983	1984	1985	1986	Total
Afrique	-	-	-	-	3	4	9	20	1 033	*1 069*
Amériques	24	14	56	264	1 032	3 134	5 989	10 424	8 336	*29 273*
Asie	-	-	1	-	1	8	4	24	30	*68*
Europe	1	-	1	6	47	235	536	1 326	1 542	*3 694*
Océanie	-	-	-	-	1	6	44	123	170	*344*
TOTAL	*25*	*14*	*58*	*270*	*1 084*	*3 387*	*6 582*	*11 917*	*11 111*	*34 448*

Sources : *Relevé Épidémiologique Hebdomadaire* du 21 novembre 1986 (OMS 1986b).

Il faudra néanmoins attendre le 21 novembre 1986 pour que les premiers chiffres concernant le nombre de cas de SIDA en Afrique soient publiés dans le REH (OMS 1986b). Dix pays africains[15] ont alors signalé des cas de SIDA et six[16] ont signalé ne pas avoir enregistré de cas. Seuls 1 069 cas de SIDA ont été notifiés sur le continent

[13] « *L'ELISA est une technique biochimique, principalement utilisée en immunologie, mais pas uniquement, afin de détecter la présence d'un anticorps ou d'un antigène dans un échantillon.* » WIKIPEDIA, *Enzyme-linked immunosorbent assay*, page web consultée le 30 mai 2007. (http://fr.wikipedia.org/wiki/Enzyme-linked_immunosorbent_assay)

[14] LAV (Lymphadenopathy-Associated Virus) et HTLV-III (Human T-Lymphotropic Virus-III) sont deux anciens noms du VIH (Virus de l'Immunodéficience Humaine), l'agent causal du SIDA.

[15] Afrique du Sud (41 cas de SIDA), Botswana (2), République Centrafricaine (202), Ghana (7), Jamaïque (5), Kenya (101), Ouganda (29), République Unie de Tanzanie (462), Zambie (217) et Zimbabwe (6).

[16] Comores, Éthiopie, Gambie, Ghana, Maurice et Nigeria.

africain au 14 novembre 1986, soit à peine 3,1 % des 34 448 cas mondiaux enregistrés à l'OMS (voir Tableau 1.1).

Comme pour les autres continents, l'Afrique subsaharienne va connaître à cette époque une croissance exponentielle du nombre de cas de SIDA notifiés (voir Figure 1.1). À la fin de la décennie 1980, il apparaissait clairement qu'elle était particulièrement touchée.

Cependant, le nombre de cas de SIDA notifiés n'est pas adapté pour rendre compte de la situation réelle de l'épidémie. Au 1er janvier 1992, l'Afrique subsaharienne concentrait 144 522 des 484 163 cas de SIDA notifiés à l'OMS (soit 29 %) tandis que le nombre cumulé d'infections, estimé à la même date, y était de 8 772 500 sur les 12 875 450 infections mondiales (soit 68 %)[17]. Le nombre de cas de SIDA notifiés s'avère donc un mauvais indicateur de la situation épidémique, d'une part en raison de sa définition, et surtout en raison de ses problèmes d'enregistrement d'autre part.

1.3 Définition des cas de SIDA

Dans un premier temps, avant que le terme de SIDA ne soit apparu, les premiers cas décrits concernaient des hommes jeunes et homosexuels atteints de pneumonie à *pneumocystis* ou bien du sarcome de Kaposi, pathologies rares et touchant jusqu'alors des personnes de plus de soixante ans[18]. Une première définition des cas de SIDA a été publiée pour la première fois en septembre dans le MMWR du CDC[19].

> « *Cette définition comporte une liste de pathologies considérées comme indiquant un état de déficit immunocellulaire, survenant chez un patient n'ayant pas de cause connue pouvant expliquer cette immunodépression. La définition n'inclut ni l'appartenance des patients à un groupe à risque particulier, ni l'existence d'une modification des tests immunologiques portant sur les lymphocytes T. […] Aucun*

[17] D'après MANN J. M., TARANTOLA D. J. M. et NETTER T. W., *AIDS in the World*, Cambridge (US) et Londres (GB), Harvard University Press, 1992. Tableau 2.4 page 30 et tableau 3.4 page 116.

[18] CDC, « Pneumocystis pneumonia - Los Angeles », *Morbidity and Mortality Weekly Report*, n°30, 1981b.
CDC, « Kaposi's sarcoma and Pneumocystis pneumonia among homosexuel men - New York City and California », *Morbidity and Mortality Weekly Report*, n°30, 1981a.

[19] CDC, « Update on Acquired Immune Deficiency Syndrome (AIDS) - United States ».

> *marqueur ne peut être considéré [au début des années 1980] comme spécifique du SIDA.* » (DGS 1983)

Les pathologies indicatrices d'un déficit immunocellulaire sont, en premier lieu, le sarcome de Kaposi et la pneumonie à *pneumocystis carinii*, avec ou sans autres infections dites *opportunistes* (CDC 1982). Par la suite, cette définition a connu des révisions mineures concernant cette liste de pathologies indicatrices (JAFFE 1983, CDC 1984, SELIK 1984).

Dès 1982, les premières données épidémiologiques sur le SIDA suggèrent la possibilité que cette pathologie soit d'origine virale. Le 20 mai 1983, une équipe de chercheurs français menée par Luc MONTAGNIER et Françoise BARRÉ-SINOUSSI annonce dans la revue *Sciences* avoir isolé un virus à partir du sérum d'un patient atteint du SIDA (BARRÉ-SINOUSSI 1983). Ce virus, baptisé initialement LAV (Lymphadenopathy Associated Virus), va s'avérer être l'agent causal du SIDA (MONTAGNIER 1984). Un an après les Français, en mai 1984, une équipe américaine dirigée par Max GALLO isole le virus HTLV-III qui s'avérera être identique au LAV. En 1986, le nom de VIH (Virus de l'Immunodéficience Humaine) est adopté par la communauté scientifique.

Suite à l'identification du LAV/HTLV-III et au développement de tests biologiques de recherche d'anticorps spécifiques à ce virus, la définition des cas de SIDA a évolué à l'initiative du CDC en 1985[20], puis fut reprise et adoptée par l'OMS en 1987[21]. La dernière grande révision de la définition du SIDA a été appliquée au 1er janvier 1993 aux États-Unis. Outre la prise en compte de trois nouveaux critères cliniques par rapport à 1987 (la tuberculose pulmonaire, les pneumopathies bactériennes récurrentes et le cancer invasif du col), un critère biologique a été introduit : un nombre de lymphocytes CD4 inférieur à 200 par millimètre cube de sang[22].

Avec la découverte du VIH, il est apparu que certains patients pouvaient être porteurs asymptomatiques de ce virus pendant plusieurs années avant d'entrer en phase SIDA. Il en résulte que le nombre de cas de SIDA est un indicateur de l'avancée de la pathologie chez les personnes infectées mais n'est pas à même de rendre compte de l'ampleur des différentes épidémies, les personnes porteuses asymptomatiques n'étant pas comptabilisées. Ainsi, les cas de SIDA peuvent au

[20] CDC, « Revision of the Case Definition of Acquired Immunodeficiency Syndrome for National Reporting - United States », *Morbidity and Mortality Weekly Report*, n°34(25), 1985.

[21] DGS, « Définition du SIDA avéré (révision 1987) », *Bulletin Épidémiologique Hebdomadaire*, n°51/1987, 1987.

[22] Après discussion, la France a décidé de retenir la définition américaine à l'exception du critère biologique sur le nombre de CD4 qui a été rejeté dans la définition française. (DGS, « Révision de la définition du SIDA en France », *Relevé Épidémiologique Hebdomadaire*, n°11/1993, 1993.)

mieux donner une image des infections avec une dizaine d'années de retard (MANN 1992, p. 39).

Lors du meeting tenu en 1985 à Bangui en République Centrafricaine, les participants avaient conclu que, les tests de dépistage du VIH n'étant pas répandus et accessibles facilement sur le continent africain, une définition des cas de SIDA basée sur des manifestations cliniques serait plus appropriée. Les cliniciens avaient décrits les différentes manifestations du SIDA en Afrique : amaigrissement accentué, fièvre, diarrhée chronique et plusieurs manifestations cliniques moins fréquentes comme une dermatose prurigineuse ou une infection au virus varicelle-zona. Un cas de SIDA fut alors défini comme la présence de deux conditions majeures et une condition mineure, ou bien par le diagnostic d'un sarcome de Kaposi ou d'une méningite à cryptocoques (WHO 1986).

Des études sur des patients hospitalisés ont montré que si cette définition des cas de SIDA était relativement spécifique (90 % des patients séronégatifs au VIH ne remplissaient pas les critères), elle n'était que peu sensible : seuls 50 % personnes infectées par le VIH correspondaient à cette définition (DE COCK 1988). La nature clinique de cette définition prêtait à confusion quant à son objectif, à savoir la surveillance de l'épidémie et non le diagnostic clinique, et elle fut rejetée par de nombreux cliniciens en raison de son incapacité à rendre compte de certaines manifestations cliniques.

Par ailleurs, si dans les pays du Nord la notification des cas de SIDA est une statistique relativement fiable, puisqu'il est très rare qu'une personne décède sans être passée par une consultation hospitalière et sans avoir été auscultée par un médecin à un moment donné ou à un autre, il n'en est pas de même dans de nombreux pays du Sud et en particulier d'Afrique. Sur ce continent, seule une faible part des personnes en phase SIDA a pu être diagnostiquée et donc notifiée — 10 à 30 % des cas de SIDA selon l'OMS en 1992[23] et 10 % selon Peter WAY[24], à la fois en raison de systèmes de notification inadéquats mais également, au tout début de l'épidémie, à la réticence de certains pays à reconnaître l'existence de cas de SIDA sur leur territoire. Cette statistique était encore estimée à hauteur de 15 % en 1997 par l'OMS (OMS 1997).

La surveillance des cas de SIDA s'est donc révélée insuffisante pour documenter la présence du SIDA dans un pays particulier, pour fournir une épidémiologie descriptive des groupes infectés et de la distribution par âge et sexe ou bien encore pour identifier les facteurs de risques. Comme nous l'avons déjà évoqué

[23] MANN, TARANTOLA et NETTER, *AIDS in the World*, p. 38.

[24] WAY P. O., *HIV/AIDS in Sub-Saharan Africa*, présentation à la National Academy of Sciences Committee on Population and Demography, 5-6 mars, Center for International Research US Bureau of the Census, 1992, p. 1.

précédemment, la notification des cas de SIDA en Afrique ne fournissait pas une statistique adéquate pour déterminer l'ampleur et les tendances de l'épidémie.

1.4 Premières mesures de la prévalence du VIH : développement de la surveillance sentinelle

Le développement de tests de dépistage basés sur la recherche d'anticorps anti-VIH a amené les épidémiologistes à mettre en place des systèmes pour mesurer, plutôt que les cas de SIDA, la prévalence du VIH, un indicateur plus adapté pour rendre compte de la situation épidémique internationale.

Les premières mesures de prévalence du VIH en Afrique subsaharienne apparaissent au milieu de la décennie 1980 et vont croître rapidement. En 1985, on compte une dizaine de sources, 28 en 1986, 121 en 1987, 183 en 1998 (voir Figure 1.2). Dès 1987, le Center for International Research (CIR) du Bureau Américain du Recensement met en place la *HIV/AIDS Surveillance Database* grâce au financement de l'USAID (United States Agency for International Development)[25]. L'ensemble des données de prévalence du VIH présentées lors des principales conférences internationales ou publiées dans des revues scientifiques est rassemblé et compilé dans une base de données unique disponible sur internet[26]. Les données peuvent être triées par pays, sexe, groupe d'âges et population enquêtée. Sont également collectés des chiffres d'incidence du VIH, de cas de SIDA et de mortalité liée au SIDA.

Au cours des années 1980, les enquêtes de prévalence ont varié aussi bien dans leur forme que dans leur qualité et leur quantité d'un pays à l'autre. Seuls quelques pays ont conduit des enquêtes à grande échelle en population générale : Ouganda, Rwanda, et Côte d'Ivoire. D'autres pays ont procédé à des enquêtes échantillonnées en population générale sur des aires géographiques plus restreintes : la Bissau en Guinée-Bissau, la zone de Bangui en République Centrafricaine, le district de Rakai en Ouganda, les provinces du Nord-Ouest, du Sud et du Sud-Ouest au Cameroun (MANN 1992, p. 40)[27].

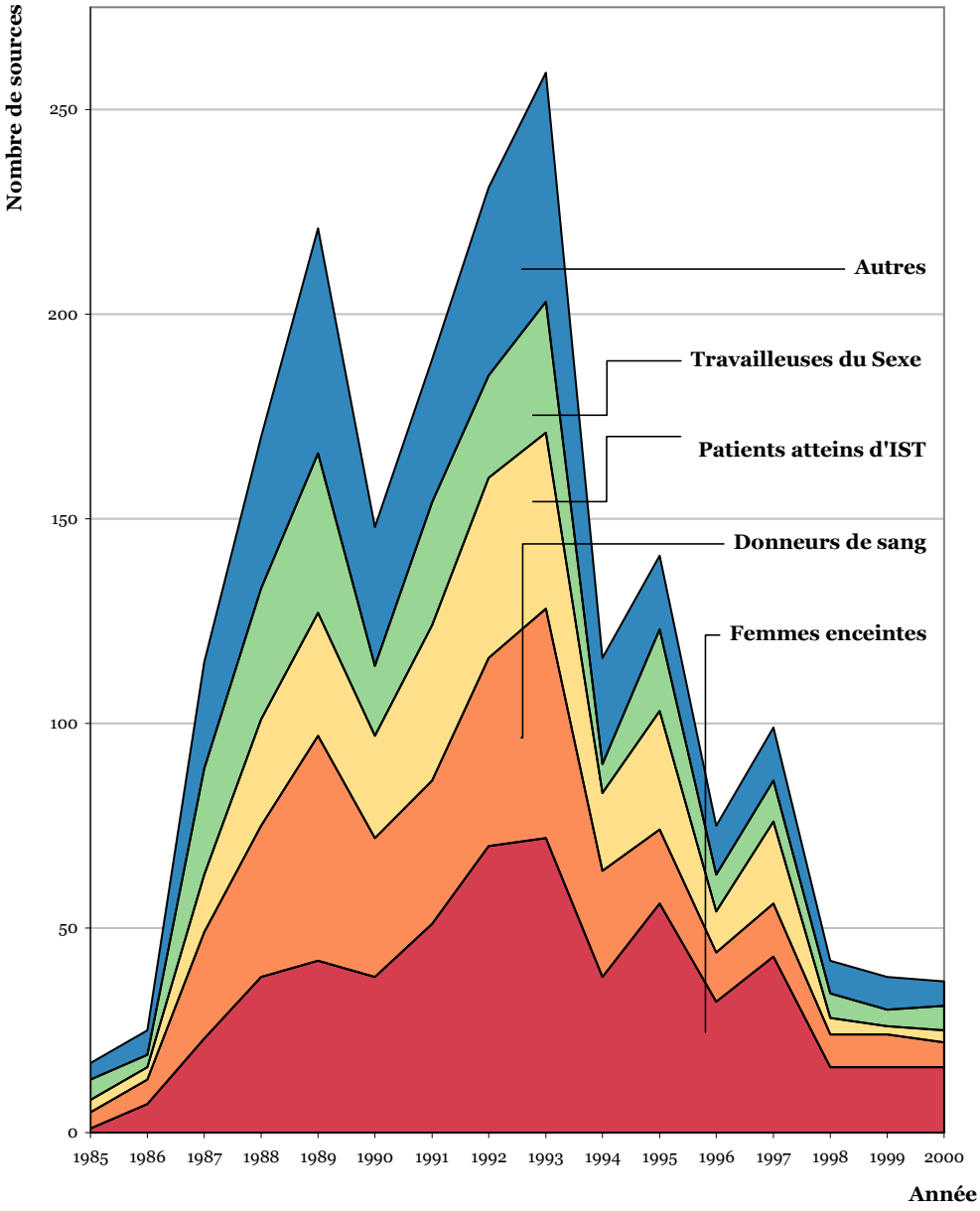
[25] Voir Encadré 1.7 page 64 pour plus de détails sur la *HIV/AIDS surveillance database*.

[26] http://www.census.gov/ipc/www/hivaidsn.html. Voir Encadré 1.7 page 64.

[27] Pour le Cameroun, DURAND J. P., MUSI S. *et. al.*, « Prévalence des porteurs d'anticorps contre les virus de l'immunodéficience humaine (VIH1 et VIH2) dans le Sud-Cameroun. Résultats des tentatives d'isolement de rétrovirus. », *Medecine Tropicale*, n°48(4), 1988.

Figure 1.2

Nombre de publications présentant des données de prévalence du VIH en Afrique subsaharienne, enregistrées dans la HIV/AIDS surveillance database, par année de publication et population enquêtée



Sources : *HIV/AIDS Surveillance Data Base – juin 2001* (US CENSUS BUREAU 2001).

Au Rwanda, l'enquête nationale menée en décembre 1986 sur 30 clusters[28] urbains et 30 clusters ruraux a montré une prévalence en milieu urbain de 14,6 % pour les hommes et de 21,0 % pour les femmes tandis qu'elle était de 1,3 % pour les hommes et 1,4 % pour les femmes en milieu rural (GODIFROID 1988, BIZIMUNGU 1989).

L'enquête nationale ougandaise a été réalisée entre septembre 1987 et janvier 1988 avec un échantillonnage visant 15 000 personnes réparties en 100 clusters. La prévalence nationale du VIH a été estimée à 4,9 % pour l'ensemble de la population (enfants inclus) avec des prévalences par sexe de 16 % pour les femmes adultes et 12 % pour les hommes adultes (BERKLEY 1989, KENGEYA-KAYONDO 1989, NAAMARA 1990).

Plusieurs enquêtes nationales ont été menées en Côte d'Ivoire entre 1986 et 1989. Lors de l'enquête de février 1989, 4 899 personnes âgées de 15 à 64 ans ont été testées pour le VIH. Les résultats montrent une prévalence comprise entre 2 et 8 % selon les régions (SANGARE 1989, BENOIT 1990, GERSHY-DAMET 1991).

Si ces différentes enquêtes nationales en population générale sont évoquées dans l'ouvrage de synthèse *AIDS in the World*[29] publié en 1992 (MANN 1992), la suite *AIDS in the World II* publiée en 1996 (MANN 1996) ne fait pas mention d'enquête nationale en population générale, même si plusieurs enquêtes en population générale sur des zones géographiques restreintes sont toujours évoquées, telles que celles menées dans le district de Rakai en Ouganda (WAWER 1991, SERWADDA 1992) ou la région de Kagera en Tanzanie (KILLEWO 1990).

Les enquêtes nationales en population générale sont des opérations lourdes, nécessitant une logistique complexe et des moyens humains et financiers importants. Une grande majorité de pays a donc mené des enquêtes sur des populations spécifiques, notamment pour identifier les groupes les plus à risques (voir Figure 1.1). Cependant, l'ensemble de ces enquêtes présente des modes opératoires très différents les uns des autres rendant les comparaisons difficiles voire impossibles. Dès 1988, l'OMS, dans le cadre du *Global Programme on AIDS (GPA)*, va formaliser les principes d'une surveillance sentinelle du VIH/SIDA.

> « *The main purpose of sentinel serosurveillance is to detect changes – i.e. to monitor trends and to provide a basis for evaluating preventive strategies and activities. However, it should be pointed out that sentinel populations need not to be "representative". At the same time, it is*

[28] Le terme de cluster renvoie aux zones géographiques sélectionnées par échantillonnage (dans le cadre d'un tirage en grappes à plusieurs degrés) et qui ont été enquêtées.

[29] Ces deux ouvrages collectifs, *AIDS in the World* et *AIDS in the World II*, préfigurent les futurs rapports mondiaux biennaux d'ONUSIDA.

> *important that sites, facilities, or populations chosen remain similar, that procedures initially chosen continue to be used (unless they are improved), and that subjects are chosen in such a way that selection is minimized. Additionally, sufficient demographic data need to be collected so that changes in the population can be detected (e.g., an influx of refugees).* » [30]

Cette surveillance sentinelle recommandée par l'OMS est censée permettre un meilleur suivi des épidémies dans un contexte où les pays disposent de relativement peu de moyens.

> « *Large-scale population serosurveys demand considerable time and resources, and their results may be of limited accuracy because of serious problems arising from selection and participation bias. Furthermore, they may become rapidly out-dated in areas where there is a high incidence of infection. WHO has therefore recommended the development of sentinel systems for routine public health surveillance of HIV infection.* » [31]

La surveillance sentinelle doit ainsi permettre un suivi annuel de populations facilement identifiables et accessibles. Comme il ne s'agit pas d'estimer l'ampleur de l'épidémie mais de suivre la manière dont cette dernière évolue au sein de différents groupes de population, la surveillance sentinelle ne nécessite pas l'exigence statistique d'être *représentative*[32]. Elle se situe résolument, à la fin des années 1980 et au début des années 1990, dans une logique de santé publique et

---

[30] SLUTKIN G., CHIN J. *et. al.*, *Sentinel surveillance for HIV infection: a method to monitor HIV infection trends in population groups*, IV International Conference on AIDS, Stockholm (SE), Juin, WHO, 1988, p. 3. Traduction par nos soins :
« L'objectif principal d'une séro-surveillance sentinelle consiste à détecter les changements, c'est-à-dire de suivre les tendances et de fournir une base pour l'évaluation des activités et des stratégies de prévention. Cependant, il doit être noté que les populations sentinelles ne nécessitent pas d'être "représentatives". Dans le même temps, il importe que les sites, les infrastructures ou les populations choisies demeurent similaires, que les procédures retenues initialement continuent d'être employées (à moins qu'elles ne soient améliorées), et que les sujets soient choisis de manière à minimiser les biais de sélection. De plus, des données démographiques suffisantes devront être collectées afin de détecter les changements survenus dans la population (par exemple, un flux de réfugiés). »

[31] CHIN J., « Public health surveillance of AIDS and HIV infections », *Bulletin of the World Health Organization*, n°68(5), 1990, p. 535. Traduction par nos soins :
« Les enquêtes de séroprévalence en population générale à large échelle requièrent des ressources considérables et beaucoup de temps, et leurs résultats peuvent être d'une précision limitée en raison de sérieux problèmes dus à des biais de sélection et de participation. En outre, elles peuvent être rapidement périmées dans les zones présentant une forte incidence de l'infection. L'OMS a donc recommandé le développement de systèmes sentinelles en vue d'une surveillance de routine de l'infection à VIH dans une optique de santé publique. »

[32] La notion de *représentativité* sera discutée plus loin dans la thèse : voir section 3.2 page 136.

une optique d'évaluation des activités de prévention et d'orientation des programmes de lutte contre le VIH/SIDA. L'hypothèse sous-jacente repose sur le fait que si la collecte des données est uniforme à travers le temps, ainsi que les biais de sélection, alors les tendances observées de l'épidémie dans les différentes populations enquêtées reflèteront l'évolution de la prévalence réelle de ces différents groupes. Le suivi sentinelle doit se concentrer en priorité sur les populations les plus à risques d'être infectées, populations alors prioritaires dans la mise en place de programmes d'actions. *« The sentinel populations selected should allow for the monitoring of major HIV risk behaviours or factors known to be prevalent in any given area.*[33] » Néanmoins, la surveillance de populations à faible risque est également suggérée, en particulier pour repérer le plus tôt possible une éventuelle montée de l'épidémie parmi celles-ci. *« Both high- and low-risk groups can be monitored to afford an indication of the range of HIV infection. Serosurveillance of high-risk groups is especially useful for targeting and should lead to suggesting acceptable intervention strategies.*[34] »

L'OMS va insister à cette époque sur la nécessité de distinguer deux objectifs distincts des tests de dépistage des anticorps anti-VIH : d'une part, la recherche de cas et, d'autre part, la surveillance de santé publique. La recherche de cas a pour objectif principal de déterminer avec certitude le statut sérologique d'un individu en vue de pouvoir lui proposer un suivi adéquat et un traitement médical approprié. Pour la surveillance en santé publique, il importe de pouvoir déterminer la prévalence, la distribution et les tendances de l'infection à VIH au sein d'un groupe ou d'une population. Pour cela, il n'est pas nécessaire de connaître l'identité des personnes enquêtées.

Il a été montré assez tôt que poursuivre simultanément ces deux objectifs produisait des données erronées concernant la prévalence réelle de l'infection du fait que les personnes refusant le test de dépistage étaient plus souvent infectées par le VIH (HULL 1988, JENUM 1988). L'étude de HULL *et. al.* a mesuré une prévalence du VIH de 1,0 % parmi 782 patients atteints d'IST ayant accepté le test de dépistage du VIH, tandis que la prévalence était de 5,4 % parmi les 167 patients

[33] CHIN, « Public health surveillance of AIDS and HIV infections », p. 534. Traduction par nos soins : *« Les populations sentinelles sélectionnées doivent permettre le suivi des comportements ou facteurs à haut risque pour le VIH connus pour être répandus dans une région donnée.*

[34] SLUTKIN G., CHIN J. *et. al.*, *Use of HIV surveillance data in national AIDS control programmes: a review of current data use with recommendations for strengthening future use*, Genève (CH), GPA/WHO, 1990, p. 5. Traduction par nos soins :
« Il est possible de surveiller à la fois des groupes à faibles et à hauts risques pour fournir une indication sur l'étendue de l'infection à VIH. La séro-surveillance des groupes à haut risque est notamment utile pour les cibler et devrait permettre de suggérer des stratégies d'intervention convenables. »

ayant refusé le test[35]. L'étude de JENUM, en Norvège, a détecté 4 femmes séropositives au VIH parmi les 36 053 femmes enceintes ayant accepté de se faire tester (soit une prévalence de 0,011 %) et une femme séropositive parmi les 50 femmes qui avaient refusé le test (soit une prévalence de 2 %).

Pour réduire les biais liés au refus de se faire tester, l'OMS a donc préconisé, pour la surveillance sentinelle, la méthode UAS (*Unliked Anonymous Screening*) qui consiste à tester des échantillons sanguins, prélevés pour d'autres raisons que la surveillance du VIH, après suppression de toute donnée nominative et tout identificateur possible (GPA 1989). Les études de surveillance sentinelle vont se développer ainsi auprès de populations qui peuvent être facilement recrutées et pour lesquelles on dispose de prélèvements sanguins[36] : patients atteints d'IST (recrutés via les services IST des services de santé), les femmes enceintes (via les cliniques prénatales), les donneurs de sang (via les agences de collecte du sang), etc.

Les femmes enceintes vont se révéler être une population particulièrement intéressante à enquêter. D'une part, dans une majorité de pays, les femmes enceintes sont suivies médicalement pour les soins prénataux, au moins en milieu urbain. D'autre part, lors du suivi prénatal, des échantillons sanguins sont collectés pour différents tests. Il est donc possible de les utiliser à des fins de surveillance épidémiologique, après anonymisation des échantillons. Elles constituent une population bien définie permettant l'utilisation d'un critère simple d'inclusion dans l'étude[37]. Enfin, elles sont considérées comme plus ou moins représentatives de la population générale. « *For many countries, data on pregnant women provide the most representative picture of HIV infection in the general population. This is particularly true when testing and pre/post counselling for the purpose of individual diagnosis being offered separately.*[38] »

[35] Les tests du VIH ont été réalisés sur des prélèvements de sang collectés pour le test de la syphilis, après suppression de l'ensemble des données individuelles nominatives.

[36] Voir Figure 1.2 page 26.

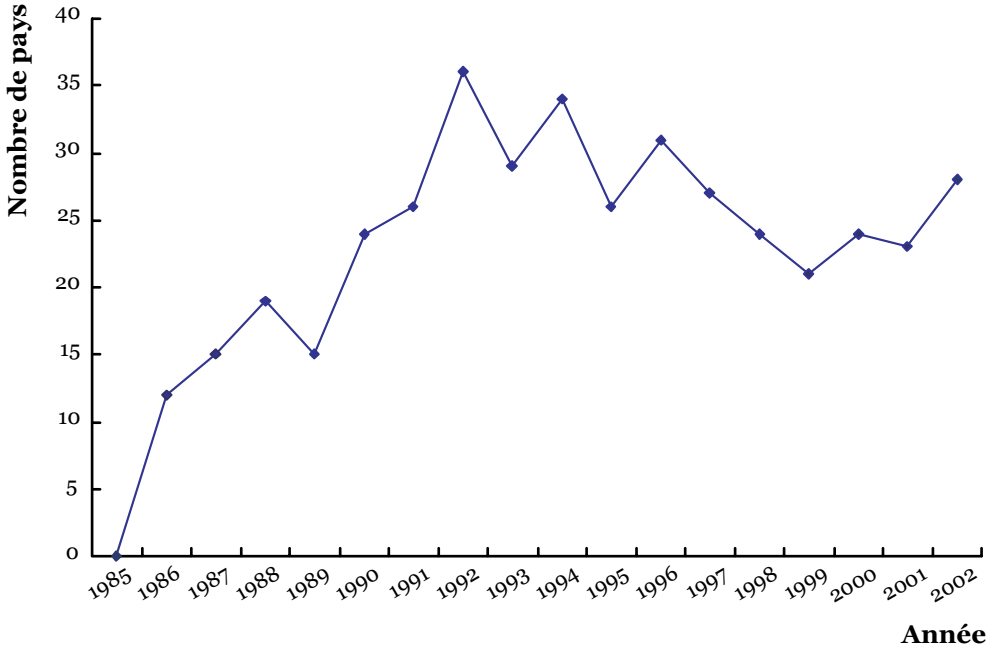
[37] Le plus souvent est inclue toute femme enceinte lors de sa première consultation en clinique prénatale.

[38] MANN, TARANTOLA et NETTER, *AIDS in the World*, p. 64. Traduction par nos soins :
« Pour de nombreux pays, les données sur les femmes enceintes fournissent l'image de l'infection à VIH la plus représentative de ce qu'il en est dans la population générale. Cela est particulièrement vrai quand les tests sont conduits de manière à ce que les résultats individuels ne puissent être reliés à la personne testée (méthode anonymat non lié), avec un test volontaire et un conseil pré- et post-test en vue d'un diagnostique individuel offerts séparément. »

De ce fait, une majorité de pays d'Afrique subsaharienne vont mettre en place un système de surveillance sentinelle incluant des enquêtes de routine auprès des femmes enceintes (voir Figure 1.3). En 2003, sur les 43 pays de la zone[39], 39 disposaient d'un tel système (UNAIDS/WHO 2003). Cependant, le nombre de pays d'Afrique subsaharienne ayant réalisé une enquête de surveillance sentinelle auprès des femmes enceintes varie dans le temps (Figure 1.3), notamment parce que certains pays ne procèdent pas à une enquête chaque année. Après un pic en 1992 où 37 pays ont réalisé une étude de surveillance sérologique des femmes enceintes, ce nombre a décru jusqu'en 1999.

Figure 1.3

Nombre annuel de pays d'Afrique subsaharienne ayant réalisé une enquête de surveillance sérologique parmi les femmes enceintes (1985-2002)



Sources : (GARCIA-CALLEJA 2004).

WALKER *et al.* ont analysé la qualité de la surveillance sentinelle, sur la période 1990-1999, en élaborant un score permettant de prendre en compte plusieurs aspects de la surveillance[40] :

- fréquence des collectes de données ;
- ancienneté des dernières données collectées ;
- pertinence des populations sous surveillance ;

[39] Pour plus de détails sur le découpage du monde en grandes régions, voir Encadré 1.2 page 38.

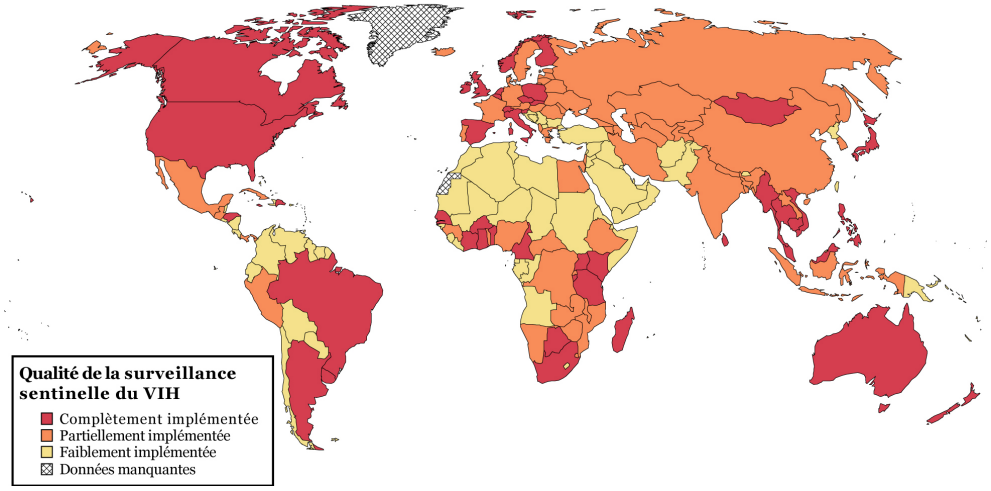
[40] WALKER N., GARCIA-CALLEJA J. M. *et. al.*, « Epidemiological analysis of the quality of HIV sero-surveillance in the world: how well do we track the epidemic? », *AIDS*, n°15(12), 2001.

- régularité des sites/populations enquêtés au cours du temps ;
- couverture et représentativité des populations enquêtées vis-à-vis de la population générale adulte.

Dans le cas des pays à épidémie généralisée[41], ce qui est le cas de presque tous les pays d'Afrique subsaharienne, une importance particulière a été accordée à la couverture de la population adulte, notamment via la surveillance des femmes enceintes. Les auteurs catégorisent les différents pays du monde en trois groupes :

- surveillance sérologique entièrement implémentée ;
- surveillance partiellement implémentée et
- surveillance faiblement implémentée.

Figure 1.4
Qualité des systèmes de surveillance sentinelle des épidémies à VIH à travers le monde en 1999



Sources : (WALKER 2001).

La Figure 1.4 montre clairement qu'en Afrique subsaharienne, la qualité de la surveillance sentinelle diverge d'un pays à l'autre. Sur 42 pays[42], 16 n'ont pas les éléments de base pour un suivi sentinelle efficace de l'épidémie, 14 présentent un système de surveillance partiellement implémenté et seuls 12 ont un système de surveillance complet.

[41] C'est-à-dire présentant une prévalence nationale supérieure à 1 % parmi les adultes. Pour plus de détails, voir Encadré 1.1 page 33.

[42] Les données sont manquantes pour un des pays de la zone.

Encadré 1.1
Les différents profils épidémiques

Pattern I, II et III

À la fin des années 1980, les épidémiologistes ont distingué trois profils épidémiques principaux (CHIN 1988) :

- Le *pattern I* correspond essentiellement à la situation observée en Amérique du Nord et en Europe de l'Ouest. L'épidémie est essentiellement masculine et urbaine, concentrée chez les homosexuels masculins et les utilisateurs de drogues.
- Le *pattern II* est typique en Afrique, dans les Caraïbes et dans une partie de l'Amérique du Sud. Le nombre d'hommes et de femmes infectés par le VIH est relativement le même. La transmission du VIH est essentiellement hétérosexuelle et, dans une moindre mesure, périnatale.
- Le *pattern III* se rencontre en Europe de l'Est, en Asie et dans certains pays d'Amérique du Sud. L'épidémie a démarré plus tardivement et touche à la fois les homosexuels masculins et les usagers de drogues par voie intraveineuse, mais également la population hétérosexuelle adulte et les enfants nés de mère séropositive. Le nombre d'hommes est supérieur à celui des femmes.

Au début des années 1990, ces trois patterns se sont révélés inadaptés pour rendre compte des évolutions des épidémies de VIH au fur et à mesure que les connaissances épidémiologiques progressaient (MANN 1992). Ils ont alors été remplacés par les Aires Géographiques d'Affinités, GAA en anglais (voir Encadré 1.2 page 38)

Épidémies généralisées, concentrées et limitées

Une nouvelle terminologie va être proposée à partir de 2000 par le groupe d'experts d'ONUSIDA (UNAIDS/WHO WORKING GROUP ON GLOBAL HIV/AIDS AND STI SURVEILLANCE 2000, p. 24) :

- les pays à *épidémie généralisée* présente une prévalence nationale du VIH supérieure à 1 % chez les adultes ;
- les pays à *épidémie concentrée* connaissent une prévalence nationale en population générale inférieure à 1 % mais certains groupes à risques présentent une prévalence supérieure à 5 % ;
- les pays à *épidémie limitée* ont une prévalence nationale faible (inférieure à 1 %) et la prévalence du VIH dans les groupes à risques n'excède pas les 5 %.

Cette dernière typologie est, depuis, largement employée dans les différentes publications de l'ONUSIDA mais n'apparaîtra dans les rapports mondiaux qu'à partir de 2004.

À partir de 2006/2007, ONUSIDA a proposé une quatrième catégorie : les *épidémies hyper-endémiques* pour rendre compte des pays d'Afrique australe où l'épidémie a atteint des prévalences élevées dans la population générale, supérieures à 15 % (UNAIDS 2007b). Le Groupe de Référence de l'ONUSIDA reste quant à lui réservé sur cette nouvelle classification, recommandant une révision plus importante de la classification des épidémies qui prendrait également en compte les populations à risques contribuant au développement des épidémies (THE UNAIDS REFERENCE GROUP ON ESTIMATES MODELLING AND PROJECTIONS 2006a).

1.5 Modéliser l'épidémie…

À la fin des années 1980, plusieurs modèles mathématiques vont être développés afin de mieux appréhender la dynamique épidémique et estimer l'impact du SIDA dans différents domaines et en particulier les conséquences démographiques. Les Nations unies et l'OMS vont organiser sur cette thématique un atelier de travail intitulé *Modelling the Demographic Impact of the AIDS Epidemic in Pattern II Countries*[43]*: Progress to Date and Policies for the Future*, du 13 au 15 décembre 1989, à New-York. Cet atelier donnera lieu à la publication d'un rapport (UN/WHO 1991).

Alberto PALLONI a présenté une synthèse des différents modèles existants alors. Il distingue quatre types de modèles (PALLONI 1991). Le premier type, relativement simple, se concentre sur l'évolution du nombre de cas de SIDA et cherche à ajuster une courbe mathématique aux données d'observations. Dans le second type, une distinction est réalisée entre les cas de SIDA reportés et les infections à VIH asymptomatiques. Ces modèles nécessitent donc l'estimation d'une fonction d'incubation traduisant l'évolution de l'infection à VIH au SIDA et procèdent à un rétro-calcul à partir des cas de SIDA notifiés pour estimer l'évolution passée de l'infection à VIH. Les modèles de type 3 tiennent compte, quant à eux, des différents modes de transmission et nécessitent donc d'estimer des probabilités de transmission. Le type 4, enfin, incorpore des hypothèses comportementales explicites. Bien que plus généraux, ces modèles nécessitent un nombre plus élevés de paramètres et sont donc soumis à des marges d'incertitudes importantes. Lors de l'atelier de travail, les résultats des principaux modèles de l'époque ont été comparés, en particulier ceux développés par Roy ANDERSON, Nicolas BROUARD, Bertran AUVERT, Rodolfo BULATAO, Klaus DIETZ et Alberto PALLONI, et le modèle IWG développé par le *United States Interagency Working Group*[44].

Ces différents modèles sont complexes à mettre en œuvre et nécessitent de nombreuses données et hypothèses en entrée. L'atelier avait conclu à la nécessité de disposer de modèles plus simples qui puissent être mis en œuvre facilement par les programmes de lutte contre le SIDA afin de fournir une aide à la décision politique. Aucun des modèles précédents n'est encore en usage de nos jours.

[43] Voir Encadré 1.1 page 33 pour plus de détails sur les différents patterns.

[44] Parmi les modèles développés à l'époque, on pourra citer le modèle de John BONGAARTS non présentés lors du workshop des Nations unies et de l'OMS. BONGAARTS J., *Modeling the Spread of HIV and the Demographic Impact of AIDS in Africa*, New York City, New York (US), The Population Council, coll. *Working Papers of the Center for Policy Studies*, 1988.

Une dernière approche a été présentée lors de cet atelier des Nations unies, celle de l'OMS (CHIN 1991b), qui retiendra plus particulièrement notre attention. Ce modèle, développé à partir de 1987 dans le cadre du Programme Global sur le SIDA, ne repose pas, comme les précédents, sur le nombre de cas de SIDA déclarés, mais sur des estimations de prévalence du VIH, notamment celles fournies par la surveillance sentinelle. Il s'agit d'un modèle relativement simple, construit de manière pragmatique afin de fournir aisément des estimations de la prévalence du VIH pour les pays présentant une épidémie élevée.

Le choix de travailler sur des données de prévalence du VIH plutôt que sur des cas de SIDA notifiés vient du fait que, dans une majorité de pays en développement et en particulier en Afrique, seule une très faible part des cas de SIDA est reportée. Les données de prévalence du VIH s'avèrent donc être de meilleure qualité.

Ce modèle a été décrit pour la première fois lors de la *III^e Conférence Internationale sur le SIDA et les cancers associés en Afrique* qui s'est tenue du 14 au 16 septembre 1988, à Arusha en République Unie de Tanzanie (LWANGA 1988). En janvier 1989, James CHIN et Jonathan MANN publiaient dans le *Bulletin de l'OMS* les premières projections à court terme du nombre de cas de SIDA dans le monde, en Europe, aux États-Unis et en Afrique (CHIN 1989).

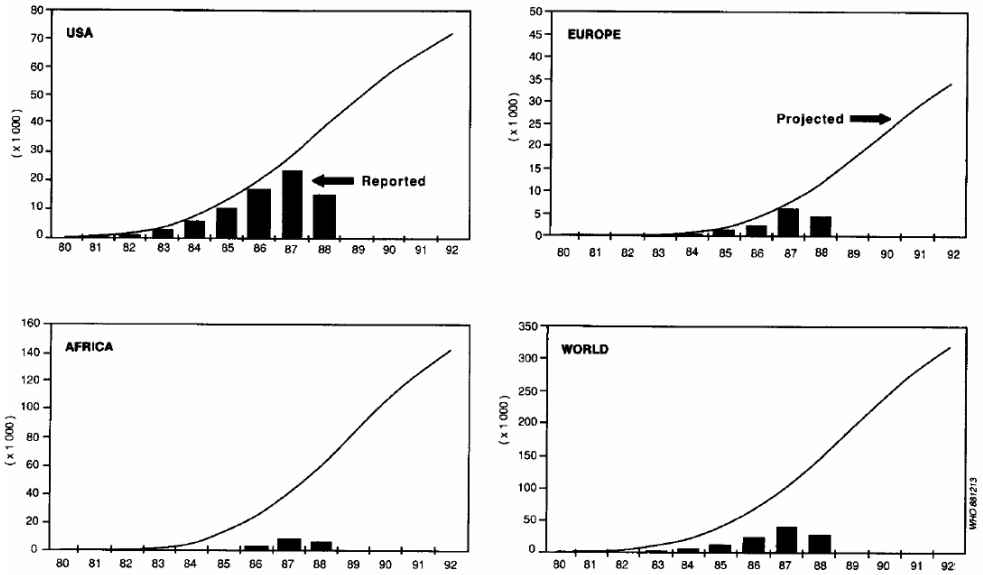
Si pour l'Europe et les États-Unis, il leur était possible de projeter le nombre de cas de SIDA à partir des tendances observées selon les notifications envoyées à l'OMS, cela n'avait pas de sens pour les données africaines (voir Figure 1.5). Ils ont donc eu recours à leur modèle épidémiologique pour estimer la courbe des prévalences du VIH à partir des données dont ils disposaient en faisant les hypothèses suivantes :

- les épidémies de VIH ont débuté à la fin des années 1970 ou au tout début des années 1980 ;
- la courbe du nombre cumulé d'infections à VIH peut être estimé à partir des données des enquêtes de séroprévalence du VIH (cette courbe est supposée, sur le long terme, atteindre une asymptote ou un plateau plutôt que de continuer à croître exponentiellement) ;
- les incidences annuelles du VIH sont calculées par décomposition de la prévalence cumulée ;
- le nombre de nouveaux cas de SIDA est estimé annuellement à partir d'une fonction d'évolution de l'infection à VIH vers le SIDA tenant compte de la durée d'infection.

Le nombre d'enfants infectés est calculé séparément, sous l'hypothèse que 25 % des enfants nés de mère infectée sont eux-mêmes infectés par le VIH. La courbe d'évolution vers le SIDA est beaucoup plus rapide et tous les enfants nés infectés décèderont avant d'atteindre l'âge adulte.

Figure 1.5

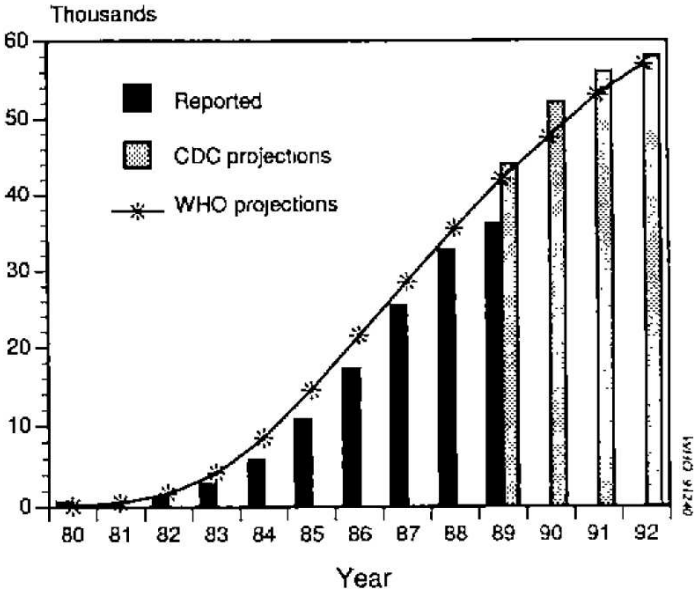
Cas de SIDA reportés et estimés (selon le modèle de l'OMS) aux États-Unis, en Europe, en Afrique et dans le monde, de 1980 à 1992 (projections pour 1988-1992)



Sources : (CHIN 1989). Les écarts entre les cas de SIDA reportés et la courbe représentent la sous-notification des cas de SIDA.

Figure 1.6

Cas de SIDA reportés (1980-1989) et estimés selon les projections de l'OMS et du CDC (1980-1992) aux États-Unis



Sources : (CHIN 1991a).

Par la suite, il a été montré que les projections réalisées aux États-Unis concordaient avec la limite inférieure des projections des cas de SIDA calculées par le CDC (voir Figure 1.6).

1.6 … pour estimer les prévalences du VIH

Fin 1991, le Programme Global sur le SIDA annonce la mise en ligne d'*EpiModel*, un logiciel simple permettant d'appliquer le modèle OMS à ses propres données et de faire varier les différents paramètres (CHIN 1991a). Le logiciel ne nécessite *a minima* qu'une estimation de la prévalence du VIH à une date donnée et une estimation de l'année où l'épidémie a démarré (usuellement 1980 ± 2 années). L'objectif premier du modèle consiste à projeter à court terme (3 ou 4 ans maximum) la courbe des cas de SIDA. Cependant, une fonction d'évolution du SIDA vers le décès a été implémentée et il est possible d'estimer ainsi le nombre de décès liés au VIH/SIDA.

En 1992, paraît *AIDS in the World*, un ouvrage collectif préfigurant les futurs rapports mondiaux d'ONUSIDA. Des estimations de prévalence et d'incidence du VIH sont alors publiées par aire géographique d'affinité (GAA, voir Encadré 1.2). Elles reposent sur le logiciel *EpiModel* et sur la méthode *Delphi*.

Il s'agit d'enquêtes d'opinion auprès d'un panel d'experts. Dans un premier temps, il leur est demandé différentes estimations chiffrées puis, dans un second temps, l'ensemble des résultats du premier tour est envoyé à chaque expert pour qu'il puisse fournir une seconde estimation. L'objectif de ce second tour consiste à faire émerger un consensus. Ce type d'enquêtes d'opinion vise à compenser l'absence de données disponibles dans certaines régions.

Fin 1995, le Programme Global sur le SIDA de l'OMS publie dans le *Relevé Épidémiologique Hebdomadaire* les premières estimations mondiales du nombre d'infections à VIH par pays à fin 1994[45]. Ces estimations reposent en particulier sur *EpiModel* pour les pays d'Afrique subsaharienne. Le rapport *AIDS in the World II*, sorti en 1996, reprend les estimations pays par pays à fin 1994 et fournit des estimations détaillées par aire géographique d'affinité à la même date (MANN 1996). Ces dernières ont été également calculées à partir d'*EpiModel*, selon la méthodologie utilisée en 1992. Cependant, les différentes hypothèses de fécondité, de ratio hommes/femmes ou urbain/rural ont été revues et corrigées, en fonction des dernières publications disponibles à l'époque et adaptées à chaque région.

[45] OMS, « Estimations de travail provisoires de la prévalence du VIH chez les adultes, à la fin 1994, par pays », *Relevé Épidémiologique Hebdomadaire*, n°70(50), 1995.

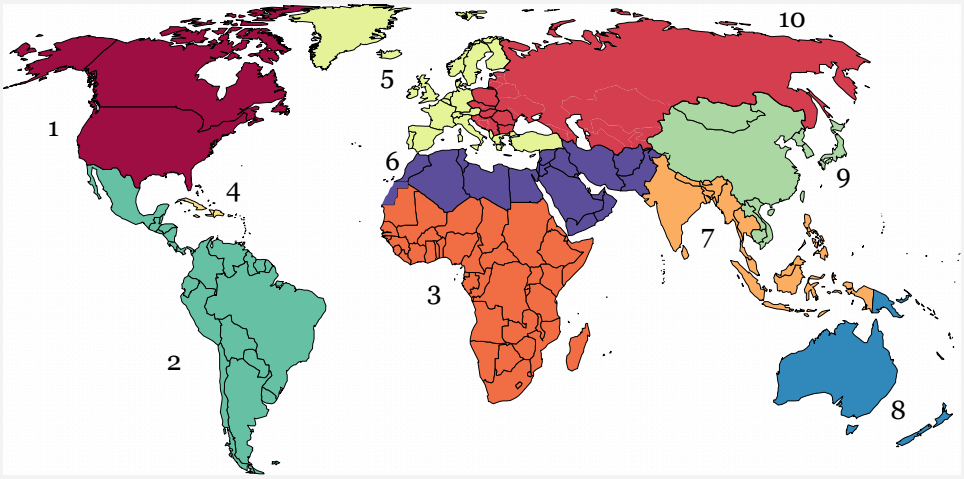
Outre une estimation de la prévalence du VIH à fin 1994 par pays, le rapport détaille, pour les grandes régions mondiales, les courbes de prévalence et d'incidence de l'infection à VIH et du nombre de cas de SIDA, ainsi que des projections jusqu'en 2001.

Encadré 1.2

Le découpage du monde en grandes régions

Les Aires Géographiques d'Affinité (Geographic Areas of Affinity – GAA)

En 1992, le rapport *AIDS in the World* (MANN 1992) propose un découpage du monde en dix grandes régions à partir de critères épidémiologiques, programmatiques et sociétaux :



1 – Amérique du Nord
2 – Amérique latine
3 – Afrique subsaharienne
4 – Caraïbes
5 – Europe de l'Ouest
6 – Sud-Est méditerranéen
7 – Sud-Est asiatique
8 – Océanie
9 – Nord-Est asiatique
10 – Europe de l'Est

Ce découpage fut ensuite repris dans le rapport *AIDS in the World II* (MANN 1996).

Les différents rapports biennaux publiés par ONUSIDA depuis 1998 présentent également leurs estimations selon un découpage proche, bien qu'il ne soit plus fait explicitement mention des aires géographiques d'affinité et que les noms des différentes régions aient été légèrement remaniés, le 'Sud-Est méditerranéen' ayant été rebaptisé 'Afrique du Nord et Moyen Orient' par exemple.

Le 1er décembre 1995 est créé le Programme Commun des Nations unies contre le VIH/SIDA plus connu sous le nom d'ONUSIDA. C'est lui qui sera chargé dorénavant du suivi mondial de l'épidémie. En novembre 1996, suite au symposium satellite sur les tendances de l'épidémie organisé lors de la XIe Conférence Mondiale sur le SIDA qui eut lieu en juillet 1996 à Vancouver, l'ONUSIDA, conjointement avec l'OMS, met en place un groupe de travail mondial sur la surveillance du VIH/SIDA et des IST (UNAIDS 2007a).

En juin 1998, ONUSIDA sort son premier rapport mondial sur l'épidémie de VIH/SIDA (UNAIDS/WHO 1998). Les estimations pays par pays de 1994 sont alors réactualisées à fin 1997, toujours en utilisant *EpiModel*. Les courbes de prévalence du VIH sont toujours estimées en tenant compte à la fois de l'année supposée de départ de l'épidémie dans le pays et de points de mesure de la prévalence du VIH.

Pour les pays où l'épidémie est concentrée dans des groupes de populations spécifiques, une courbe de prévalence est calculée, pour chaque groupe à risques, et la taille de chaque groupe est estimée, afin de calculer une prévalence nationale.

Dans les pays où l'épidémie s'est généralisée au sein de la population générale, les données de surveillance auprès des femmes enceintes suivies en clinique prénatale ont été utilisées pour estimer la prévalence de l'ensemble des femmes[46], puis la prévalence des hommes a été calculée selon une hypothèse concernant le ratio hommes/femmes. Ce ratio a été estimé selon la littérature disponible à l'époque et fut considéré comme valant 1:1 en Afrique subsaharienne. Les estimations portant sur les femmes enceintes ont été réalisées en distinguant deux groupes de cliniques prénatales : les cliniques situées dans les principales zones urbaines (*major urban areas*) et celles situées en-dehors (*outside major urban areas*). Cette distinction résulte de la quasi-absence de données en milieu rural pour une majorité de pays. Pour les pays disposant de plusieurs cliniques prénatales enquêtées, la valeur médiane des prévalences observées a été utilisée.

Enfin, des ajustements divers ont pu être appliqués selon les pays. Les valeurs de ces ajustements ont été obtenues selon la méthode *Delphi* qui consiste à demander à différents experts, de l'OMS, de l'ONUSIDA, des institutions de recherche et des programmes nationaux de chaque pays, d'ajuster les résultats obtenus en fonction de leur connaissance du pays et de leur jugement sur la représentativité des sites sentinelles enquêtés (SCHWARTLANDER 1999).

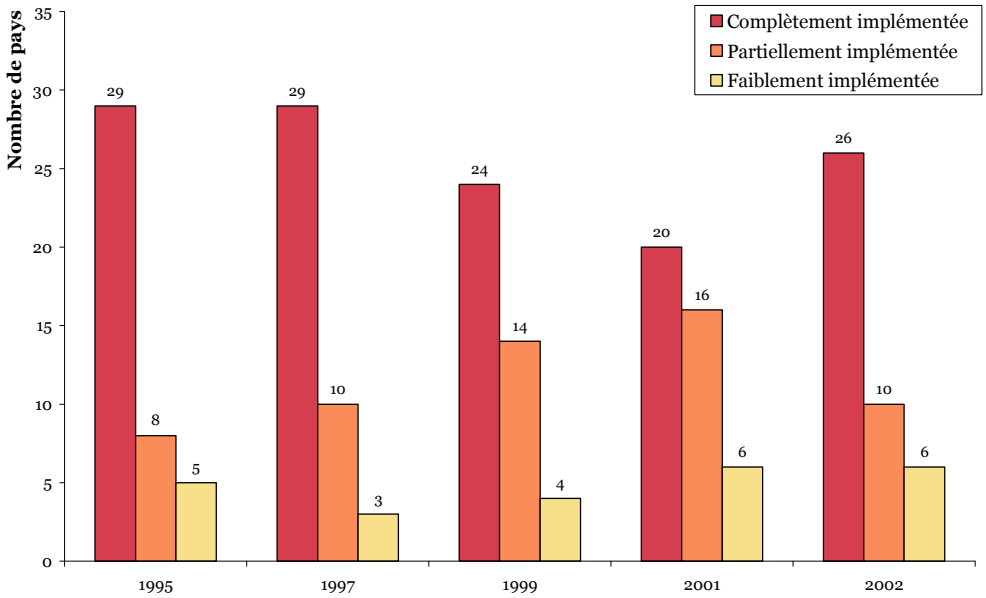
En 1999, ONUSIDA créé un *Groupe de Référence en Épidémiologie* qui deviendra en 2002 le *Groupe de Référence d'ONUSIDA sur les Estimations, la Modélisation et les Projections*. L'objectif de ce groupe consiste à faire évoluer les outils utilisés pour les estimations des tendances de l'épidémie pays par pays. Cependant, les travaux de ce groupe ne seront pas encore suffisamment avancés pour le second rapport mondial sorti en juin 2000 (UNAIDS 2000). Les estimations pays par pays ont donc été actualisées à fin 1999 selon la même méthodologie que le précédent rapport. Néanmoins, ce rapport propose pour la première fois des estimations hautes et basses du nombre d'individus infectés par le VIH ainsi que du nombre de décès dus au SIDA durant l'année 1999. Enfin, sont publiées les mesures de

[46] Les femmes enceintes suivies en clinique prénatale ont donc été considérées comme représentatives de l'ensemble des femmes en population générale.

prévalence du VIH effectuées auprès des jeunes adultes qui sont considérés comme un proxy de l'incidence du VIH.

À la même période, l'ONUSIDA décide de lancer une initiative pour une surveillance sentinelle de seconde génération (UNAIDS/WHO WORKING GROUP ON GLOBAL HIV/AIDS AND STI SURVEILLANCE 2000). En effet, à la fin de la décennie 1990, le nombre de pays africains ayant conduit annuellement une enquête de séroprévalence du VIH auprès des femmes enceintes a diminué (voir Figure 1.3 page 31). La qualité globale de la surveillance sentinelle en Afrique subsaharienne a également diminué sur la même période (voir Figure 1.7). Plusieurs systèmes de surveillance se sont focalisés uniquement sur certaines populations au détriment d'autres groupes. Par ailleurs, en Afrique, la majorité des données de surveillance des femmes enceintes sont urbaines, les zones rurales étant largement sous-représentées. Les données et enquêtes socio-comportementales sont inexistantes, non suivies ou peu exploitées.

Figure 1.7
Implémentation des systèmes de surveillance sentinelle du VIH dans les 42 pays d'Afrique subsaharienne de 1995 à 2002



Sources : (GARCIA-CALLEJA 2004). La méthodologie employée est la même que celle de la précédente étude présentée à la section 1.4 (WALKER 2001). Voir Figure 1.4 page 32.

Les objectifs de cette surveillance sentinelle dite de seconde génération visent l'amélioration du suivi épidémiologique et une meilleure exploitation des différentes sources d'informations. Pour les pays à épidémie généralisée (prévalence supérieure à 1 % en population générale), le groupe de travail OMS/ONUSIDA sur la surveillance globale du VIH/SIDA et des IST préconise les

points suivants (UNAIDS/WHO WORKING GROUP ON GLOBAL HIV/AIDS AND STI SURVEILLANCE 2000) :

- une surveillance sentinelle des femmes enceintes, en milieu urbain ***et*** en milieu rural, en augmentant le nombre de cliniques prénatales afin de couvrir une plus grande part de la population ;
- une collecte des caractéristiques sociodémographiques des femmes testées afin de pouvoir comparer la population enquêtée à la population générale et extraire les jeunes femmes (15-24 ans) afin de disposer d'une approximation de l'incidence ;
- développer des enquêtes socio-comportementales répétées en population générale et auprès des plus jeunes pour suivre les changements comportementaux, afin d'évaluer l'impact des programmes d'actions (ce suivi comportemental peut, entre autres, s'appuyer sur des enquêtes déjà existantes telles que les Enquêtes Démographiques et de Santé) ;
- maintenir une surveillance sentinelle du VIH et comportementale auprès de groupes à hauts risques (travailleuses du sexe par exemple) ;
- recueillir des données de morbidité et de mortalité liées au SIDA.

Ce document d'ONUSIDA pointe par ailleurs de nombreux soucis méthodologiques concernant la représentativité des femmes enceintes suivies en clinique prénatale afin de limiter, si possible, les biais de sélection[47]. Ce rapport mentionne enfin la possibilité d'utiliser des enquêtes de séroprévalence en population générale afin de contrer ce biais, bien que ce type d'enquêtes n'ait pas sa place dans une surveillance de routine en raison de son coût et de sa complexité.

> *« Population-based serosurveillance attempts to get around selection bias associated with sentinel surveillance sites by testing specimens taken after obtaining informed consent from people randomly selected from the general population. Sampling is usually household-based.*
>
> *Population-based serosurveillance requires informed consent. Experience differs across countries and cultures, but refusal and therefore participation bias has been shown to vary substantially, even when specimens are taken by non-invasive procedures—saliva or urine as opposed to blood.*

[47] Nous ne développerons pas ici les différents points de discussion sur la question de la représentativité des femmes enceintes en clinique prénatale. Cela sera abordé ultérieurement dans la thèse (voir notamment la section 3.4 page 173.

> *General population-based serosurveys can be very helpful in indicating possible sources of bias in sentinel populations. They are expensive and difficult to conduct, and are not recommended as a routine part of serosurveillance. However, where they have been carried out for research or other purposes, their results should definitely be used to calibrate the results of routine surveillance systems.*
>
> *Some countries conduct regular population-based studies for research or planning purposes in which blood is drawn (e.g. National Health Surveys, studies on Hepatitis B, malaria, etc., Demographic and Health Surveys drawing blood for anaemia testing). In such cases, the samples collected can be used for unlinked anonymous HIV testing. National AIDS programmes should, where feasible, make use of any population-based specimen samples for HIV testing. They will need to try to ensure that the population sampled can be correlated with existing sentinel sites, so that the two data sets can be compared reliably. This may require supporting oversampling of the population in the area of one or more sentinel sites.*[48] »

Suite à cette initiative pour une surveillance sentinelle de seconde génération, plusieurs pays africains vont améliorer leur système de surveillance, notamment en

[48] UNAIDS/WHO WORKING GROUP ON GLOBAL HIV/AIDS AND STI SURVEILLANCE, *Guidelines for Second Generation HIV Surveillance*, Genève (CH), UNAIDS/WHO, 2000, p. 19. Traduction par nos soins :
« *La sérosurveillance en population générale tente de contourner le biais de sélection associé aux sites de surveillance sentinelle en testant des échantillons prélevés après consentement auprès de personnes sélectionnées aléatoirement en population générale. L'échantillonnage repose le plus souvent sur les ménages.*
La surveillance en population générale nécessite un consentement éclairé. Les expériences divergent selon les pays et les cultures, mais il a été montré que les refus et, de ce fait, les biais de participation variaient substantiellement, même lorsque les échantillons sont prélevés de manière non invasives (salive ou urine plutôt que du sang).
Les enquêtes de séroprévalence en population générale peuvent être particulièrement utiles pour mettre en évidence les sources de biais possibles de la surveillance sentinelle. Elles sont onéreuses et difficile à conduire et ne sont donc pas recommandés en tant qu'enquête de routine pour la sérosurveillance. Cependant, quand elles ont été réalisées pour des raisons de recherche ou autres, leurs résultats doivent absolument être utilisées pour calibrer les résultats de la surveillance sentinelle de routine.
Certains pays réalisent régulièrement des études en population générale, à des fins de recherche ou de planification, comportant un prélèvement sanguin (par exemple les Enquêtes National de Santé, les études sur l'Hépatite B, le paludisme, etc., les Enquêtes Démographiques et de Santé qui prélèvent du sang pour les tests d'anémie). Dans ces cas là, les échantillons collectés peuvent être utilisés pour des tests du VIH anonymes et non liés. Les programmes nationaux contre le SIDA devraient, lorsque cela est réalisable, de tout échantillon prélevé en population générale pour des tests du VIH. Ils auront besoin de s'assurer que l'échantillon en population générale pourra être corrélé avec les sites sentinelles existant afin que les deux sources de données puissent être comparées avec fiabilité. Cela pourra nécessiter un support pour sur-échantillonner les populations proches d'un ou de plusieurs sites sentinelles. »

augmentant le nombre de sites sentinelles en milieu rural. Par ailleurs sur la Figure 1.3 page 31 et la Figure 1.7 page 40, nous pouvons observer une remontée du nombre de pays ayant procéder à une enquête auprès des femmes enceintes ainsi que de la qualité des systèmes de surveillance à partir de 2001/2002.

Poursuivant dans cette lignée, l'ONUSIDA et l'OMS sortiront en 2002 un guide pratique pour la mise en place d'une surveillance de seconde génération (UNAIDS/WHO WORKING GROUP ON GLOBAL HIV/AIDS AND STI SURVEILLANCE 2002). La même année, au mois de mai, MEASURE DHS+ met en ligne la *HIV/AIDS Survey Indicators Database*[49] dont l'objectif est de centraliser dans une base de données unique différents indicateurs socio-comportementaux calculés à partir de différentes enquêtes et répondre ainsi à l'amélioration de l'utilisation de données comportementales dans le suivi épidémiologique (MEASURE DHS 2002).

Dans le mouvement initié par la surveillance de seconde génération, le Groupe de Référence en Épidémiologie décide de faire évoluer les modèles utilisés par ONUSIDA afin d'améliorer les estimations et le suivi de l'épidémie. Lors d'une réunion de travail à Rome, du 8 au 10 octobre 2000[50], il pose les bases d'une nouvelle approche qui doit permettre, entre autres, des projections du nombre de nouvelles infections à VIH à court terme (5 ans). L'approche qui était utilisée jusqu'alors, celle d'*EpiModel*, reposait sur un ajustement visuel aux données de surveillance sentinelle en clinique prénatale d'une courbe gamma représentant l'incidence du VIH et n'ayant que deux paramètres ajustables : le début de l'épidémie et un paramètre d'amplitude noté α. Deux objectifs principaux émergent de cette réunion de travail :

1. Utiliser un modèle épidémiologique plutôt qu'une courbe gamma. Un tel modèle pourrait présenter un nombre plus important de paramètres facilement interprétables. Un modèle présenté par Roy ANDERSON a retenu plus particulièrement l'attention. Il utilise 4 paramètres d'ajustement : deux paramètres épidémiologiques (un taux de progression et l'année de départ de l'épidémie) et deux paramètres démographiques (mortalité générale et taux d'entrée des individus dans une sexualité active). En outre, un cinquième paramètre pourrait être ajouté qui permettrait de refléter l'impact dans le temps des programmes d'intervention.

[49] Avec le soutien de l'USAID, de l'UNICEF, du CDC, de l'ONUSIDA, de l'OMS, du bureau américain du recensement, de FHI, de MEASURE Evaluation, du Synergy Project et du US Census Bureau International Programs Center. Pour plus de détails, voir Encadré 1.7 page 64.

[50] THE UNAIDS EPIDEMIOLOGY REFERENCE GROUP, *Recommendations*, UNAIDS Epidemiology Reference Group Meeting, Rome (IT), 8-10 octobre, UNAIDS, 2000.

2. Établir des incidences du VIH par âge afin de pouvoir réaliser des projections de l'impact démographique du VIH/SIDA, ce que ne permet pas *EpiModel*.

Le Groupe ONUSIDA de Référence en Épidémiologie a ainsi posé les recommandations suivantes :

- Continuer d'utiliser les séries temporelles de prévalence du VIH fourni par la surveillance sentinelle des femmes enceintes âgées de 15 à 49 ans.
- Ne pas inclure la structure par âge de la population dans le modèle utilisé par ONUSIDA pour réaliser des projections à court terme.
- Projeter les estimations de prévalence et d'incidence du VIH tous âges confondus selon une structure par âge préétablie pour projeter l'impact démographique.
- Passer d'*EpiModel* à un modèle épidémiologique simple dont les paramètres auraient une signification épidémiologique, biologique ou démographique.
- Avoir la possibilité de diviser la population nationale en différentes cohortes et de procéder à une modélisation séparée dans chaque sous-population.

Lors de la réunion de travail suivante de ce groupe en janvier 2001 à Gex[51] (France), six modèles différents sont présentés et discutés. Les spécifications du nouveau modèle de l'ONUSIDA sont posées. C'est ce modèle qui servira pour les estimations pays par pays à fin 2001 du rapport mondial de 2002 (UNAIDS 2002). Dans la mesure où ce modèle ne fournit pas d'estimation distribuée par âge, c'est le modèle *Spectrum*[52] et son module *AIM* (*AIDS Impact Model*) développés par the Futures Group International qui ont été utilisés pour estimer l'impact démographique des épidémies.

En juin 2002, juste avant la sortie du rapport mondial, le Groupe de Référence ONUSIDA en Épidémiologie, rebaptisé Groupe de Référence d'ONUSIDA sur les Estimations, la Modélisation et les Projections[53], publie la méthodologie retenue pour les estimations à fin 2001[54]. Un logiciel informatique implémentant ce modèle

[51] THE UNAIDS EPIDEMIOLOGY REFERENCE GROUP, *Recommended methodology for the estimation and projection of HIV prevalence and AIDS mortality in the short-term*, Meeting of the UNAIDS Epidemiology Reference Group, Gex (FR), Janvier, 2001.

[52] La première version de Spectrum est sortie en 1997, le projet ayant démarré en 1995. Pour plus de détails, voir Encadré 1.7 page 64.

[53] The UNAIDS Reference Group on Estimates, Modelling and Projections.

[54] THE UNAIDS REFERENCE GROUP ON ESTIMATES MODELLING AND PROJECTIONS, « Improved methods and assumptions for estimation of the HIV/AIDS epidemic and its impact: Recommendations of the UNAIDS Reference Group on Estimates, Modelling and Projections », *AIDS*, n°16(9), 2002.

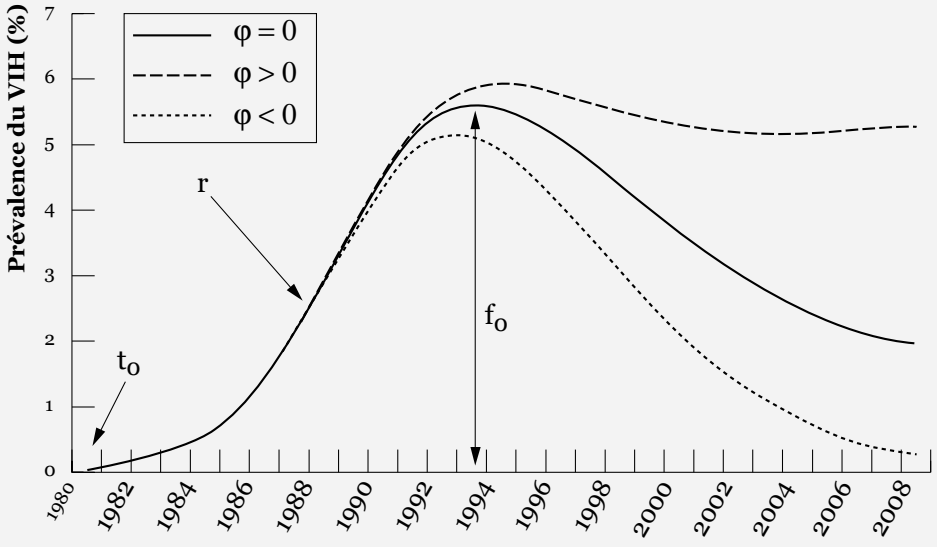
et appelé *Epidemic Projection Package* (*EPP*) est mis en ligne sur les sites web de l'OMS et de l'ONUSIDA.

Encadré 1.3
Modèle épidémiologique d'ONUSIDA implémenté dans EPP

Le modèle utilisé par EPP est un modèle épidémiologique simple qui divise la population des quinze ans et plus en trois groupes : les personnes sans risque d'être infectée, les personnes à risques et les personnes infectées. Le modèle prend en compte les décès liés au SIDA et les décès non liés au SIDA. Les nouvelles entrées dans le modèle sont constituées par les personnes arrivant à l'âge de quinze ans.

Le modèle peut être résumé en quatre paramètres (GROUPE DE RÉFÉRENCE DE L'ONUSIDA SUR LES ESTIMATIONS MODÈLES ET PROJECTIONS 2003) :

- t_o : l'année de début de l'épidémie ;
- f_o : la proportion initiale de la population faisant partie des personnes à risque ;
- r : l'intensité de l'infection, ce paramètre déterminant la vitesse à laquelle les personnes à risque deviennent infectées ;
- φ : paramètre comportemental qui affecte la répartition des nouvelles entrées dans les catégories à risque et sans risque. Si φ est négatif, la population réduit globalement ses prises de risques au cours du temps, en réponse à l'épidémie, et l'épidémie décroît après un pic. Si φ est nul, la proportion de personnes à risques reste constante, et la prévalence diminue après un pic en raison de la mortalité liée au VIH/SIDA. Si φ est positif, la population augmente au contraire ses prises de risques et la prévalence décroîtra lentement ou se stabilisera à un niveau élevé.



Source du graphique : (GHYS 2004, p. i7)

Pour les épidémies limitées ou concentrées, un cinquième paramètre *d* a été ajouté dans des versions récentes d'EPP afin de prendre en compte la mobilité importante (*turnover*) des populations à hauts risques (UNAIDS 2005).

Plusieurs nouveautés apparaissent dans le rapport mondial de 2004 (UNAIDS 2004a). Pour les pays à épidémie généralisée, *EPP*, rebaptisé *Estimation and Projection Package*, est de nouveau utilisé. L'interface du logiciel a été améliorée et les résultats obtenus peuvent être exportés vers *Spectrum*. Les courbes de prévalence du VIH sont estimées à partir des données de surveillance sentinelle des femmes enceintes en distinguant milieu urbain et milieu rural (GHYS 2004).

EPP peut être également utilisé pour les pays à épidémie concentrée, à partir de données de surveillance sentinelle dans les différents groupes à risque, en procédant à des ajustements séparément pour chaque groupe défini. Pour les pays à épidémie limitée ou concentrée, en l'absence de séries temporelles, une autre approche a été développée : l'approche *Worbook*, qui se présente sous la forme d'un classeur Excel, repose sur une estimation haute et basse de la prévalence de chaque groupe à risques et sur une estimation de leur taille respective au sein de la population générale (WALKER 2004).

Pour la première fois, le rapport mondial fournit des estimations à deux dates, en l'occurrence à fin 2001 et à fin 2003. De plus, des marges d'incertitude autour des différentes estimations sont publiées (GRASSLY 2004). La méthode n'a par contre pas encore été implémentée dans EPP. L'impact démographique de l'épidémie a été estimée à l'aide du module *AIM* du logiciel *Spectrum* (STOVER 2004). Dorénavant, *Spectrum* peut importer les estimations et les projections de prévalence du VIH réalisées avec *EPP* ou l'approche *Workbook*, puis les ventiler par âge en fonction d'un pattern préétabli. Enfin, un nouveau modèle, *Asian Epidemic Model* (*AEM*), basé à la fois sur des données épidémiologiques et comportementales, a été élaboré pour explorer l'impact de différentes politiques anti-VIH sur les épidémies asiatiques (BROWN 2004).

1.7 Retour des enquêtes nationales en population générale

En 2000, le groupe de travail ONUSIDA/OMS sur la surveillance globale du VIH/SIDA et des IST avait suggéré la possibilité de tester des échantillons sanguins prélevés lors d'enquêtes nationales en population générale[55]. Certaines enquêtes de ce type avaient déjà été réalisées à la fin des années 1980 mais elles avaient été abandonnées en raison de leur coût et de leur complexité (voir section 1.4). Pendant les années 1990, seules des enquêtes en population générale sur des zones géographiques limitées ont été maintenues. Ce fut le cas notamment au Bénin

[55] UNAIDS/WHO WORKING GROUP ON GLOBAL HIV/AIDS AND STI SURVEILLANCE, *Guidelines for Second Generation HIV Surveillance*.

(BUVE 2001a), au Cameroun (BUVE 2001a, GLYNN 2001a), au Kenya (BUVE 2001a, GLYNN 2001a, 2001b), au Malawi (CRAMPIN 2003), en Ouganda (KILIAN 1999), en Tanzanie (CHANGALUCHA 2002), en Zambie (BUVE 2001a, FYLKESNES 2001, GLYNN 2001a, 2001b) et au Zimbabwe (GREGSON 2002a, 2002b).

En 2001, au Mali, est conduite la première Enquête Démographique et de Santé (EDS) incorporant un dépistage du VIH (pour plus de détails sur les EDS voir Encadré 1.4). Une seconde EDS incluant un dépistage du VIH sera réalisée fin 2001-début 2002 en Zambie. Les EDS sont des enquêtes de routine conduites tous les quatre à cinq ans dans de nombreux pays en développement et en particulier en Afrique. Ce sont des enquêtes standardisées présentant d'un pays à l'autre un échantillonnage similaire (voir Encadré 1.5).

Elles permettent donc de réaliser des comparaisons dans le temps et entre pays. Leur réalisation, en collaboration avec les autorités de chaque pays, suit une procédure relativement bien rodée. Par ailleurs, elles incluaient dès les années 1990 des questions sur les connaissances du VIH/SIDA et sur les comportements sexuels et reproductifs. Elles sont ainsi, pour de nombreux pays, la principale source de données socio-comportementales. Ce n'est donc pas un hasard si la *HIV/AIDS Survey Indicators Database* est gérée par MEASURE DHS+, l'organisme chargé de coordonner l'ensemble des EDS et dont certains membres font partie du Groupe de Référence OMS/ONUSIDA sur la Surveillance Globale du VIH/SIDA et des IST. Enfin, de nombreuses EDS incluaient déjà des prélèvements sanguins à des fins de dépistage de l'anémie et/ou du paludisme. Les EDS, qui bénéficiaient en outre d'un soutien financier important de l'USAID, étaient donc tout indiquées pour développer une mesure de la prévalence du VIH en population générale.

Les résultats des deux premières EDS avec dépistage du VIH ne pouvaient être liés aux données comportementales des questionnaires afin de garantir l'anonymat des personnes enquêtes. En effet, outre un questionnaire ménage, un questionnaire femme, un questionnaire homme et un questionnaire couple, les EDS collectent les coordonnées GPS du point central des zones d'enquêtes. Le croisement de l'ensemble de ces informations permettrait de pouvoir retrouver et identifier les individus enquêtés.

Cependant, pour les enquêtes suivantes, MEASURE DHS+ a mis au point une procédure permettant de lier les résultats des tests de dépistage du VIH et les données des questionnaires. Cette procédure, annoncée en juillet 2005 lors du XXV^{e} Congrès International de la Population de l'UIESP à Tours, consiste à décaler aléatoirement les coordonnées géographiques des zones d'enquêtes dans un rayon de 2 kilomètres en milieu urbain et de 5 kilomètres en milieu rural[56]. Cette

[56] MEASURE DHS, *Methodology - Collecting Geographic Data*, 2006, page web consultée le 10 juillet 2006. (http://www.measuredhs.com/topics/gis/methodology.cfm)

procédure permet de garder une précision géographique suffisante pour une analyse nationale. À partir de 2004, MEASURE DHS+ a élaboré des enquêtes comparables aux EDS mais spécifiques à la problématique VIH/SIDA et donc avec un questionnaire allégé. Il s'agit des *AIDS Impact Surveys* (*AIS*).

Encadré 1.4

Historique des Enquêtes Démographiques et de Santé (EDS)

Le projet *DHS* (*Demographic and Health Surveys* soit *Enquêtes Démographiques et de Santé* en français) est né en 1984 et constitue le troisième grand projet de recherche mondial initié par l'USAID. Les DHS succèdent ainsi aux *World Fertility Surveys* (*WFS* ou *Enquêtes Mondiales de Fécondité*) conduites dans plus de 60 pays dans le monde de 1972 à 1984 et aux *Contraceptive Prevalence Surveys* (*CPS* ou *Enquêtes sur la Prévalence de la Contraception*) réalisées entre 1977 et 1985.

De 1984 à 1989, le projet DHS a été coordonné par l'*Institute for Resource Development Inc.*, une filiale de la *Westinghouse Electric Company*, puis fut repris par *Macro International Inc.* qui coordonne toujours ce projet actuellement.

En 1997, le projet DHS, devenu l'un des quatre composants du programme *MEASURE* (*Monitoring and Evaluation to Assess and Use Results*) de l'USAID, est rebaptisé *MEASURE DHS+*.

Depuis 1984, plus de 200 enquêtes ont été réalisées dans plus de 75 pays. Aux questions relatives à la fécondité, au planning familial et à la mortalité infanto-juvénile, ont été progressivement ajoutés, selon les pays et les enquêtes, des modules sur la santé de la mère et de l'enfant, les connaissances et comportements vis-à-vis du VIH/SIDA et des IST, les violences domestiques, les mutilations génitales féminines, le géoréférencement des zones d'enquêtes (latitude et longitude), des mesures de poids et de taille des enfants, des tests d'anémie, etc. Depuis 2001 certaines enquêtes incorporent des tests de dépistage du VIH en population adulte.

Il s'agit d'enquêtes auprès des ménages, en population générale, représentatives usuellement au niveau national et régional. Le nombre de ménages par enquête se situe entre 5 000 et 30 000. Les EDS comportent un questionnaire ménages, un questionnaire femmes et un questionnaire hommes ainsi qu'une enquête couples. Les questionnaires sont standardisés afin de faciliter des comparaisons dans le temps et entre pays.

Elles sont conduites le plus souvent tous les cinq ans, par les instituts nationaux de la statistique, avec l'appui technique de Macro International Inc. Des enquêtes intermédiaires peuvent être réalisées, avec un questionnaire allégé, sur une thématique donnée. C'est le cas notamment des *AIDS Impact Survey* (*AIS*) et des *Interim DHS*.

L'ensemble des rapports finaux ainsi que les bases de données sont disponibles sur un site dédié : http://www.measuredhs.com (pour plus de détails, voir Encadré 1.7 page 64).

Encadré 1.5
Échantillonnage des Enquêtes Démographiques et de Santé (EDS)

Les EDS sont conçues pour fournir une image de la population générale nationale. Nous présentons ici le plan d'échantillonnage usuel des EDS et des enquêtes de même type (comme les AIS), ce dernier pouvant différer légèrement d'une enquête à l'autre.

Enquête auprès des ménages

Il s'agit d'enquêtes auprès des ménages, représentatives au niveau national, par milieu de résidence et par région. Le nombre de ménages enquêtés se situe le plus souvent entre 5 000 et 30 000. Les EDS présentent un plan de sondage comparable dans chaque pays. Il s'agit d'enquêtes stratifiées avec un tirage à deux degrés.

Stratification du pays

Le pays est divisé en plusieurs strates, une par région administrative et par milieu de résidence. Le plus souvent la capitale du pays et, éventuellement, les grandes villes de même taille sont considérées comme une seule strate. Usuellement, on compte entre 5 et 20 régions par pays. Le tirage au premier degré est réalisé séparément pour chaque strate. Selon les enquêtes, certaines strates, faiblement peuplées, peuvent être surreprésentées.

Tirage au premier degré

La base de sondage des unités primaires est typiquement composée des zones de dénombrement au dernier recensement de la population effectué dans le pays. Au premier degré, les unités primaires ou grappes sont tirées au sort, séparément dans chaque strate, avec une probabilité proportionnelle au nombre de ménages de la grappe lors du dernier recensement de population. La répartition spatiale de ces grappes peut être considérée comme une approximation (grossière) de la densité de la population. Dans certaines enquêtes, les coordonnées latitude/longitude du centre de chaque grappe sont collectées par GPS. Depuis l'arrivée des tests de dépistage du VIH, les coordonnées des grappes sont décalées aléatoirement dans un rayon de 2 kilomètres en milieu urbain et de 5 kilomètres en milieu rural.

Tirage au second degré

Après un recensement exhaustif des ménages de chaque grappe, un nombre prédéterminé de ménages est sélectionné au second degré, par tirage au sort simple, pour l'enquête ménages et le questionnaire individuel femmes (15-49 ans). Suivant le pays, seule une partie des ménages enquêtés est retenue pour le questionnaire hommes (15-59 ans). Si l'enquête comporte un dépistage du VIH, le test est alors proposé à l'ensemble des femmes et des hommes éligibles appartenant aux ménages sélectionnés pour l'enquête hommes.

Pondération des résultats

Afin de tenir compte du plan d'échantillonnage complexe des EDS, chaque base de données contient une variable de pondération statistique à appliquer aux individus afin de rendre l'échantillon représentatif au niveau national et régional. Cette variable de pondération est proportionnelle à l'inverse de la probabilité de sondage de chaque ménage, c'est-à-dire à la probabilité que le ménage en question soit enquêté.

Depuis 2001, plus d'une quinzaine d'EDS et d'AIS avec dépistage du VIH ont été conduites en Afrique subsaharienne et plusieurs autres enquêtes sont actuellement en cours. Par ailleurs, certains pays ont conduit d'autres enquêtes nationales en population générale adulte avec dépistage du VIH : Burundi en 2002, Congo en 2003 (milieu urbain uniquement), Guinée Équatoriale en 2004, Niger en 2002, Afrique du Sud en 2002 (15-24 ans uniquement) et en 2005 (enfants et adultes), Sierra Leone en 2005 et Zimbabwe en 2002 (15-29 ans uniquement)[57].

Les résultats de ces enquêtes ont parfois produit des résultats divergents fortement avec les estimations réalisées par ONUSIDA jusqu'alors. Par exemple, l'EDS réalisée en 2003 au Burkina Faso a mesuré une prévalence nationale de 1,8 % tandis que l'estimation à fin 2003, dans le rapport mondial de 2004, était de 4,2 %. Au Kenya, l'EDS réalisée en 2003 a mesuré une prévalence de 6,7 %, identique à l'estimation à fin 2003 réalisée par ONUSIDA. Mais la précédente estimation à fin 2001 publiée en 2002 était de 15,0 % (voir Figure 1.8 page 52). La concordance entre l'EDS et l'estimation à fin 2003 n'est pas due au hasard : les résultats de l'EDS étaient disponibles dès janvier 2004, avant la publication du rapport de l'ONUSIDA, tandis qu'au Burkina Faso les résultats de l'EDS ont été publiés plus tardivement, après la publication du rapport mondial de 2004. Les résultats kenyans ont suscité une polémique sur la qualité des EDS et des estimations de l'ONUSIDA. ONUSIDA a alors précisé dans un communiqué de presse[58] du 13 janvier 2004 que l'on ne pouvait conclure à une baisse de la prévalence du VIH, que les estimations réalisées dépendaient de la qualité de la couverture sentinelle, cette dernière étant sous-représentée en milieu rural, et enfin qu'il était nécessaire de mieux investiguer le biais dû aux refus de se faire tester dans les EDS (de l'ordre de 15 % au Kenya).

Dès 2003, trois chercheurs de l'OMS et de l'ONUSIDA publient un article dans le *Lancet* où ils précisent la nécessité de mieux comprendre les différences observées entre les enquêtes nationales en population générale et les données issues de la surveillance sentinelle des femmes enceintes (BOERMA 2003). Ils analysent plusieurs sources de biais potentiels.

D'une part, les hypothèses utilisées pour l'estimation des prévalences nationales à partir des femmes enceintes mériteraient d'être affinées selon les pays (réduction de 20 % de la prévalence observée en milieu rural pour prendre en compte la sous-représentation des cliniques prénatales, ratio femmes/hommes de la prévalence du

[57] GARCIA-CALLEJA J. M., GOUWS E. et GHYS P. D., « National population based HIV prevalence surveys in sub-Saharan Africa: results and implications for HIV and AIDS estimates », *Sexually Transmitted Infections*, n°82(suppl_3), 2006.

[58] UNAIDS, *UNAIDS response to Kenyan HIV prevalence survey*, Press statement, Genève (CH), 13 janvier, 2004b.

VIH de 1,2). D'autre part, des taux de participation au test de dépistage du VIH relativement faibles dans les enquêtes en population générale (76,5 % en Zambie, 80,7 % au Mali et 62,1 % en Afrique du Sud[59]) laissent à penser que les prévalences du VIH mesurées dans ce type d'enquêtes sous-estiment les prévalences réelles, les personnes migrantes et celles ayant pris le plus de risques étant également les plus susceptibles d'être absentes au moment de l'enquête ou d'avoir refusé le test.

Il est apparu clairement que l'arrivée des enquêtes nationales en population générale avec mesure de la prévalence du VIH impliquait le développement de « *nouvelles stratégies de surveillance du VIH* »[60].

Lors d'un meeting tenu en mai 2004[61], le Groupe de Référence ONUSIDA sur les Estimations, la Modélisation et les Projections a émis une série de remarques et de recommandations à la fois sur la surveillance sentinelle des femmes enceintes et les enquêtes nationales en population générale :

- Les données sur les femmes enceintes sont pertinentes pour l'analyse des tendances dans le temps et des différentiels régionaux mais non représentatives pour déterminer les niveaux nationaux.
- Par contre, ces données ont l'avantage d'être produites annuellement tandis que les enquêtes en population générale sont répétées beaucoup plus rarement.
- La couverture sentinelle en milieu rural doit être augmentée.
- Les rapports annuels devraient ventiler les résultats par âge et par site selon un modèle standardisé afin de faciliter les comparaisons.
- Les coordonnées géographiques (longitude/latitude) des sites sentinelles devraient être collectées.

---

[59] CENTRAL STATISTICAL OFFICE, CENTRAL BOARD OF HEALTH et ORC MACRO, *Zambia Demographic and Health Survey 2001–2002: preliminary report*, Calverton, Maryland (US), Central Statistical Office, Central Board of Health, ORC Macro, 2002.
CELLULE DE PLANIFICATION ET DE STATISTIQUE, MINISTÈRE DE LA SANTÉ *et. al.*, *HIV testing in Mali: findings from the 2001 Mali Demographic and Health Survey*, Calverton, Maryland (US), Cellule de Planification et de Statistique, ministère de la Santé, Direction Nationale de la Statistique et de l'Informatique (DNSI), ORC Macro, 2002.
SHISANA O. et SIMBAYI L., *Nelson Mandela/HSRC study of HIV/AIDS: South African national HIV prevalence, behavioural risks and mass media-household survey 2002*, Cape Town (ZA), Human Sciences Research Council, 2002.

[60] DIAZ T., DE COCK K. *et. al.*, « New strategies for HIV surveillance in resource-constrained settings: an overview », *AIDS*, n°19 Suppl 2, 2005.

[61] THE UNAIDS REFERENCE GROUP ON ESTIMATES MODELLING AND PROJECTIONS, *Responding to surveillance: Methods and software to produce HIV/AIDS estimates in the era of population-based prevalence surveys*, Technical Report and Recommendations - Report of a meeting on the UNAIDS Reference Group, Glion (CH), May 10-11th, UNAIDS, 2004a.

Figure 1.8

Prévalences du VIH selon différentes sources pour quinze pays d'Afrique subsaharienne (1998-2006)

		1998	1999	2000	2001	2002	2003	2004	2005	2006
Burkina Faso	ONUSIDA			6,4		6,5		4,2		2,0
	EDS							1,8		
	CPN urbain	7,4 (5)	7,3 (4)	6,3 (4)		5,3 (4)	4,0 (2)	4,2 (2)		
	CPN rural	4,3 (3)	6,0 (3)	4,2 (4)		4,2 (4)	1,8 (8)	2,1 (10)		
		1998	1999	2000	2001	2002	2003	2004	2005	2006
Cameroun	ONUSIDA			7,7		11,8		6,9		5,4
	EDS								5,5	
	CPN urbain			12,0 (2)	6,6 (1)	7,0 (6)				
	CPN rural	9,2 (3)	3,3 (1)	10,0 (25)		5,6 (57)	9,1 (1)			
		1998	1999	2000	2001	2002	2003	2004	2005	2006
Côte d'Ivoire	ONUSIDA			10,8		9,7		7,0		7,1
	AIS									4,7
	CPN urbain	12,9 (2)	13,3 (1)	11,3 (2)	6,0 (7)	6,3 (4)		9,8 (4)		
	CPN rural	9,8 (9)		8,8 (9)	7,8 (19)	7,0 (24)		8,3 (8)		
		1998	1999	2000	2001	2002	2003	2004	2005	2006
Ethiopie	ONUSIDA			10,6		6,4		4,4		nd
	EDS									1,4
	CPN urbain			14,9 (4)	16,2 (4)	13,7 (4)	11,8 (5)			
	CPN rural	9,2 (5)	11,8 (8)	3,1 (5)	10,3 (32)	11,5 (31)	7,4 (61)			
		1998	1999	2000	2001	2002	2003	2004	2005	2006
Ghana	ONUSIDA			3,6		3,0		3,1		2,3
	EDS							2,2		
	CPN urbain	3,4 (3)	2,2 (5)	3,7 (4)	2,5 (5)	4,1 (6)		3,9 (4)	2,7 (23)	
	CPN rural	3,4 (14)	2,8 (19)	2,1 (17)	3,2 (18)	3,3 (18)		3,2 (29)	3,0 (17)	
		1998	1999	2000	2001	2002	2003	2004	2005	2006
Guinée	ONUSIDA			1,5		nd		3,2		1,5
	EDS								1,5	
	CPN urbain				5,0 (1)			4,1 (5)		
	CPN rural				2,4 (4)			4,3 (13)		
		1998	1999	2000	2001	2002	2003	2004	2005	2006
Kenya	ONUSIDA			14,0		15,0		6,7		6,1
	EDS							6,7		
	CPN urbain	13,7 (2)	16,4 (5)	15,0 (7)	14,3 (6)	10,0 (4)	12,0 (6)	10,0 (6)		
	CPN rural	14,7 (21)	22,7 (22)	14,4 (17)	11,6 (20)	6,0 (1)		6,0 (34)		
		1998	1999	2000	2001	2002	2003	2004	2005	2006
Lesotho	ONUSIDA			23,6		31,0		28,9		23,2
	EDS								23,5	
	CPN urbain			42,2 (1)			35,0 (1)			
	CPN rural			19,0 (5)			27,0 (5)			
		1998	1999	2000	2001	2002	2003	2004	2005	2006

Sources :

EDS (Enquêtes Démographiques et de Santé) et *AIS* (AIDS Impact Survey) : rapport final de chaque enquête. Prévalence du VIH observée chez les hommes et les femmes adultes (15-49 ans). Le rectangle indique la date de publication du rapport final. La pointe de la flèche indique le milieu de la période de collecte des données (qui dure en général de trois à six mois).

ONUSIDA : il s'agit des estimations de la prévalence adulte pays par pays réalisées par ONUSIDA et publiées dans son rapport biennal. Les estimations reportées ici correspondent aux estimations à fin 1999 du rapport 2000, fin 2001 du rapport 2002, fin 2003 du rapport 2004 et fin 2005 du rapport 2006. Les rapports d'ONUSIDA sont disponibles sur www.unaids.org.

CPN (Cliniques Pré-Natales) : les données sont issues des *Epidemiological Fact Sheets on HIV/AIDS and sexually transmitted infections 2006 Update* de chaque pays publiés par l'OMS, l'UNICEF et l'ONUSIDA. Valeur médiane de la prévalence observée chez les femmes enceintes suivies en clinique prénatale dans le cadre de la surveillance sentinelle. Le nombre entre parenthèses représente le nombre de sites.

CPN urbain : sites considérés comme appartenant à une zone urbaine principale (*major urban areas*).

CPN rural : sites considérés comme n'appartenant pas à une zone urbaine principale (*outside major urban areas*).

- Une typologie standard devrait être élaborée pour décrire et suivre dans le temps les populations consultant chaque clinique sentinelle.
- Les enquêtes en population générale devraient réaliser des analyses précises des caractéristiques des non-testés (refus ou absence) afin de pouvoir procéder le cas échéant à des ajustements.
- Les résultats des enquêtes en population générale devraient être ajustés au minimum en tenant compte de l'âge et du sexe des non répondants.
- Des comparaisons devraient être effectuées systématiquement entre les données issues de la surveillance sentinelle et celles issues des enquêtes en population générale, d'une part à un niveau local et, d'autre part, en isolant les femmes enceintes ayant consulté en suivi prénatal dans les enquêtes en population générale.
- Pour les pays à épidémie généralisée, les mesures effectuées en population générale restent néanmoins *a priori* plus proches de la prévalence que les observations effectuées auprès des femmes enceintes. Pour les pays à épidémie concentrée et, plus globalement, les pays où la prévalence nationale est inférieure à 3 %, des enquêtes spécifiques auprès des populations à risques restent plus que nécessaires, les enquêtes en population générale étant susceptibles de sous-estimer la prévalence réelle de l'épidémie dans la mesure où ces populations à risques sont plus difficiles à enquêter dans le cadre des enquêtes nationales.
- La taille des populations non enquêtées devrait être estimée dans la mesure du possible (internat, prisons, camps militaires, etc.).
- Les résultats des tests de dépistage des enquêtes en population générale devraient pouvoir être liés aux données issues des questionnaires, si possible.
- Les questionnaires ménages devraient collecter un maximum d'information sur les membres du ménage absents lors de l'enquête, notamment les détails de leur migration (lieu, durée, raison, etc.)
- Les courbes de prévalence calculée par *EPP* doivent toujours être calculées à partir des données de surveillance sentinelle des femmes enceintes puisqu'il s'agit des seules données pour lesquelles on dispose de séries temporelles. Cependant, *EPP* devrait permettre de pouvoir les ajuster par la suite en tenant compte des informations fournies par les enquêtes en population générale.
- Les patterns de distribution des infections à VIH par âge et sexe utilisés dans *Spectrum* doivent continuer à être calculés sur des bases régionales, les données issues des enquêtes nationales étant soumises à de plus grandes variations aléatoires.
- Pour les épidémies concentrées, en l'absence de séries temporelles, l'approche *Workbook* reste préconisée.

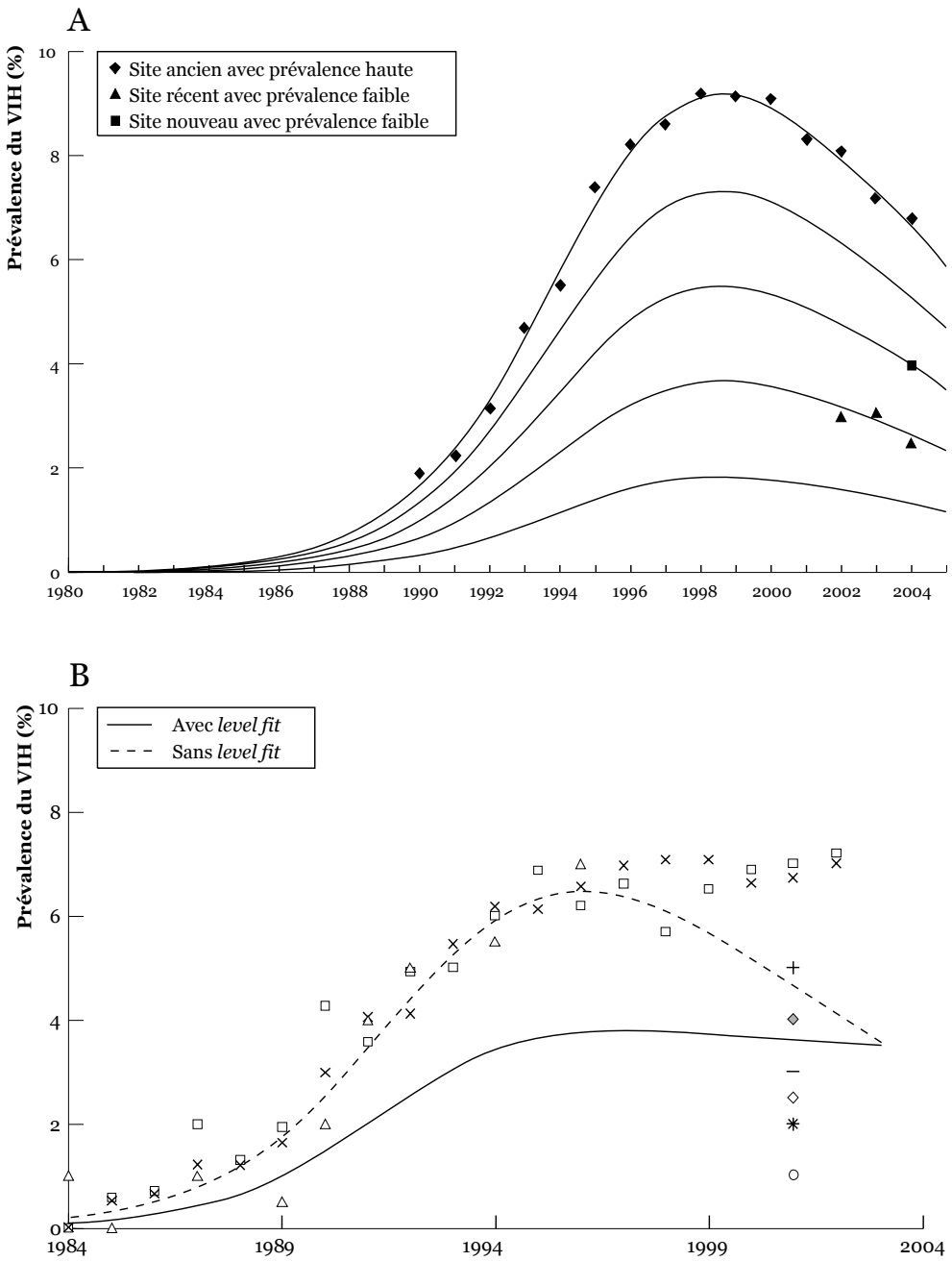
Pendant l'année 2004, plusieurs évolutions ont été apportées au logiciel *EPP* au point de parler d'une version V2[62]. Ce travail va aboutir à ce qui sera nommé *EPP 2005*, une version améliorée du modèle, utilisé pour les estimations du rapport mondial 2006 (UNAIDS 2006). À l'occasion des dix ans de l'ONUSIDA, le rapport 2006 fournit, en plus des estimations usuelles à fin 2003 et fin 2005, une fiche technique pour chaque pays du monde présentant une synthèse de différents indicateurs épidémiologiques, démographiques, politiques et économiques. En décembre 2006, un numéro spécial de la revue *Sexually Transmitted Infections* (supplément 3) a été consacré aux nouvelles méthodes utilisées par ONUSIDA pour ses différentes estimations. Outre une présentation des nouveautés introduites dans l'approche *Workbook* (LYERLA 2006), un nouveau modèle a été présenté, *Modes of Transmissions* (*MoT*), destiné à estimer les incidences du VIH par modes de transmission afin de suivre et de mesurer l'impact des programmes d'actions sur les différentes populations à risques (GOUWS 2006). Nous nous arrêterons plus particulièrement sur les nouveautés introduites dans *EPP 2005*, à savoir la procédure *level fit* et le calibrage des courbes de prévalence (BROWN 2006).

La procédure *level fit*, ou ajustement de niveau en français, a été développée afin de prendre en compte l'extension progressive du système de surveillance sentinelle, notamment en milieu rural. Jusqu'alors, la courbe des prévalences était ajustée à l'ensemble des observations. Or, dans nombre de pays, pour les premières années de l'épidémie, on ne dispose que de mesures dans les zones les plus touchées. Ainsi, lors de l'extension du système de surveillance sentinelle au milieu rural, les années plus récentes présentent un nombre de points de mesure plus important et des prévalences souvent plus faibles. En cas d'ajustement à l'ensemble des observations, la courbe des prévalences s'avère être élevée les premières années, correspondant à ce qui a été observé dans les zones les plus touchées, puis diminue artificiellement pour les années récentes du fait de la présence des nouvelles observations (courbe en pointillée sur la Figure 1.9 B).

La procédure *level fit* effectue un ajustement séparé pour chaque site sentinelle en considérant que tous les sites d'une même région suivent un même pattern (voir Figure 1.9 A). Les sites ajoutés récemment au système de surveillance sentinelle influent donc sur l'ensemble de la courbe produite, les sites les plus anciens permettant d'estimer la forme de la courbe (voir courbe en trait plein sur la Figure 1.9 B).

[62] THE UNAIDS REFERENCE GROUP ON ESTIMATES MODELLING AND PROJECTIONS, *Development of EPPv2 and Spectrum, and Measuring and tracking concentrated HIV epidemics*, Technical Report and Recommendations - Report of a meeting on the UNAIDS Reference Group, Sintra (PT), December 8-10th 2004, 2004b.

Figure 1.9
Procédure « level fit » implémentée dans EPP 2005



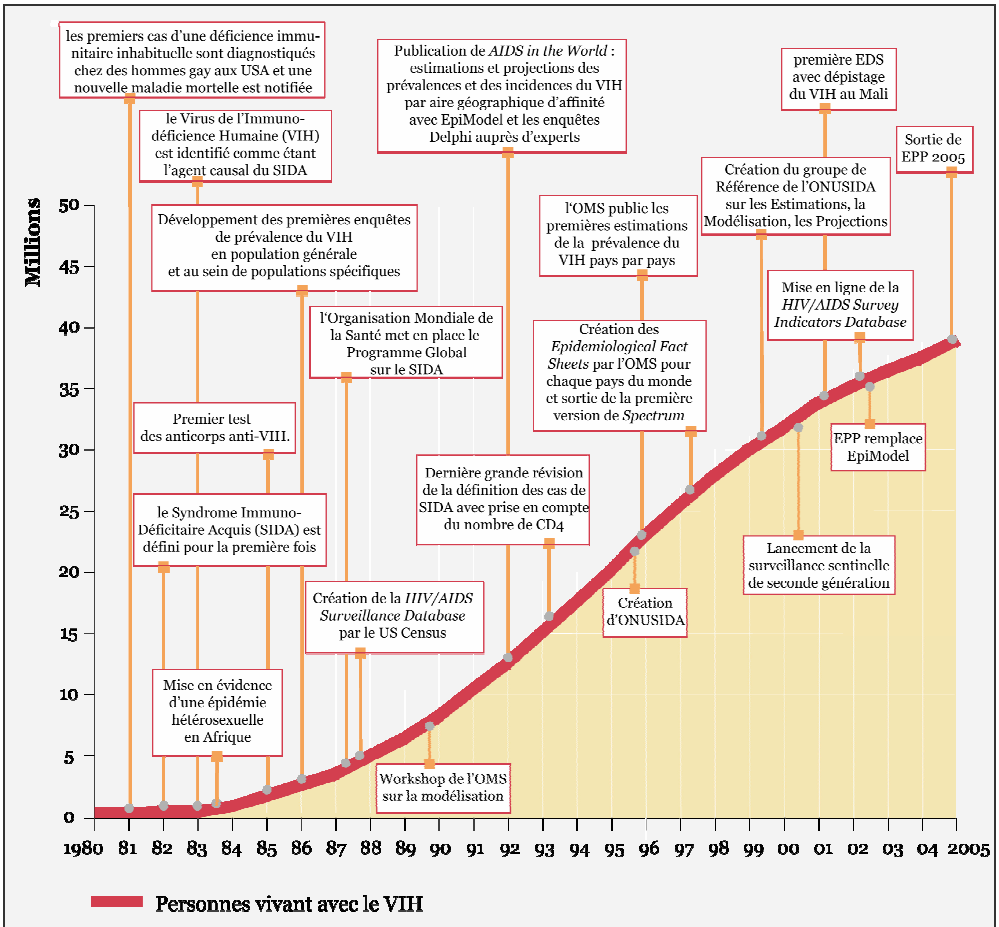
Sources : (BROWN 2006). La procédure *level fit* considère que la prévalence du VIH suit un même pattern dans chaque site d'une même région (A). La courbe globale est obtenue par combinaison des courbes de chaque site (B), tandis qu'en l'absence de la procédure *level fit* la courbe est ajusté à l'ensemble des observations.

Une fois les courbes de prévalence ajustées, il est possible de les calibrer à partir des résultats fournis par une enquête en population générale. Typiquement, les prévalences observées en milieu rural et en milieu urbain vont être spécifiées dans le modèle, ainsi que l'année d'observation, puis un facteur correctif multiplicatif va être appliqué à l'ensemble des courbes de manière à ce que les courbes urbaine et rurale passent par les points spécifiés. Les tendances de l'épidémie sont alors estimées à partir de la surveillance sentinelle des femmes enceintes tandis que le niveau est, quant à lui, déterminer à partir d'une enquête nationale en population générale.

La Figure 1.8, qui présente simultanément les résultats de quinze EDS ou AIS, les observations réalisées auprès des femmes enceintes et les différentes estimations présentes dans les rapports mondiaux d'ONUSIDA, montre comment les résultats des enquêtes en population générale ont été progressivement pris en compte dans les estimations biennales de l'ONUSIDA. On remarquera également au passage un effet pervers des EDS : la surveillance sentinelle des femmes enceintes a diminué ces dernières années, certains pays n'ayant pas renouvelé le suivi sentinelle en clinique prénatale l'année de l'EDS ou l'année suivante.

Encadré 1.6

Chronologie récapitulative : 25 ans de surveillance du VIH/SIDA



Sources : pour le nombre de personnes vivant avec le VIH, Rapport Mondial de 2006 (UNAIDS 2006), figure 1.1 page 4.

1980	Première description d'un cas de dépression du nombre de lymphocytes T4 chez un homosexuel masculin à Los Angeles.
1981	Des cas analogues sont décrits aux USA et en Europe. Le terme AIDS est utilisé par des médecins américains.
1982	Le terme AIDS apparaît pour la première fois dans le MMWR. Le CDC définit le SIDA comme une liste de pathologies indicatrices d'un déficit immunocellulaire.
1983	Découverte du LAV par une équipe française menée par Luc MONTAGNIER et Françoise BARRÉ-SINOUSSI. Le terme SIDA apparaît dans le REH de l'OMS et le BEH de la DGS. L'OMS publie une première série de chiffres européens (153 cas au 30 juin 1983).

1984	Max GALLO isole le HTLV-III. Premières publications de données de prévalence du VIH portant sur l'Afrique subsaharienne (selon la *HIV/AIDS Surveillance Database*).
1985	Premier atelier de l'OMS sur le SIDA en Afrique, à Bangui du 22 au 25 octobre 1985 (neuf pays présents). Élaboration d'une définition « clinique » des cas de SIDA. Premières publications de prévalence du VIH sur des femmes enceintes en Afrique subsaharienne (selon la *HIV/AIDS Surveillance Database*).
1986	Élaboration d'un plan d'action pour la lutte contre le SIDA en Afrique subsaharienne, à Brazzaville, du 3 au 7 mars 1986 (41 pays présents). 7 pays africains avaient, au 6 mars 1986, enregistré des cas de SIDA. Le terme de VIH est adopté par la communauté scientifique pour désigner le LAV/HTLV-III. Publication des premiers chiffres africains sur le nombre de cas de SIDA dans le REH de l'OMS. Au 14 novembre, dix pays avaient notifié 1 069 cas. Publication des premiers travaux de modélisation mathématique de la dynamique des épidémies de VIH. Enquête nationale en population générale de séroprévalence du VIH au Rwanda (décembre).
1987	Révision par l'OMS de la définition des cas de SIDA avec prise en compte de la sérologie anti-VIH, suite au changement de définition par le CDC en 1985. Création du Programme Global (GPA) sur le SIDA à l'OMS. Création de la *HIV/AIDS surveillance database* par le Center for International Research (CIR) of the United States Bureau of the Census avec le support de l'USAID. Réalisation d'une enquête nationale de séroprévalence du VIH en population générale en Ouganda (septembre 1987 – janvier 1988). Le Programme Global sur le SIDA de l'OMS développe un modèle simple d'estimations des cas de SIDA à partir des données des enquêtes de prévalence.
1989	Enquête nationale en population générale de séroprévalence du VIH en Côte d'Ivoire (février) suite à d'autres enquêtes du même type menées dans le pays depuis 1986. Le modèle épidémiologique du Programme Global sur le SIDA est présenté à la III[e] Conférence Internationale sur le SIDA et les cancers associés en Afrique. L'OMS organise du 13 au 15 décembre, à New York, un Workshop sur la modélisation de l'impact démographique du SIDA. Alberto PALLONI présente une synthèse des différents modèles développés et en distingue six types.
1991	Le Programme Global de l'OMS développe *EpiModel*, un logiciel permettant simplement d'appliquer son modèle épidémiologique à ses propres données.

1992	Publication de *AIDS in the World*. Des estimations de prévalence et d'incidence du VIH par aire géographique d'affinité (GAA) sont réalisées, à partir d'*EpiModel* et des enquêtes *Delphi*, du début de l'épidémie au 1er janvier 1992 ainsi que des projections jusqu'en 1995.
1993	Révision par le CDC de la définition des cas de SIDA avec prise en compte du taux sanguin de lymphocytes T4.
1995	Création le 1er décembre du programme commun des Nations unies sur le VIH/SIDA (ONUSIDA). Démarrage du projet *Spectrum*. Le programme global de l'OMS réalise les premières estimations de l'épidémie pour chaque pays du monde à fin 1994. Ces dernières sont basées sur les données de surveillance disponibles, en faisant plusieurs hypothèses, et sur le logiciel *EpiModel*. Les résultats provisoires sont publiés dans le REH du 15 décembre.
1996	Publication de *AIDS in the World II*. Les estimations par pays à fin 1994, affinées, y sont publiées. Le rapport fournit également, par aire géographique détaillée, des estimations détaillées de 1977 à 1996 et des projections jusqu'en 2001. Symposium satellite sur les tendances de l'épidémie organisé par l'ONUSIDA lors de la XIe Conférence Mondiale sur le SIDA, à Vancouver, en juillet. Novembre : création du *groupe de travail ONUSIDA/OMS sur la surveillance globale du VIH/SIDA et des IST*.
1997	Mise en place par l'OMS, conjointement avec ONUSIDA, des *Epidemiological Fact Sheet* pour chaque pays. Sortie de la première version de *Spectrum*. Celle-ci comporte un module sur le VIH/SIDA : *AIM* (*AIDS Impact Model*).
1998	Juin : premier rapport d'ONUSIDA sur l'épidémie mondiale avec des estimations à fin 1997, pays par pays, selon la même méthodologie que celle utilisée en 1995.
1999	ONUSIDA crée un *Groupe de Référence en Épidémiologie* qui deviendra plus tard le *Groupe de Référence d'ONUSIDA sur les Estimations, la Modélisation et les Projections*.
2000	L'ONUSIDA et l'OMS lancent une initiative pour une surveillance sentinelle de seconde génération et recommandent, entre autres, le développement de la surveillance comportementale et l'extension de la surveillance sentinelle des femmes enceintes au milieu rural. Juin : second rapport d'ONUSIDA avec des estimations pays par pays à fin 1999 selon la même méthodologie qu'en 1995 et 1998. Désormais, une estimation haute et une estimation basse sont publiées en plus de l'estimation centrale. Octobre : réunion du Groupe de Référence en Épidémiologie d'ONUSIDA à Rome : le groupe recommande de faire évoluer les modèles utilisés pour l'estimation des prévalences du VIH.

2001	Janvier : réunion du Groupe de Référence en Épidémiologie d'ONUSIDA à Gex (France) : après comparaison de plusieurs modèles épidémiologiques, le groupe a élaboré des recommandations pour les futures estimations à fin 2001. Réalisation de la première EDS avec dépistage du VIH au Mali (rapport publié en 2002).
2002	Mai : mise en ligne de la *HIV/AIDS Survey Indicators Database*. Juillet : troisième rapport d'ONUSIDA avec les estimations pays par pays à fin 2001 réalisées selon les recommandations du Groupe de Référence en Épidémiologie. Le nouveau modèle utilisé a donné naissance au logiciel *EPP* (*Epidemic Projection Package*) qui remplace *EpiModel*. L'impact des épidémies est estimé à l'aide du module *AIM* de *Spectrum*.
2003	Le Groupe de Référence de l'ONUSIDA développe l'approche *Wordbook* pour estimer les prévalences dans les pays à épidémie limitée ou concentrée et l'*Asian Epidemic Model* (*AEM*) pour estimer l'impact de différentes politiques d'intervention en Asie.
2004	Juillet : quatrième rapport d'ONUSIDA avec les estimations par pays à fin 2001 et fin 2003. *EPP*, devenu *Estimation and Projection Package*, est utilisé pour estimer les prévalences et l'impact des épidémies est estimé à l'aide du module *AIM* de *Spectrum*.
2005	Août : sortie de *EPP* 2005 (2.0). *EPP* incorpore désormais une procédure dite *level fit* et permet de calibrer les projections sur les données des EDS. L'ONUSIDA développe un nouveau modèle, *Modes of Transmission* (*MoT*), pour estimer à court terme l'incidence du VIH selon le mode de transmission. MEASURE DHS+ met en place une procédure permettant de lier, dorénavant, dans les EDS et les AIS, les données VIH et les coordonnées GPS.
2006	Cinquième rapport d'ONUSIDA avec les estimations par pays à fin 2003 et fin 2005 et publication d'une fiche profil par pays. *EPP* 2005 a été utilisé.
2007	Sortie de EPP 2007 (release 9 en mars et release 10 en juin) : calibrage des données à partir de plusieurs sources (par exemple 2 EDS), analyse d'incertitude (modèle bayésien). Implémentations en cours dans EPP : permettre aux différents paramètres d'évoluer au cours du temps.

1.8 Sens et portée des différentes sources de données : une problématique de la mesure

La surveillance du VIH/SIDA, en particulier en Afrique, a connu plusieurs changements majeurs depuis le début de l'épidémie (voir Encadré 1.6 page 58). Le retour des enquêtes nationales en population générale au début des années 2000 a profondément changé l'image que l'on avait de l'épidémie sur le continent africain. Plusieurs hypothèses jusque là bien admises, telles que la représentativité des femmes enceintes, ont volé en éclats. Parmi les personnes engagées dans la lutte contre le SIDA (politiques, chercheurs, militants associatifs…), nombreuses sont celles qui ne savent plus sur quelles données se baser pour avoir une image précise de la situation actuelle et les critiques concernant les estimations des prévalences sont parfois virulentes.

Les mauvaises interprétations des chiffres ne sont pas rares. Certains ont conclu rapidement et à tort à une baisse de la prévalence du VIH dans certains pays alors que cette baisse apparente, d'un rapport mondial à l'autre, provient avant tout d'une évolution de la méthodologie employée pour réaliser ces estimations. Les chiffres ne peuvent alors être comparés directement. Certains considèrent que les biais liés aux enquêtes nationales en population générale sont tels qu'il est préférable de continuer à se baser sur la surveillance sentinelle des femmes enceintes. Les experts de l'ONUSIDA ont, quant à eux, adopté une position pragmatique, utilisant les prévalences du VIH observées chez les femmes enceintes pour estimer les tendances des épidémies et les résultats des enquêtes en population générale pour en déterminer les niveaux.

Une autre difficulté consiste à distinguer population générale et sous-groupes à risque, les seconds faisant partie de la première mais dans des proportions telles que leur poids sur l'épidémie nationale est difficile à identifier. Pendant des années, la surveillance du VIH s'est avant tout focalisée sur ces groupes les plus à risque d'être infectés. Par la suite, alors que l'épidémie de VIH se révélait être endémique sur le continent africain, la surveillance s'est étendue aux femmes enceintes, considéré comme un proxy de la population générale, essentiellement urbaines et seulement plus récemment rurales. Le passage d'une vision multiple en plusieurs sous-groupes de taille plus ou moins connue à une vision globale n'est pas aisé, les comparaisons ne pouvant être directes du fait des problèmes méthodologiques liés au changement d'échelles.

Face à différentes sources d'informations (enquêtes en population générale, surveillance sentinelle auprès de femmes enceintes et/ou de populations spécifiques, estimation par modélisation, voir également l'Encadré 1.7 page 64), il est nécessaire de mieux comprendre la portée de chacune d'elles et d'en appréhender le sens et les limites.

⁂

L'objet de cette thèse consistera donc à déterminer, pour chaque source de données, quelles sont les informations pertinentes qu'elle fournit, dans quelle mesure elle reflète une certaine réalité épidémique et, cela, à quelles conditions et/ou sous quelles hypothèses. Notre propos se limitera aux données de prévalence du VIH, essentiellement en population générale, dans des pays présentant une épidémie généralisée, en Afrique subsaharienne. Par ailleurs, nous travaillerons sur des données standardisées, en particulier les Enquêtes Démographiques et de Santé et la surveillance sentinelle des femmes enceintes, afin que nos résultats soient le plus génériques possible.

Nous nous situons résolument dans un contexte de statistique imparfaite. Comment, à partir de données parcellaires, pouvons-nous obtenir des informations sur les sociétés humaines ? Plus généralement, notre problématique sera celle de la mesure dans le domaine des sciences : quelle signification objective peut-on déduire d'une observation du monde et en quoi ce processus peut-il nous permettre d'appréhender le réel ?

Dans un premier temps, nous ferons un détour par l'épistémologie et la philosophie des sciences afin de mieux comprendre le processus intellectuel à l'œuvre dans tout acte d'observation et de se doter de critères, de concepts pour appréhender la signification objective d'un énoncé d'observation. Cela nous amènera à préciser le sens que nous donnerons au concept de domaine de validité d'un énoncé d'observation (Chapitre 2).

Puis, nous aborderons la notion, usuelle en statistique, de représentativité et, conjointement, celle de biais d'observation. Nous montrerons notamment que la surveillance sentinelle des femmes enceintes n'a qu'une représentativité limitée et locale tandis que les EDS sont conçues pour être représentatives au niveau national et régional (Chapitre 3).

Cela nous amènera à développer une approche méthodologique et cartographique afin de reconstruire, à partir des EDS, les variations spatiales de la prévalence du VIH (Chapitre 4).

De la discussion de nos résultats se dégageront les différentes « dimensions » de la mesure, que ce soit la question des échelles d'analyse, géographiques ou populationnelles, celle du niveau de notre indicateur ou encore de ses tendances (Chapitre 5). Que dire alors de la démarche actuelle de l'ONUSIDA pour la réalisation de ses estimations nationales biennales à partir du logiciel EPP ?

Encadré 1.7

Les principales sources de données épidémiologiques sur le VIH/SIDA en Afrique

Estimations et modèles ONUSIDA

http://www.unaids.org

ONUSIDA publie un rapport biennal depuis 1998 comportant des estimations de la prévalence du VIH pays par pays. Ces estimations sont réalisées par le Groupe de Travail ONUSIDA/OMS sur la Surveillance mondiale du VIH/SIDA et des IST créé en 1996, en tenant compte des recommandations du Groupe de Référence d'ONUSIDA sur les Estimations, la Modélisation et les Projections, et en collaboration avec les programmes nationaux de lutte contre le SIDA. Elles sont basées sur les différentes sources de données disponibles.

Les rapports d'ONUSIDA sont disponibles sur *http://www.unaids.org* ainsi que les logiciels utilisés pour la réalisation des estimations :

- *EPP (Estimation and Projection Package)*, pour estimer et projeter les prévalences du VIH chez les adultes à partir de données de surveillance sentinelle, pour les pays à épidémie généralisée ou concentrée ;
- *Spectrum*, un ensemble de modèles permettant, entre autres, de réaliser des projections démographique (module *DemProj*) ou d'estimer l'impact démographique du VIH/SIDA (module *AIM*) ;
- *Workbook Method*, sous la forme d'un classeur Excel, pour réaliser des estimations de la prévalence nationale à partir de données de surveillance pour les pays à épidémie concentrée ou à faible prévalence ;
- *Modes of Transmission (MoT)*, également sous la forme d'un classeur Excel, pour calculer des projections des nouvelles infections à VIH selon le mode de transmission.
- Nous mentionnons également *AEM (Asian Epidemic Model*) bien que ce dernier ne soit pas téléchargeable sur le site de l'ONUSIDA. Ce modèle a été développé pour tester l'impact de différentes politiques d'intervention en Asie.

Entre deux rapports, ONUSIDA publie des *Epidemic Update* avec des estimations globales par grande région géographique.

Les méthodologies utilisées par ONUSIDA pour ses estimations évoluent en fonction des recommandations de son Groupe de Référence sur les Estimations, la Modélisation et les Projections. Les rapports des différentes réunions de ce groupe sont disponibles sur un site dédié : *http://www.epidem.org*.

HIV/AIDS Surveillance Database

http://www.census.gov/ipc/www/hivaidsn.html

Créée en 1987 par le Center for International Research (CIR) of the United States Bureau of the Census avec le support de l'USAID, cette base rassemble l'ensemble des données de prévalence et d'incidence du VIH publiées dans des revues internationales médicales et scientifiques ou présentées lors de conférences sur le VIH/SIDA.

Pour chaque chiffre entré dans la base de données, il est précisé l'effectif de l'échantillon, la population concernée, la tranche d'âges, la région géographique, la date d'observation et la source de l'information. La base de données, disponible sous la forme d'une base Access, permet d'effectuer une recherche multicritères. Il est possible d'exporter les résultats de la requête au format Excel ou texte.

Les données sont mises à jour annuellement et la dernière révision date de décembre 2006. Des tables synthétiques ainsi que des cartes peuvent être téléchargées sur le site de la HIV/AIDS Surveillance Database. La base n'est plus téléchargeable en ligne mais peut être commandée par courriel.

Epidemiological Fact Sheets (EFS)

http://www.who.int/hiv/pub/epidemiology/pubfacts/en/

Publiés par l'OMS et l'ONUSIDA depuis 1997, les EFS sont une émanation du groupe ONUSIDA/OMS sur la Surveillance Globale du VIH/SIDA et des IST. Réactualisés régulièrement, ces Fact Sheets fournissent une image synthétique de la situation épidémique de chaque pays dans le monde et rassemblent à la fois les estimations réalisées par ONUSIDA, une synthèse des données disponibles dans la HIV/AIDS Surveillance Database, des informations sur les services de santé implémentés dans le pays, 13 indicateurs démographiques, 11 indicateurs socio-économiques, 9 indicateurs d'accès aux services de santé, 10 indicateurs sur les connaissances et les comportements et des indicateurs sur les IST.

UNAIDS/WHO Global HIV/AIDS online database

http://www.who.int/globalatlas/default.asp

La *UNAIDS/WHO Global HIV/AIDS online database* est un sous projet du *Global Health Atlas* de l'OMS. Cette base de données reprend essentiellement les données publiées dans les *Epidemiological Fact Sheets*. Il est possible de réaliser ses propres requêtes, d'exporter les résultats, mais également de réaliser des cartes interactives.

Par ailleurs, certaines cartes et tableaux synthétiques, déjà réalisés, sont directement téléchargeables.

Au moment où nous écrivons (mi 2007), il semble que la base de données n'ait pas été mise à jour et qu'elle ait été détériorée (enregistrements manquants).

Enquêtes Démographiques et de Santé (EDS) & AIDS Impact Surveys (AIS)

http://www.measuredhs.com/

L'ensemble des rapports finaux de ces enquêtes en population générale sont disponibles en ligne (voir Encadré 1.4 page 48 pour une présentation plus détaillée de ces enquêtes). Par ailleurs, les différentes bases de données peuvent être téléchargées après en avoir fait la demande en ligne.

Les principaux indicateurs calculés à partir de ces enquêtes (un peu moins de 200) peuvent être consultés à partir d'un outil appelé *STATcompiler*, disponible en anglais et en français, permettant de créer ses propres tableaux puis de les exporter au format Excel. *http://www.statcompiler.com/statcompiler/*

Les données concernant le VIH/SIDA ne sont pas accessibles directement dans *STATcompiler* mais peuvent être consultés via un outil équivalent, disponible via le même site : la *HIV/AIDS Survey Indicators Database* (voir plus bas).

Les données de *STATcompiler* peuvent être également cartographiées en ligne à l'aide de *STATmapper* : *http://statmapper.mapsherpa.com*

HIV/AIDS Survey Indicators Database

http://www.measuredhs.com/hivdata/

Mise en ligne en mai 2002, la *HIV/AIDS Survey Indicators Database* a pour vocation de réunir dans une même base de données un peu plus de 150 indicateurs comportementaux, épidémiologiques ou portant sur les connaissances relatives au VIH/SIDA. Les indicateurs retenus ont été déterminés en fonction de guides et de besoins de l'ONUSIDA, de l'Assemblée Générale des Nations unies réunie en session spéciale sur le VIH/SIDA (UNGASS), des Objectifs du Millénaire (Millennium development Goals ou MDGs), du plan présidentiel américain PEPFAR (President's Emergency Plan for AIDS Relief) et du Fond Mondial contre le VIH/SIDA, le Paludisme et la Tuberculose.

La base de données réunie les résultats d'environ 200 enquêtes menées dans plus de 75 pays. Il s'agit essentiellement d'EDS, mais également de Multiple Indicator Cluster Surveys (MICS), de Reproductive Health Surveys (RHS), de Sexual Behaviour Surveys (SBS) et de Behavioral Surveillance Surveys (BSS).

La base de données est supervisée par un comité international incluant des représentants de l'USAID (principal financeur), l'UNICEF, le CDC, l'ONUSIDA, l'OMS, le US Bureau of the Census, FHI, MEASURE Evaluation, le Synergy Project, le US Census Bureau International Programs Center et MEASURE DHS+.

Comme pour *STATcompiler*, l'utilisateur peut construire en ligne ses propres tableaux puis les exporter au format Excel.

Depuis mi-2007, des cartes au niveau national et régional des différents indicateurs peuvent être réalisées en ligne à l'aide de *HIVMapper* : http://www.hivmapper.com/.

Chapitre 2
Domaine de validité d'une observation

Sur le terrain, le chercheur quantitativiste est amené à réaliser des enquêtes, à effectuer des mesures, à produire des informations chiffrées puis à les interpréter. Tout au long de ce processus, il y a le plus souvent une part d'implicite dans la manière dont il va donner sens à une suite de chiffres alignés dans un tableau. À partir d'observations forcément fragmentaires (une observation parfaitement exhaustive d'un phénomène relevant de l'impossible), il va en déduire une connaissance plus générale sur le monde.

Cependant, il est possible de basculer facilement vers une interprétation abusive et de produire ainsi un sens non contenu dans les données observées et ne correspondant pas à une réalité du phénomène étudié.

L'objectif du présent chapitre, au travers d'un détour théorique par l'épistémologie des sciences, vise à mieux comprendre les processus intellectuels à l'œuvre dans l'acte d'observation et d'interprétation. Il s'agira de nous doter de critères et de concepts pour préciser la signification objective que l'on peut attribuer à une mesure quantitative donnée, critères que nous pourrons réutiliser par la suite afin de comprendre et d'expliciter les divergences apparentes entre les différentes mesures de la prévalence du VIH en Afrique subsaharienne.

Nous limiterons dans ce chapitre notre propos aux observations dites numériques, c'est-à-dire aux actes d'observations du monde se traduisant par la construction de

grandeurs chiffrées. Nous tenons à préciser ici qu'il ne s'agit aucunement d'un déni des apports des méthodes dites qualitatives. Bien au contraire, nous croyons fermement que les approches quantitatives et qualitatives sont non seulement complémentaires, mais également nécessaires l'une à l'autre pour essayer de démêler la complexité du monde social qui nous environne. Cependant, afin de ne pas trop nous éloigner de notre sujet, à savoir la mesure des prévalences du VIH, il nous semble opportun de limiter notre discours, bien que certains des aspects abordés ci-après dans le présent chapitre concernent tout aussi bien les observations qualitatives que les observations quantitatives.

2.1 À la recherche d'une signification objective de la mesure

Les phénomènes qui nous préoccupent, et plus largement l'ensemble des phénomènes des sciences sociales et d'autres sciences, présentent deux caractéristiques majeures. D'une part, ils s'inscrivent dans l'espace-temps physique et se manifestent sous une forme *régionalisée*. Ils se sont produits pendant un laps de temps donné, à une époque donnée, dans une zone donnée. D'autre part, il s'agit de phénomènes *uniques* en ce sens qu'ils ne se réalisent qu'une seule et unique fois dans des conditions parfaitement identiques. Nous nous situons en dehors du champ expérimental d'un laboratoire où il serait possible de répéter à loisir un même phénomène dans des circonstances *a priori* identiques.

Face à ce type de phénomènes, le scientifique est amené à estimer certaines grandeurs dont il n'a pas connaissance mais qui existent néanmoins indépendamment de lui. Ne pouvant procéder à un recensement exhaustif parfait de l'ensemble du phénomène et des propriétés du contexte dans lequel ce dernier intervient, il a alors recours à des méthodologies, des modèles pour tenter de répondre aux questions qu'il se pose. Fréquemment, le scientifique utilise la statistique et les modèles aléatoires, et implicitement fait appel à la *Loi des Grands Nombres*. Cette dernière, formalisée au XVIII[e] siècle lors de la découverte de nouveaux langages mathématiques[1], indique que lorsque l'on fait un tirage aléatoire dans une série de grande taille, plus on augmente la taille de l'échantillon,

[1] La Loi des Grands Nombres a été tout d'abord formalisée par Jacob BERNOUILLI en 1713 dans un ouvrage intitulé *Ars Conjectandi: Usum & Applicationem Praecedentis*. Bernouilli l'avait alors nommée « théorème d'or ». En 1835, Siméon Denis POISSON la décrit comme « Loi des Grands Nombres », nom sous laquelle elle est aujourd'hui la plus connue. Les mathématiciens distinguent aujourd'hui deux énoncés : la *loi forte des grands nombres* et la *loi faible des grands nombres*.

plus les caractéristiques statistiques du tirage se rapprochent des caractéristiques statistiques de la population.

La plupart des sondages et des enquêtes repose sur cette loi. En interrogeant un nombre suffisant de personnes il devient alors possible de connaître ce qu'il en est (probablement) pour la population entière. C'est également grâce à cette loi qu'a pu se développer le système assurantiel en permettant de déterminer la probabilité des sinistres dont les assureurs se portent garants.

Cette loi porte par ailleurs en elle une sorte de paradoxe. Si un évènement considéré isolément semble soumis au hasard (obtenir 1 000 fois « pile » lors de 1 000 lancers d'une pièce de monnaie n'est pas impossible), cette loi postule l'absence de hasard global (sur 1 000 lancers, on observe en réalité un nombre de « pile » proche de 500), comme s'il existait une loi d'équilibre naturel, comme si le chaos était impossible et les catastrophes improbables.

Le recours à la statistique et aux modèles probabilistes a fait couler beaucoup d'encre et en particulier la notion de probabilité entre *objectivistes* (ou *fréquentistes*) et *subjectivistes*. Une vision objectiviste considère la probabilité d'un évènement comme la fréquence à laquelle ce dernier survient lorsque le nombre de répétitions tend vers l'infini. « *Parler de la probabilité pour qu'une proposition donnée soit vraie n'a aucun sens pour un objectiviste, car on ne peut parler de la probabilité d'un évènement par nature unique : la proposition est vraie ou fausse, en logique aristotélicienne.* » (COURGEAU 2004, p. 29) Dans le cadre d'une approche subjectiviste, « *la probabilité ε a la signification d'une évaluation personnelle (subjective) du "degré de croyance rationnelle" (selon une formule due à Keynes) dans la réalisation de l'évènement auquel elle s'applique* » (BONITZER 1984, p. 14).

Le débat entre ces deux approches est loin d'avoir été toujours fécond et la faiblesse commune au fréquentisme et au subjectivisme tient entre autres à une focalisation excessive sur le seul concept de probabilité d'un évènement particulier (BONITZER 1984). Par ailleurs, la distinction entre subjectivisme et objectivisme recoupe d'autres points de divergence, que ce soit en statistique, en épistémologie ou en sciences sociales : relativisme contre réalisme, positivisme ou anti-fondamentalisme, sociologie compréhensive face à sociologie explicative, DURKHEIM ou WEBER, approche qualitative ou approche quantitative…

Il nous semble peu opportun d'aller plus avant sur ces questions pour le moment ni même de nous positionner de suite. Nous risquerions de nous perdre dans des considérations purement théoriques et stériles d'un point de vue pratique ou opérationnel. Néanmoins, nous y reviendrons brièvement à la fin de ce chapitre.

Dans son *Essai sur la pratique des probabilités*[2], Georges MATHERON pose cette problématique de manière différente :

> « *Je ne peux donc, en aucune façon, éluder la fameuse question, si mal formulée : "la probabilité est-elle subjective ou objective ?". En fait, il n'y a pas, il ne saurait y avoir, de probabilité en soi. Il n'y a que des modèles probabilistes. Ou, si l'on préfère, l'"aléatoire" n'est en aucune façon une propriété, univoquement définie, ni même définissable, du phénomène lui-même. Mais, uniquement, une caractéristique du, ou des, modèle(s) que nous choisissons pour le décrire, l'interpréter, et résoudre tel ou tel problème que nous nous posons à son sujet. Et, suivant la nature de ces problèmes, nous pouvons fort bien adopter, tout à tour, des modèles différents, attribuant selon le contexte, des probabilités différentes à un évènement donné dont l'énoncé, apparemment reste le même. Le seul problème réel est de savoir si un modèle donné, dans un contexte donné, possède, ou non, une signification objective et, le cas échéant, de faire le tri : je veux dire, parmi les concepts, énoncés, paramètres, etc. figurant dans le modèle, distinguer ceux qui possèdent dans (ce que nous appelons) la réalité une contrepartie objective, observable, mesurable, etc. et les autres : ces derniers, que j'appellerai conventionnels (plutôt que subjectifs) pourront jouer un rôle heuristique utile dans l'élaboration et la mise en œuvre du modèle. Mais ils devront disparaître de la formulation ultime de nos conclusions et de nos résultats. Car, dans la mesure même où les problèmes que nous cherchons à résoudre sont des problèmes réels, les solutions que nous proposons doivent, elles aussi, sous peine de se révéler illusoires, respecter le principe de réalité. […]*
>
> *Cette position méthodologique se reflète dans la terminologie que j'adopte. Les statisticiens orthodoxes [ou objectivistes] auraient dit : estimer et prédire. Les bayesiens, ou subjectivistes : évaluer et prévoir. J'ai préféré : estimer et choisir. […] En somme, on estime des grandeurs (objectives), on choisit des méthodes, des modèles ou des paramètres conventionnels et on convient de critères. Pour les bayesiens, il n'y a pas d'estimation, mais seulement des "choix". Ils évaluent des probabilités (subjectives) et en tirent des prévisions (également subjectives) concernant les évènements et les grandeurs inconnues. Les orthodoxes, ou objectivistes, à l'inverse, réservent le terme "estimation" au choix des paramètres de leurs modèles, sans se demander au préalable si ces paramètres possèdent ou non, une contrepartie objective dans la réalité.* » (MATHERON 1978, p. 2-4)

[2] MATHERON G., *Estimer et Choisir : essai sur la pratique des probabilités*, Fontainebleau (FR), Centre de Géostatistique, École Nationale Supérieure des Mines de Paris, 1978.

MATHERON nous invite ainsi à nous interroger sur la signification objective de nos mesures et de nos estimations, à savoir dans quelle mesure ces dernières reflètent ou non un certain aspect du monde réel. Vis-à-vis des modèles que l'on utilise, il distingue *objectivité externe* et *objectivité interne*. Certaines méthodologies ont permis de traiter et de résoudre nombre de problèmes et ont connu ainsi des succès réels confirmés par la pratique. Ayant fait leurs preuves, il est possible de considérer que ces techniques et méthodes disposent d'une certaine objectivité que MATHERON nomme *objectivité externe*, c'est-à-dire la « *sanction de la pratique* ». Mais les objets étudiés restant uniques (il n'existe pas deux sociétés parfaitement identiques, deux territoires identiques, deux populations identiques en tout point), MATHERON se demande « *dans quelle mesure une estimation ou un modèle probabiliste concernant ce gisement-ci, cette forêt-là (et non un gisement ou une forêt en général) possède-t-il une signification objective ?* » (MATHERON 1978, p. 2). C'est cette signification qu'il nomme *objectivité interne*.

Si MATHERON s'interroge essentiellement sur l'utilisation de modèles probabilistes, nous pouvons étendre son interrogation aux opérations de mesures d'un phénomène, étant donné que, comme nous le verrons plus loin, toute opération visant à quantifier le réel nécessite le passage par un modèle plus ou moins implicite. Nous nous interrogerons donc sur l'objectivité interne de nos mesures ou, exprimé autrement, sur la *signification objective* que nous pouvons accorder à celles-ci.

2.2 Le critère poppérien d'objectivité : la falsifiabilité

Si l'on cherche à déterminer la signification objective d'un énoncé, encore faut-il pouvoir définir ce que nous pouvons appeler *objectivité*. Pour cela, nous pouvons nous référer à ces mots de SÉNÈQUE LE JEUNE régulièrement cités par différents auteurs :

> « *La différence entre nous et les Toscans, les plus habiles interprètes des tonnerres, c'est que, selon nous, l'explosion de la foudre a lieu par suite de la collision des nuages, et que, suivant eux, la collision n'a lieu que pour amener l'explosion. En effet, comme ils rapportent tout à Dieu, ils croient non pas que les foudres annoncent l'avenir parce qu'elles sont formées, mais qu'elles sont formées parce qu'elles doivent annoncer l'avenir. Néanmoins elles se produisent de la même manière, que leur pronostic soit ou la cause ou l'effet de leur formation.* »
>
> (SÉNÈQUE vers 62, Questions Naturelles II, 32-2)

En l'occurrence, l'argument des Toscans est imparable. En expliquant la foudre comme étant la résultante d'une volonté divine, quelque soit le phénomène

observé, il sera toujours possible de répondre que ce dernier correspondait au choix divin. Mais dans le même temps, cette explication ne permet pas de prévoir quoi que ce soit. Il est toujours possible, *a posteriori*, de fournir une explication *ad hoc*. Par contre, *a priori*, une observation du ciel ne permet pas de déterminer si la foudre tombera ou non, la volonté divine restant hors de portée. La position de SÉNÈQUE, à savoir que la foudre naît de la collision entre nuages, implique au contraire que toute observation de foudre doit être concomitante avec un ciel chargé. Elle présuppose donc qu'en présence d'un ciel dégagé, il ne peut y avoir de foudre. Elle permet donc de réaliser une prédiction *a priori*, celle-ci pouvant ou non se vérifier dans les faits.

Nous pouvons alors considérer comme *purement subjective* (c'est-à-dire dépourvue d'objectivité) une opinion dès lors qu'elle sera compatible avec toute chose et son contraire. À l'inverse, elle ne permettra pas de prévoir un phénomène, quel qu'il soit. Cela revient à dire qu'elle n'apporte aucune information réelle. L'objectivité d'un énoncé est alors liée à la possibilité que nous avons d'en contrôler l'exactitude ou, à tout le moins, de procéder à une tentative de vérification empirique. Selon Karl POPPER, « *l'objectivité des énoncés scientifiques réside dans le fait qu'ils peuvent être intersubjectivement soumis à des test*.[3] » Il précise alors ce qu'il entend par « tests intersubjectifs ». Il introduit alors la notion de *régularité* qui, par la reproduction d'observations identiques, permet d'envisager des tests pour vérifier la validité des dites observations.

> *« Seules de telles répétitions peuvent nous convaincre que nous n'avons pas à faire à une simple "coïncidence" isolée mais à des évènements qui, en raison de leur régularité et de la possibilité qu'ils ont d'être reproduits, peuvent en principe être soumis à des tests intersubjectifs. »*[4]

Précision importante, POPPER n'exige pas qu'un énoncé scientifique *ait été soumis* à des tests intersubjectifs, mais que ce dernier *puisse* l'être. C'est la possibilité de pouvoir être testé intersubjectivement qui confère une objectivité à un énoncé. Cela ne signifie pas pour autant que cet énoncé soit juste ou véridique, mais cela permet d'envisager la possibilité d'une vérification empirique. Un énoncé ne pouvant être soumis à des tests intersubjectifs ne concernera donc pas le scientifique. Cela n'implique pas que cet énoncé soit faux mais, dans la mesure où l'on ne pourra ni confirmer ni infirmer cet énoncé, il sort du champ de la recherche scientifique.

---

[3] POPPER K. R., *La Logique de la découverte scientifique*, Paris (FR), éditions Payot, 1968, p. 41. Chapitre premier, section 8.

[4] *Idem*, p. 42.

POPPER distingue deux types d'énoncés pour lesquels la question de l'objectivité se pose différemment : les *énoncés universels* et les *énoncés singuliers*[5].

Figure 2.1
Déploiement du Quilt à Washington en 1992

Source : Plaquette du *Patchwork des Noms*, photographie de Jean FOREST.

Un énoncé singulier décrit « *une occurrence*[6] ». Il s'agit d'énoncé du type « *constatation d'un fait*[7] ». Un tel énoncé décrit un évènement particulier, en un lieu donné, à une date donnée. Par exemple, "*le 21 octobre 1992, 22 000 panneaux*

5 Parfois appelés *énoncés de base* ou *énoncés existentiels*.

6 POPPER, *La Logique de la découverte scientifique*, p. 86. Chapitre 4, section 23.

7 MATHERON, *Estimer et Choisir*, p. 32.

du Quilt[8] *sont déployés en cercle autour du Mémorial de Washington (USA)*". Il est possible, selon POPPER, d'arriver à un consensus concernant les énoncés de base pour décider de les accepter ou de les réfuter. « *Le critère d'objectivité réside dans le fait qu'une fois réunie toute la documentation nécessaire, un consensus se réalise entre les « gens sensés » concernant la vérité ou la fausseté de l'énoncé en question : il s'agit d'énoncés décidables, c'est-à-dire dont il serait univoquement possible de décider s'ils sont vrais ou faux, pourvu seulement que l'on dispose des informations voulues.*[9] » Un énoncé *décidable* peut rester *indécidé* si l'on ne peut avoir accès à des sources permettant d'arriver à un consensus. Par exemple, on ne pourra trancher un énoncé tel que "*il pleuvait à Lutèce le 14 juillet de l'an 52 avant Jésus-Christ*". Pour autant, cet énoncé garde son objectivité dans la mesure où, si nous avions à disposition la documentation adéquate, nous pourrions arriver à un consensus quand à sa vérité ou sa fausseté. Un énoncé objectif peut très bien être faux. Il suffit qu'il soit possible de le déclarer faux. Par ailleurs, cette définition des énoncés de base revient :

> *« à s'arrêter à des énoncés sur l'acceptation ou le rejet desquels les divers chercheurs peuvent s'entendre. Et s'ils ne s'entendent pas, ils poursuivront tout simplement leurs tests ou les recommenceront tous. S'ils n'obtiennent pas plus de résultat de cette manière, nous pourrions alors dire que les énoncés en question ne pouvaient pas être soumis à des tests intersubjectifs ou, qu'après tout, ils ne traitaient pas d'évènements observables. S'il devait un jour n'être plus possible pour les observateurs scientifiques de s'entendre au sujet des énoncés de base, cela équivaudrait à l'échec du langage comme moyen de communication universel.* » (POPPER 1968, p. 104, Chapitre 5, section 29)

POPPER distingue ensuite des énoncés qu'il qualifie d'*énoncés universels*. Ce type d'énoncés recoupe généralement ce que l'on nomme *énoncés théoriques* ou *lois naturelles*. Ils portent sur une classe illimitée d'objets ou d'évènements et sont valables en tous temps et en tous lieux. Un énoncé universel (par exemple : "*deux corps s'attirent en raison du carré inverse de leur distance*"[10]) affirme que tous les éléments d'une classe (en l'occurrence tous les objets ayant une masse) vérifient une certaine propriété. On ne peut vérifier de tels énoncés, c'est-à-dire montrer

[8] Le Quilt est né à San Francisco en 1987 sous l'impulsion de Cleve JONES. Il est aujourd'hui maintenu dans le monde par le *Names Project* et s'est traduit en France avec la création en 1989 du *Patchwork des Noms*. « *C'est une mosaïque immense constituée de panneaux cousus où les gens ont exprimé quelque chose d'important sur celle ou celui qu'ils ont perdu.* » Jonathan MANN, *L'Autre Journal*, oct. 1990, cité par le Patchwork des Noms.

[9] MATHERON, *Estimer et Choisir*, p. 32-33.

[10] Version simplifiée de la Loi Universelle de la Gravitation, formulée pour la première fois en 1684 par Isaac NEWTON.

qu'ils sont vrais dans la mesure où il est impossible de procéder à un inventaire exhaustif de l'Univers (nous ne pouvons prouver matériellement que tous les corps de l'Univers s'attirent effectivement en raison inverse du carré de leur distance). Par contre, il est possible de montrer que de tels énoncés sont faux. Il suffit pour cela d'une seule observation ne vérifiant pas la propriété posée par l'énoncé. Par exemple, si nous prenons l'énoncé "*tous les chats sont noirs*", nous ne pouvons montrer que cet énoncé est juste puisque nous ne pouvons vérifier la couleur de tous les chats passés, présents et à venir. Par contre, il nous suffit de trouver un seul chat blanc ou roux pour affirmer que l'énoncé "*tous les chats sont noirs*" est faux. C'est justement cette possibilité de pouvoir être *falsifié* qui déterminera l'objectivité d'un énoncé universel.

> « *En matière scientifique, s'agissant d'un énoncé, d'un modèle, d'une théorie, etc. (considérés en tant qu'ils se rapportent à un secteur bien défini du monde que nous appelons réel) nous disons qu'ils ont une signification objective si, et dans la mesure où, il est possible de les soumettre au contrôle d'expériences ou d'observations dont le résultat soit définissable sans équivoque (ce qui veut dire que l'énoncé de ces résultats doit être susceptible de réaliser le consensus des spécialistes). Le plus souvent, un énoncé scientifique est du genre universel, et ne peut par suite faire l'objet d'une vérification (logique) rigoureuse, qui impliquerait la réalisation effective d'une infinité d'observations ou d'expériences. Par contre, il doit toujours être possible de déduire d'un énoncé scientifique général des énoncés plus particuliers (prédiction du résultat d'expériences ou d'observations effectuées dans des conditions bien définies) susceptibles, eux, d'être confirmés ou infirmés (vérifiés ou falsifiés). S'ils sont confirmés, il n'en résulte pas que l'énoncé général soit vérifié, mais seulement qu'il est "corroboré" (non réfuté). Par contre, si l'un d'eux est infirmé, l'énoncé général est par la même falsifié (réfuté). Autrement dit, l'énoncé général a une signification objective dans la mesure où il est falsifiable, et possède une validité (relative et toujours provisoire) dans la mesure où il a été corroboré, c'est-à-dire a résisté victorieusement à toutes les tentatives de falsification auxquelles il a été soumis jusqu'à présent : et nous lui attribuerons un degré de validité d'autant plus élevé que ces tentatives auront été plus nombreuses et plus sévères. Tel est le critère de falsifiabilité que propose K. R. POPPER comme ligne de démarcation entre énoncés "métaphysiques" et énoncés "empiriques" ou objectifs.* » (MATHERON 1978, p. 33-34)

Nous pouvons donc poser le postulat[11] suivant :

> **Postulat 2.1**
> *Un énoncé aura une signification objective s'il est potentiellement falsifiable.*

De nombreuses critiques ont été formulées à l'encontre du critère de falsifiabilité et plus généralement à l'encontre du *falsificationisme*. Nous ne les détaillerons pas toutes ici mais en retiendrons trois principales : l'incapacité de ce critère à rendre compte de l'histoire des sciences et des théories scientifiques, la complexité des situations réelles de tests et enfin la faillibilité des énoncés d'observations.

Si l'on étudie l'histoire des théories scientifiques, et en particulier l'histoire de la physique, les exemples sont nombreux de situations où une théorie a été maintenue alors qu'elle s'avérait contradictoire avec des observations. Par exemple :

> « *Dans les années qui suivirent sa formulation, la théorie de la gravitation de NEWTON fut falsifiée par des observations de l'orbite de la Lune. Cinquante ans environ s'écoulèrent avant que l'on écarte cette falsification en la mettant au compte d'autres facteurs que de la théorie newtonienne. Plus tard, cette même théorie se révéla en désaccord avec les valeurs précises trouvées pour la trajectoire de la planète Mercure et les savants ne l'abandonnèrent pas pour autant. Pourtant on ne parvint jamais à expliquer cette falsification d'une façon qui aurait préservée la théorie de NEWTON.* » (CHALMERS 1976, p. 116)

Dans une logique falsificationiste naïve, si un énoncé de base s'avère contradictoire avec ce que prédit les énoncés universels constituant une théorie, la dite théorie se retrouvant réfutée devrait être rejetée. Or, nombre de découvertes scientifiques importantes n'ont pas suivi cette règle, fort heureusement. Si cette critique est pertinente dans une optique historienne, visant à rendre compte de la science en tant que phénomène social inscrit dans un contexte historique, il n'en est pas de même dans le cadre d'une réflexion épistémologique sur les meilleures manières de produire du savoir. Lorsque POPPER pose le critère de falsifiabilité, il ne s'agit pas d'un critère historique mais d'un critère politique ou philosophique sur ce que doit

[11] Nous avons préféré utiliser dans ce chapitre le terme de *postulat* plutôt que les mots *axiome*, *théorème* ou *loi*. Un théorème ou une loi est une assertion rigoureusement démontrée à partir d'axiomes et/ou de postulats. Un axiome est une vérité évidente en soi sur laquelle une connaissance peut se reposer. Un postulat est un principe utilisé dans la construction d'un système déductif, mais qu'on ne démontre pas lui-même, sans pour autant s'interdire la possibilité de s'y essayer plus tard. En ce sens, le postulat se distingue de l'axiome, ce dernier étant toujours posé au départ comme un élément fondamental qu'on ne cherchera pas à démontrer. Le postulat n'est pas forcément *évident*, contrairement à l'axiome. Un postulat sera donc considéré comme légitime et temporairement accepté, bien que non démontré.

être la science, selon lui, en tant que forme particulière d'acquisition de connaissances.

L'autre grande critique adressée à POPPER porte sur la complexité des situations réelles de tests. En effet, une théorie scientifique est composée de tout un ensemble, plus ou moins complexe, d'énoncés universels et d'assertions. Pour tester expérimentalement une théorie, il est nécessaire d'avoir également recours à d'autres énoncés ou postulats, telles que les lois et théories employées pour l'élaboration et l'utilisation des différents outils d'observation utilisés, ainsi que divers paramètres tels que les conditions initiales ou la description du descriptif expérimental. La prédiction qui sera testée découlera de l'ensemble de ces prémisses. Ainsi, si le test s'avère non concluant, cela ne signifie pas pour autant que la théorie testée est réfutée. Cela implique simplement que l'un des prémisses entrant dans le cadre de cette expérience est faux. Encore faut-il pouvoir déterminer lequel.

> « *Un [...] exemple nous est fourni par l'argument de l'astronome danois Tycho BRAHÉ qui affirmait avoir réfuté la théorie copernicienne quelques dizaines d'années après sa publication. Si la Terre tourne en orbite autour du Soleil, disait BRAHÉ, alors la direction dans laquelle on observe une étoile fixe à partir de la Terre doit varier au cours de l'année pendant que la Terre se déplace d'une face du Soleil à une autre. Mais les tentatives de BRAHÉ de détecter cette parallaxe prévue au moyen de ses instruments, les plus sensibles et les plus précis qui existaient à l'époque, se soldèrent par un échec. BRAHÉ fut ainsi amené à conclure que la théorie copernicienne était fausse. Avec le recul, on s'aperçoit que la prédiction erronée provient non pas de la théorie de COPERNIC, mais de l'une des hypothèses auxiliaires de BRAHÉ. Son estimation de l'ordre de grandeur de la distance des étoiles fixes était bien trop sous-évaluée. Lorsqu'on lui substitua une valeur plus réaliste, on se rendit compte que la parallaxe prévue était beaucoup trop faible pour avoir pu être détectée par les instruments de BRAHÉ.* » (CHALMERS 1976, p. 113-114)

Autrement dit, pour pouvoir réfuter une théorie, il est nécessaire de pouvoir montrer qu'il s'agit bien des énoncés posés par la théorie qui induisent une prédiction erronée. Par ailleurs, une théorie étant composée de tout un ensemble d'énoncés, certains étant plus fondamentaux que d'autres, il se peut que l'expérience amène à réfuter un énoncé secondaire particulier sans remettre en cause les énoncés centraux. À ce moment-là, ce n'est pas l'ensemble de la théorie qui doit être rejetée, une amélioration de cette dernière pouvant éventuellement amener à une prédiction correcte sans modification des paradigmes fondamentaux.

Nous retrouvons ici la troisième critique majeure adressée au falsificationisme : la faillibilité des énoncés d'observations. Outre les problèmes liés à la subjectivité des individus dans l'acte d'observation, de nombreuses expériences ayant montré à

quel point nos perceptions peuvent être trompeuses (*cf.* les illusions optiques par exemple), toute observation doit être précédée d'une théorie pour pouvoir être formulée de manière précise. Un cadre conceptuel est nécessaire afin de communiquer ce que l'on a observé.

> « *Les énoncés d'observations seront toujours formulés dans le langage d'une théorie et seront aussi précis que le cadre théorique ou conceptuel qu'ils utilisent. Le concept de "force" utilisé en physique est précis parce qu'il acquiert sa signification de par le rôle qu'il joue dans une théorie précise, relativement autonome, la mécanique newtonienne. L'utilisation du même mot dans le langage de tous les jours (la force des circonstances, les vents de force 8, la force de l'argumentation, etc.) est imprécise seulement parce que les théories correspondantes sont fort variées et imprécises. Des théories précises, clairement formulées, sont une condition préalable pour que des énoncés d'observation soient précis. En ce sens, la théorie précède l'observation.* »
>
> (CHALMERS 1976, p. 61)

Les écrits de POPPER, FEYERBEND et KUHN abondent en arguments et en exemples de ce que la théorie précède l'observation. Les théories scientifiques étant faillibles, il en résulte que les énoncés d'observations peuvent également être faillibles (voir l'exemple cité précédemment de Tycho BRAHÉ à propos de la parallaxe des étoiles fixes). Ces différents exemples nous montrent qu'une application naïve du falsificationisme peut amener à des erreurs. Quel que soit le résultat d'une expérience ou d'une observation, un examen critique rigoureux est nécessaire et il est parfois nécessaire de repousser la décision de rejeter ou d'accepter les résultats obtenus. Cependant, cette critique ne porte pas sur le critère de falsifiabilité d'un énoncé en tant que définition de l'objectivité d'un énoncé mais sur le processus visant à déterminer à partir de quel moment une théorie doit être considérée comme falsifiée.

Si, personnellement, nous serons prudent sur la question du rejet d'une théorie ou d'un ensemble d'énoncés dès lors qu'un test de falsification aura été positif, nous retiendrons par contre de POPPER la nécessité qu'un énoncé soit falsifiable, c'est-à-dire qu'il soit possible d'imaginer un test intersubjectif permettant, *a priori*, de le falsifier. Nous posons ainsi la falsifiabilité comme une exigence permettant de limiter les inférences abusives et les interprétations *ad hoc*.

2.3 À la quête de vérité : le philosophe et l'ingénieur

Plusieurs buts ont été assignés ou attribués à la philosophie au cours des âges. Pour certains, la philosophie s'apparente à la quête de la *vérité*[12], le philosophe recherchant à percer la nature véridique du monde qui nous entoure. La vérité a été posée très tôt comme étant l'objet de la science. Pour ARISTOTE, « *l'objet de la science au sens propre est quelque chose qui ne peut pas être autre qu'il n'est*. »[13] Autrement dit, la science porte sur l'essence même du monde qui nous entoure, au-delà de nos perceptions sensitives qui peuvent être trompeuses. Par ailleurs, le propos de la science se doit d'être universel, valable en tous lieux et en tous temps : « *la science consiste dans la connaissance universelle*[14] ». La recherche de lois universelles a marqué la progression du savoir scientifique et notamment la physique jusqu'au début du XX^{e} siècle. Dans cette optique, toute théorie qui aurait été réfutée par quelque expérience que ce soit (moyennant les réserves évoquées à la section précédente) doit être rejetée : une telle théorie ne rendrait pas compte de l'ensemble des phénomènes observables et par là-même serait erronée quant à l'essence de ce qui est.

Un des exemples de réfutation le plus souvent cité concernant la pensée poppérienne est la réfutation de la mécanique classique newtonienne, remplacée par la relativité restreinte développée en 1905 par Albert EINSTEIN et la relativité générale publiée en 1915. À la fin du XVII^{e} siècle, Isaac NEWTON formule la loi universelle de la gravitation qui stipule que deux corps s'attirent mutuellement proportionnellement à leur masse et en raison du carré inverse de leur distance. Cette avancée majeure va se situer à l'origine de la mécanique dite classique et va être vérifiée empiriquement à de nombreuses reprises jusqu'à la fin du XIX^{e} siècle. La théorie newtonienne va être ainsi en capacité de prédire avec précision les déplacements des astres dans le ciel, les trajectoires d'objets divers tels que des obus ou bien encore de prévoir les éclipses du Soleil et de la Lune. Les principes de la physique classique vont être universellement acceptés jusqu'à la fin du XIX^{e} siècle. Cependant, certaines observations étaient en contradiction avec la mécanique classique, cette dernière étant dans l'incapacité d'expliquer par exemple les variations de la trajectoire de la planète Mercure. L'orbite de Mercure connaît

[12] Nous employons ici le terme de *vérité* dans un sens volontairement naïf : est vérité une proposition vraie, c'est-à-dire adéquate avec ce qui est.

[13] ARISTOTE, *Les Seconds Analytiques - Organon IV*, Wikisource, nouvelle traduction pour Internet par sœur Pascale NAU, à partir de la version grecque, de la traduction Vrin et de celle de G. R. G. Mure. Livre I, Chapitre 2.

[14] ARISTOTE, *Les Seconds Analytiques*. Livre I, Chapitre 31.

une très lente précession du périhélie[15] autour du Soleil. En d'autres termes, son orbite est elle-même en rotation autour du Soleil. Toutes les planètes connaissent une précession, causée par l'influence gravitationnelle des autres corps du système solaire, et celle-ci s'explique par la mécanique newtonienne pour chacune d'elle, sauf Mercure. En effet, Mercure connaît une précession légèrement plus rapide que celle à laquelle on peut s'attendre en appliquant les lois de la mécanique céleste, et se trouve en avance d'environ 43 secondes d'arc par siècle. Ce phénomène constitue ainsi une réfutation expérimentale de la théorie newtonienne. La théorie de la relativité générale d'EINSTEIN permet quant à elle de décrire avec précision ce phénomène. Cette nouvelle théorie modifie la structure de l'espace et du temps tels qu'ils étaient pensés jusqu'alors. Alors que selon NEWTON l'Univers est euclidien (lisse et plat), la théorie relativiste institue un espace-temps à quatre dimensions (trois dimensions spatiales et une dimension temporelle) dans lequel tous les évènements de l'Univers s'inscrivent. Ce continuum espace-temps est courbe à proximité des objets massifs[16] ce qui implique que le chemin le plus court entre deux points n'est pas toujours une droite. Ainsi, la trajectoire d'un rayon lumineux sera courbe, selon cette théorie, à proximité d'un objet lourd tel que le Soleil. Ce phénomène, prédit par EINSTEIN, fut vérifié expérimentalement le 29 mai 1919, par des observations réalisées par Sir Arthur EDDINGTON lors d'une éclipse de Soleil. Cela constitua la première grande vérification expérimentale de la théorie relativiste. La relativité générale s'avéra ainsi à la fois capable d'expliciter les mêmes phénomènes que ceux décrits par la physique newtonienne mais également de prédire, avec justesse, des phénomènes dont la mécanique newtonienne était incapable de rendre compte. POPPER écrivit que « *la théorie newtonienne a été réfutée par des expériences cruciales qui n'ont pu réfuter la théorie d'EINSTEIN*[17] ».

Nous pouvons alors nous demander si la théorie newtonienne est caduque. Il est clair qu'EINSTEIN a profondément bousculé les conceptions de l'espace et du temps et qu'il n'est plus possible aujourd'hui de considérer que nous vivons dans un univers plat. Cependant, dans notre vie quotidienne, les implications de la relativité ne sont pas perceptibles. En effet, les effets relativistes n'ont une ampleur importante qu'à des vitesses proches de celle de la lumière. Par ailleurs, les équations relativistes sont beaucoup plus complexes à manipuler que les équations classiques. En d'autres termes, pour envoyer un satellite dans l'espace ou pour calculer la trajectoire d'un missile balistique, les ingénieurs ont toujours recours à

[15] Le périhélie d'une planète est le point de son orbite le plus proche du Soleil. La précession est le nom donné au changement graduel de l'axe de rotation d'un objet, en l'occurrence le changement de l'orientation du grand axe de l'orbite de la planète.

[16] Et notamment à proximité du Soleil. Mercure étant la planète la plus proche de ce dernier, la précession de son orbite est plus affectée que celle des autres planètes.

[17] POPPER K. R., *Conjectures et Réfutations : la croissance du savoir scientifique*, Paris (FR), Payot, 1963, p. 172.

la théorie de NEWTON. Il en découle que si la mécanique classique est fausse quant aux conceptions de la matière, de l'espace et du temps qu'elle implique, elle n'en conserve pas moins une certaine part de vérité dans des contextes particuliers.

> « *Si on limite les équations de la théorie de la Relativité Générale au cas où les champs de gravitation doivent être considérés comme faibles et où toutes les masses se déplacent, par rapport au système de coordonnées, avec des vitesses qui sont petites comparées à celle de la lumière, on obtient tout d'abord la théorie de NEWTON comme première approximation.* » (EINSTEIN 1917, p. 14)

Albert EINSTEIN introduit ici un concept intéressant, celui de *validité limitée* d'une théorie qui pourrait s'exprimer également sous la forme du *domaine d'application* de cette dernière. Le terme « vrai » désigne, pour EINSTEIN, « *la concordance avec un objet réel* » (EINSTEIN 1917, p. 2). Nous pourrions qualifier cela comme la « posture de l'ingénieur », dans la mesure où cette approche de la notion de vérité s'apparente à une démarche fonctionnelle. D'un point de vue pratique, le physicien aura recours à la mécanique classique dès lors que cette dernière sera suffisamment précise pour l'usage qui en sera fait, c'est-à-dire dans un contexte où les déformations de l'espace-temps pourront être négligées. Nous voyons poindre ici deux idées importantes sur lesquelles nous reviendrons par la suite : la *précision* d'un énoncé, d'une part, et la notion d'*effets négligeables*, d'autre part.

Si POPPER ne parle pas de validité limitée, il existe dans sa pensée le concept de *vérité approchée*, la vérité étant chez lui toujours approchée et jamais définitivement atteinte.

> « *Tout en déclarant qu'EINSTEIN réfute NEWTON et que la théorie newtonienne peut être considérée comme "réfutée"*[18]*, c'est-à-dire "son système d'idées et le système formel qui en découle", POPPER affirme, par ailleurs, qu'il est "possible d'admettre, comme partie intégrante du savoir constitué, la vérité approchée, dans certaines limites, de ses formules quantitatives".* » (KREMER-MARIETTI 2005)

En fait, quelle que soit la théorie envisagée, il existe toujours une limitation à celle-ci. Et la théorie de la relativité générale n'échappe pas à cette règle, notamment du fait de son incompatibilité avec la mécanique quantique. Il s'agit de la seconde grande théorie physique du XX^{e} siècle. Développée dans les années 1910-1930 par quelques physiciens visionnaires tels que Max PLANCK, Niels BOHR, Werner HEISENBERG, Wolfang PAULI ou Louis de BROGLIE, elle rend compte du comportement des particules élémentaires et des phénomènes subatomiques.

[18] Les citations de POPPER sont extraites de POPPER, *Conjectures et Réfutations*, p. 354.

> « *Ces deux grandes théories, vérifiées à maintes reprises par de nombreuses mesures et observations, fonctionnent extrêmement bien tant qu'elles demeurent séparées et cantonnées dans leurs domaines respectifs. La mécanique quantique décrit précisément le comportement des atomes et de la lumière, quand les deux forces nucléaires forte et faible et la force électromagnétique mène le bal et que la gravité est négligeable. La relativité rend bien compte des propriétés de la gravité à l'échelle cosmique de l'Univers, des galaxies, des étoiles et des planètes, quand celle-ci occupe le devant de la scène et que les forces nucléaires et électromagnétique ne jouent plus le premier rôle. Mais la physique connue s'essouffle et perd ses moyens quand la gravité, d'ordinaire négligeable à l'échelle subatomique, devient aussi importante que les trois autres forces. Or c'est exactement ce qui est arrivé aux premiers instants de l'Univers. […]*
>
> *L'infiniment petit a […] accouché de l'infiniment grand, et pour comprendre l'origine de l'Univers et, par conséquent, notre propre origine, il nous faut une théorie physique qui soit capable d'unifier la mécanique quantique avec la relativité et de décrire une situation où les quatre forces fondamentales qui contrôlent l'Univers sont sur un pied d'égalité.*
>
> *Or cette tâche d'unification n'est pas aisée car il existe […] une incompatibilité fondamentale entre la mécanique quantique et la relativité générale en ce qui concerne la géométrie de l'espace. Selon la relativité, l'espace à grande échelle où se déploient les galaxies et les étoiles est lisse et dépourvu de toute rugosité. Par contre, l'espace à l'échelle subatomique de la mécanique quantique n'est plus lisse, mais devient une sorte de mousse sans forme définie, remplie d'ondulations et d'irrégularités, surgissant et disparaissant sur des temps infinitésimalement petits, perpétuellement en mouvement et perpétuellement changeante.* » (THUAN 2000)

Ainsi, même ces deux grandes théories physiques que sont la relativité générale et la mécanique quantique sont limitées. Plusieurs physiciens planchent sur des théories en capacité de les englober toutes deux. La plus connue est probablement la théorie des cordes[19]. Cependant, nous sommes encore loin d'une théorie ultime permettant de rendre compte de l'ensemble des phénomènes observables.

[19] Cette dernière suppose que les particules élémentaires de la matière ne seraient pas des particules ponctuelles mais des petites cordes vibrantes dont les différentes modalités vibratoires détermineraient les particularités des éléments constituants de la matière.

Cette notion de *validité limitée* ne s'applique pas qu'à la physique mais à toute théorie scientifique, y compris en sciences humaines. Certes, ces dernières ne disposent pas encore de théories aussi globales que peuvent l'être celles de la physique. Néanmoins, les différentes constructions de ces disciplines sont, elles aussi, des « vérités approchées ». Par exemple, la validité de la loi économique de l'Offre et de la Demande ne tient que sous des hypothèses précises :

- en présence d'*homo œconomicus*, c'est-à-dire d'acteurs ayant un comportement rationnel de maximisation de leurs satisfactions conformément aux principes de l'utilitarisme ;
- en supposant une *concurrence « pure et parfaite »* sur les marchés des facteurs de la production et des produits (atomicité de l'offre et de la demande, information parfaite, etc.).

En toute autre circonstance, cette théorie ne donne qu'une approximation grossière du fonctionnement des marchés. Nous voyons de suite apparaître les limites de cette loi qui ne peut rendre compte dans les faits de la complexité des phénomènes économiques, ces deux hypothèses fondamentales n'étant jamais parfaitement remplies.

Nous retiendrons l'impossibilité de concevoir une théorie *universelle*. Par contre, pour certaines théories, il est possible d'envisager une *validité limitée* sur une ou plusieurs sous-régions de l'espace-temps. Il nous faut alors envisager le principe de réfutabilité sous un angle différent. Si l'exigence de réfutabilité doit être maintenue, afin de conférer une objectivité aux énoncés et théories étudiés, un test de réfutation non concluant, la théorie étant incapable de rendre compte du phénomène étudié, n'impliquera alors pas nécessairement une réfutation complète de la théorie, mais une limitation accrue du domaine d'applicabilité de cette dernière, la théorie pouvant rester non falsifiée et corroborée expérimentalement en d'autres endroits ou dans d'autres contextes.

2.4 À propos des énoncés singuliers et universels

Nous avons vu précédemment que POPPER distinguait énoncés singuliers et énoncés universels. Si une large part du débat épistémologique a porté sur la manière de valider, corroborer ou réfuter des énoncés universels et des théories en tant qu'ensemble d'énoncés universels, relativement peu d'écrits portent sur des critères ou des méthodes pratiques pour statuer quant à la validité d'un énoncé d'observation. Par ailleurs, à la suite de ce qui a été évoqué à la section précédente à propos de la notion de validité limitée, la distinction entre énoncés universels et

énoncés singuliers semble trop binaire et restrictive. Ne pourrait-on envisager des énoncés ayant une universalité limitée ?

Il nous semble plus pertinent d'établir un continuum et de réinscrire chaque énoncé au sein d'une portion donnée de l'espace-temps pour laquelle nous pourrons le considérer comme valable, ce qui correspondrait à son domaine d'applicabilité. L'énoncé singulier le plus simple portera sur un unique point de l'espace-temps. Par exemple, "*le lundi 13 août 2007, à 17h52, il y avait un renard au pied de l'arbre situé face à l'entrée du CEPED au Jardin Tropical de Paris*". Un énoncé un peu plus général sera représenté non plus par un point mais par une superficie d'espace-temps : "*le lundi 13 août 2007, de 17h30 à 18h, il y avait un renard tournant autour de l'arbre situé face à l'entrée du CEPED*". Un énoncé très général portera quand à lui sur une portion d'espace-temps particulièrement importante. Ainsi, la loi de la gravitation universelle de NEWTON pourra être associée à l'ensemble des domaines de l'espace-temps de notre univers tels que l'espace-temps en question puisse être considéré comme « plat ». La théorie de la transition démographique, quant à elle, s'inscrit dans une région d'espace-temps plus limitée, à savoir la planète Terre, sur la période XVIII^{e}-XXI^{e} siècle.

En procédant ainsi, nous situons chaque énoncé dans le temps et l'espace et par là-même commençons à décrire le contexte au sein duquel il s'applique. Il y a un autre élément de tout énoncé qui nous semble fondamental, à savoir les objets sur lesquels il porte, ce que nous pouvons appeler la *population associée* à cet énoncé. Dans le cadre de la loi de la gravitation universelle, il s'agit par exemple de l'ensemble des objets possédant une masse. En informatique, les microprocesseurs forment la population concernée par la loi de MOORE[20]. La théorie de la transition démographique porte sur des populations humaines suffisamment grandes pour que des tendances démographiques y soient visibles.

Il nous semble important de préciser un point de vocabulaire afin d'éviter des confusions sur le terme de population, notamment lorsque l'on travaille en sciences humaines. Dans le cas de la théorie de la transition démographique, la *population associée* à cet énoncé est constituée de *populations humaines*, à savoir de groupes d'êtres humains. Les « individus » ou « éléments » constituant ce que nous avons appelé la « population sur laquelle porte cette théorie » sont donc eux-mêmes des agrégats d'êtres humains, que nous appelons couramment des « populations humaines ». Il importe donc de ne pas confondre ces deux emplois du mot « population ». En langage mathématique, la population associée à un énoncé

[20] En 1975, Gordon MOORE, l'un des trois fondateurs de la société Intel, a formulé ce que l'on nomme la seconde Loi de MOORE. Cette dernière stipule que le nombre de transistors des micro-processeurs doit doubler tous les deux ans. Cette loi, bien qu'intuitive, a été à peu près vérifiée depuis 1973. Dans la mesure où il est possible de la tester intersubjectivement, nous pouvons considérer qu'elle a une signification objective.

s'avère être un ensemble. Dans le cadre de la théorie des ensembles, un *ensemble* désigne une collection d'objets, que l'on nomme *éléments* de l'ensemble. Et les éléments d'un ensemble peuvent être eux-mêmes des ensembles. Prenons l'énoncé suivant : "*la prévalence du VIH était de 1,8 % au Burkina Faso en 2003*"[21]. Il s'agit d'un énoncé simple dont la population associée est un ensemble ne comportant qu'un seul élément : la population burkinabé. Il s'avère, dans le cas présent, que cet élément est lui-même un ensemble composé des êtres humains vivant au Burkina Faso. Par ailleurs, les femmes enceintes du Burkina Faso forment également un ensemble et ce second ensemble s'avère être un sous-ensemble de la population burkinabé : toute personne faisant partie des femmes enceintes burkinabé fait également partie de la population burkinabé. Néanmoins, l'ensemble constitué par les femmes enceintes burkinabé ne fait pas partie de l'ensemble associé à l'énoncé stipulant une prévalence de 1,8 %. Exprimé sous une forme mathématique, si un ensemble E est un élément d'un ensemble U, tout sous-ensemble S de E n'est pas un élément de U[22]. Un ensemble élément d'un autre ensemble n'est pas un sous-ensemble de ce dernier. Autrement dit, l'énoncé "*la prévalence du VIH était de 1,8 % au Burkina Faso en 2003*" ne nous informe en rien sur ce qu'il en était parmi les femmes enceintes. De la même manière, un énoncé tel que "*la prévalence mondiale du VIH est inférieure à 5 %*" ne nous informe en rien sur la situation du Burkina Faso, alors qu'un énoncé comme "*la prévalence du VIH est inférieure à 5 % dans tous les pays du monde*" implique, si cet énoncé est juste, que la prévalence du VIH au Burkina Faso doit être elle aussi inférieure aux 5 % mentionnés.

À chaque énoncé, nous pouvons alors associer un ensemble *population-espace-temps*, que nous noterons *PET*, définissant les éléments sur lesquels porte cet énoncé et les inscrivant à la fois dans l'espace et dans le temps. Une fois ces ensembles population-espace-temps définis, nous pouvons établir certaines règles. L'énoncé le plus singulier possible sera associé à un ensemble population-espace-temps ne comportant qu'un seul élément. Un énoncé universel au sens strict sera associé à un ensemble population-espace-temps correspondant à l'Univers dans sa totalité, incluant tous les objets le composant. Un énoncé A pourra être considéré comme étant plus universel ou plus général qu'un énoncé B s'il porte sur un nombre plus important d'éléments et/ou sur une plus grande portion de l'espace-temps, c'est-à-dire si l'ensemble population-espace-temps associé à A inclut l'ensemble population-espace-temps associé à B. Ainsi, l'énoncé "*tout système*

[21] Il s'agit du résultat mesuré par l'Enquête Démographique et de Santé de 2003 au Burkina Faso auprès des femmes et des hommes âgés de 15 à 49 ans. Par souci de simplicité, nous ne faisons pas mention de l'âge dans le corps du texte.

[22] Alors que si E était inclus dans U (c'est-à-dire si E était un sous-ensemble de U), alors tout sous-ensemble de E aurait également été un sous-ensemble de U.

économique connaît au moins une crise par période de 50 ans" sera plus général que l'énoncé "*la France a connu une crise économique entre 1960 et 1990*".

> **Postulat 2.2**
> *L'énoncé A sera plus universel que l'énoncé B si et seulement si* $PET(B) \subset PET(A)$, *PET représentant l'ensemble population-espace-temps associé à tout énoncé et définissant son domaine d'applicabilité.*

Dès lors, un énoncé A ne pourra être corroboré ou réfuté par un autre énoncé B, sous réserve de la validité de ce dernier, qu'à la condition que l'ensemble population-espace-temps du premier contienne l'ensemble population-espace-temps du second. L'énoncé "*les ornithorynques pondent des œufs*" ne sera pas un falsificateur de l'énoncé "*les euthériens*[23] *ne pondent pas d'œufs*" dans la mesure où les ornithorynques appartiennent aux protothériens. Par contre, cet énoncé sera un falsificateur possible de "*les mammifères ne pondent pas d'œufs*" dans la mesure où les protothériens font partie des mammifères. Nous avons évoqué ici le cas particulier où l'ensemble population-espace-temps de B est entièrement inclus dans l'ensemble population-espace-temps de A. Plus généralement, l'énoncé B sera un falsificateur possible de A si l'intersection des ensembles population-espace-temps de A et de B est non vide (il existe au moins un élément commun à ces deux ensembles) et la falsification ne pourra être réalisée que sur les assertions de ces deux énoncés à cette intersection. L'énoncé "*les femmes utilisant un moyen contraceptif ne souhaitent pas avoir d'enfants*" pourra être un falsificateur de l'énoncé "*les femmes enceintes souhaitent avoir un enfant*" dans la mesure où parmi les femmes utilisant un moyen contraceptif, certaines se retrouvent néanmoins enceintes. On ne peut comparer deux énoncés qu'à l'intersection de leurs ensembles population-espace-temps.

> **Postulat 2.3**
> *Un énoncé B ne pourra être un falsificateur de l'énoncé A qu'à la condition que* $PET(B) \cap PET(A) \neq \varnothing$ *(l'intersection des ensembles population-espace-temps de A et B n'est pas vide).*

[23] Les euthériens ou placentaires forment l'infra-classe la plus importante des mammifères. Les embryons des espèces de cette infra-classe se développent entièrement dans le corps de leur mère, et sont alimentés pendant la grossesse grâce au placenta. Les primates, les canidés ou encore les cétacés appartiennent à cette infra-classe tandis que l'ornithorynque appartient à la sous-classe des protothériens.

La falsification d'un énoncé ne portant que sur une partie de son ensemble population-espace-temps, il est possible de modifier cet énoncé afin qu'il ne soit plus falsifié en modifiant son ensemble population-espace-temps de telle sorte que l'intersection de ce dernier avec celui de l'énoncé falsificateur devienne vide. Si l'on transforme l'énoncé précédemment cité en "*les femmes enceintes, à l'exception de celles utilisant un moyen de contraception, désirent avoir un enfant*", il ne pourra plus être falsifié par "*les femmes utilisant un moyen contraceptif ne souhaitent pas avoir d'enfants*". L'exemple présenté ici pourra paraître trivial. Il correspond à des énoncés ayant un faible niveau de généralité. Dans le cadre d'énoncé universel, il conviendra d'être prudent sur la manière de réduire l'ensemble population-espace-temps afin d'éviter les modifications *ad hoc* peu porteuses de sens. Si l'on revient à la physique newtonienne réfutée par la précession du périhélie de Mercure, un énoncé tel que "*deux corps possédant une masse (à l'exception de la planète Mercure) s'attirent proportionnellement à leur masse et en raison inverse du carré de leur distance*" serait inadéquat et correspondrait au type de modifications *ad hoc* que l'on doit éviter, en l'occurrence une réduction de l'ensemble population-espace-temps par la suppression d'un élément singulier plutôt que d'une classe d'éléments. La réfutation de la mécanique classique par les vérifications expérimentales de la relativité générale a amené à limiter la validité des équations newtoniennes aux portions de l'espace temps "*où les champs de gravitation doivent être considérés comme faibles et où toutes les masses se déplacent, par rapport au système de coordonnées, avec des vitesses qui sont petites comparées à celle de la lumière*"[24]. Il s'agit ici d'une réduction pertinente de l'ensemble population-espace-temps de la théorie newtonienne, renvoyant à la notion de validité limitée d'EINSTEIN.

2.5 La notion de concept opératoire

La recherche de l'objectivité a conduit depuis longtemps les physiciens à n'accepter que l'usage de *concepts opératoires*. Selon ULLMO, « *une définition opératoire est une définition qui comporte la description d'un procédé régulier pour repérer, mesurer, plus généralement atteindre et identifier le concept défini*[25] ». Cela ne signifie pas seulement que la définition d'un concept doit être fondée sur des critères permettant d'effectuer des mesures ; mais, plus profondément, que le

[24] *EINSTEIN A., La Théorie de la Relativité restreinte et générale, Paris (FR), Dunod, 1917, p. 14.*

[25] ULLMO J., *La Pensée scientifique moderne*, Paris (FR), Flammarion, 1969, p. 24-25.

concept est défini ou constitué par le système même des « *relations répétables*[26] » qui permettent de le dégager.

> « *Nous remarquons que cette définition de la définition comporte un postulat de répétition. Il faut que n'importe qui*[27] *puisse répéter les opérations incluses dans la définition opératoire et soit assuré alors d'aboutir aux mêmes constatations qui ont été initialement comprises en elle. D'où viendra la garantie de cette répétition ? Justement du choix des êtres particuliers sur lesquels portent les définitions de la science ; c'est pourquoi nous avons parlé […] de la nécessité où celle-ci se trouve de construire ses objets. Il ne s'agit pas comme le dit E. MACH*[28]*, de "classer et de rassembler des faits qui sont individuellement donnés ; [le savant] doit avant tout trouver les caractères dont il faut tenir compte".*
>
> *En se refusant à employer d'autres concepts qu'opératoires, la science s'impose de reconnaître, dans la diversité mouvante des apparences sensibles, des objets qui se prêtent à la répétition.*
>
> *Cette répétition ne saurait être celle des faits d'observation ou phénomènes. Proprement, un phénomène ne se répète pas. Un fait observé dans la nature présente toujours des conditions si complexes que leur répétition identique et détaillée ne se produira pratiquement jamais. Même l'observation dirigée qu'on nomme expérience ne peut nous offrir la répétition d'un phénomène, du moins à ce stade de la recherche qui correspond aux définitions. […]*
>
> *Au stade initial de la construction de la science où nous sommes, la répétition n'est pas manifeste. L'obtenir, c'est le premier effort, et peut-être le plus dur, de la recherche.* » (ULLMO 1969, p. 25-26)

La notion de concept opératoire renvoie, dans un vocabulaire plus familier aux sciences humaines et sociales, à l'idée d'*indicateur*. Un indicateur définit son objet de recherche et précise également, dans le cadre d'une approche quantitative, les opérations mathématiques qui sous-tendent son calcul. Ainsi, la définition de l'Indicateur Synthétique de Fécondité (ISF) en démographie incorpore la manière de le calculer. L'ISF est obtenu par l'addition des taux de fécondité par âge, entre 15 et 50 ans exacts, observés sur une période de temps donnée, le plus souvent une année. Il apparaît qu'un indicateur peut faire référence à d'autres indicateurs. En

[26] ULLMO, *La Pensée scientifique moderne*, p. 26.

[27] Note de Jean ULLMO : « *C'est ici que s'établit le fossé radical entre l'expérience scientifique et tout autre type d'expérience, par exemple l'expérience mystique ; la science doit rejeter celle-ci, non parce qu'elle conteste qu'elle soit réelle, mais parce qu'elle n'est pas répétable.* »

[28] MACH E., *La Connaissance et l'erreur*, Paris (FR), Flammarion, 1905, p. 307.

l'occurrence, l'ISF fait appel à la notion de taux de fécondité par âge, ces derniers étant définis comme le rapport entre le nombre de naissances vivantes issues de mères d'un certain groupe d'âges rapporté au nombre de femmes de ce même groupe d'âges[29]. Les concepts opératoires, ou indicateurs, tels que définis, sont la résultante d'une théorie et, en ce sens, précèdent l'observation, conformément à ce que nous avons évoqué précédemment (voir la section 2.2). Les indicateurs qui seront retenus dans le cadre d'une recherche détermineront, en partie, la manière dont les observations seront effectuées. Par exemple, afin de pouvoir calculer l'ISF, il importe de procéder à un dénombrement des naissances en fonction de l'âge de la mère ainsi que des personnes soumises au risque de fécondité.

> « *Il n'y a donc, dans le concept opératoire, rien de plus (mais aussi bien : rien de moins) qu'un système d'opérations, effectivement réalisables par le physicien, qui se contrôlent et se recoupent mutuellement. Et ce sont justement les invariants qui mettent en évidence ces contrôles et recoupements mutuels qui constituent les concepts opératoires.*
>
> *Il en résulte aussitôt que la valeur, et même la signification objective d'un concept opératoire, sont toujours relatives et limitées : relatives à l'échelle et au secteur de la réalité où les opérations qui le constituent ont un sens ; limitées par la précision des mesures qui le définissent. Et il ne s'agit pas seulement de dire que nous ne pouvons jamais avoir qu'une connaissance approximative des "vraies" valeurs qui existeraient par ailleurs, bien que nous ne les connaissions pas. Des concepts parfaitement opératoires à notre échelle, comme la longueur ou la vitesse, s'estompent dans une sorte de flou et perdent peu à peu toute signification objective au fur et à mesure que nous descendons vers les échelles microscopiques[30]. C'est ici que la distinction nécessaire du modèle et de la réalité prend toute son importance. Car, une fois que les concepts opératoires et les lois physiques qui les sous-tendent ont été rassemblés dans le cadre d'un modèle mathématique, la tentation est grande d'oublier ces limites et, se fiant aveuglément au formalisme mathématique, de tirer du modèle des déductions qui vont bien au-delà de son domaine de validité objective. […] L'existence de ce [seuil d'objectivité], et la tentation de le franchir, constituent un danger*

[29] L'ISF et les taux de fécondité peuvent également être appliqués aux hommes de la même manière. Nous parlons ici de « femmes » et de « mères » par souci de commodité. Par ailleurs, nous remarquerons que la notion de taux de fécondité par âge fait elle-même référence à d'autres indicateurs ou concepts opératoires tels que l'âge ou la notion de naissance vivante.

[30] Note de MATHERON : « *La longueur d'une règle est définissable, disons, au dixième de millimètre près, mais non avec 15 décimales exactes : à plus forte raison, la question de savoir si cette longueur s'exprime en centimètres par un nombre rationnel ou irrationnel n'a absolument aucun sens pour un physicien.* »

> *permanent que nous devons tout particulièrement garder présent à l'esprit lorsque nous mettons en œuvre des modèles*[31]. »
>
> (MATHERON 1978, p. 34-35)

Les relations répétables qui définissent les concepts opératoires permettent de pouvoir procéder à une observation empirique et, par là-même, de pouvoir tester une théorie ou un énoncé. Cela implique que, pour pouvoir être soumis à une vérification empirique, un énoncé devra être traduit, d'une part, en indicateurs et, d'autre part, en une conditionnalité sur ces derniers, afin que cet énoncé puisse être testé. Cette opération, à savoir la *transformation d'un énoncé en une conditionnalité sur des concepts opératoires*, correspond à ce qui est usuellement nommé en sciences sociales sous le terme d'*opérationnalisation des hypothèses*. Le chercheur part d'une *problématique* générale et pose un cadre théorique fournissant une explication possible des phénomènes définis par cette problématique. De ce cadre théorique appliqué à cette problématique vont découler un certain nombre d'*hypothèses* sur ce qui devrait être observé si ce cadre conceptuel s'avérait exact. La formulation des hypothèses correspond à l'élaboration d'énoncés d'universalité moindre découlant logiquement de la théorie proposée. Le chercheur va alors transformer ces différentes *hypothèses* en *hypothèses opérationnelles* en choisissant et en définissant différents indicateurs et en précisant, dans un second temps, le comportement de ces derniers si chacune de ces hypothèses étaient confirmées. Ce n'est qu'à partir de ce moment là que le scientifique pourra les vérifier empiriquement en confrontant ce que prédit sa théorie aux données collectées et observées. Cette conditionnalité sur des concepts opératoires précisera les dimensions des indicateurs qui seront porteuses de sens. Un indicateur de niveaux présentera une conditionnalité différente de celle d'un indicateur de tendances.

> **Postulat 2.4**
> *Pour pouvoir être testé empiriquement, un énoncé doit être traduit en une conditionnalité sur des concepts opératoires ou indicateurs, la définition d'un concept opératoire devant comporter la définition d'un procédé régulier, d'une relation répétable, permettant de mesurer et d'identifier le concept ainsi défini.*

Prenons un exemple concret. En différents endroits, la prévalence du VIH observée chez les femmes en population générale diffère de celle observée chez les femmes enceintes : en Tanzanie (KWESIGABO 2000, CHANGALUCHA 2002), au Malawi

[31] Matheron écrit précisément « *modèles probabilistes* ». Il nous semble que ce qu'il écrit peut s'appliquer à tout modèle, qu'il soit probabiliste ou non.

(CRAMPIN 2003), en Zambie (FYLKESNES 1998), au Zimbabwe (GREGSON 1995, GREGSON 2002b) et en Ouganda (CARPENTER 1997, GRAY 1998), où des données provenant à la fois de sites sentinelles et d'enquêtes en population générale sont disponibles (voir l'annexe 1 pour plus de détails). Afin d'expliquer ce fait, nous pouvons émettre la théorie suivante : la séropositivité au VIH induirait une sous-fécondité des femmes, les femmes VIH+ étant alors moins souvent enceintes que les femmes VIH-[32]. Ce différentiel permettrait d'expliquer à lui seul, selon cette théorie, les différences observées entre les mesures en population générale et les mesures effectuées auprès des femmes enceintes[33].

De cette théorie nous allons pouvoir en déduire deux hypothèses :

1. d'une part, dans les enquêtes en population générale, nous devrions observer une proportion plus faible de femmes enceintes parmi les femmes VIH+ ;
2. d'autre part, les observations effectuées parmi les femmes enceintes en clinique prénatale devraient pouvoir être déduites de notre connaissance de l'épidémie parmi l'ensemble des femmes et du différentiel de fécondité constaté.

Pour procéder à une vérification empirique, nous devons opérationnaliser nos hypothèses et, pour se faire, nous allons avoir recours à deux indicateurs principaux :

- la *prévalence du VIH*, définie comme la proportion de personnes séropositives au sein d'un groupe, que nous calculerons à la fois pour l'ensemble des femmes (P_{toutes}) et pour les femmes enceintes uniquement (P_{enc}) ;
- l'*Odds Relatif de l'Infection*[34] *(ORI)* qui correspond au ratio entre la proportion de femmes enceintes parmi les femmes VIH+ et la proportion de femmes enceintes parmi les femmes VIH-[35].

Une fois ces deux indicateurs (ou concepts opératoires) définis et choisis, nos deux hypothèses vont pouvoir être réécrites sous la forme suivante :

1. l'ORI, calculé en population générale, doit être inférieur à 1, traduisant ainsi que les femmes VIH+ sont moins susceptibles d'être enceintes que les femmes VIH- ;

[32] Nous utilisons l'abréviation *VIH+* pour désigner les personnes séropositives au VIH et *VIH-* pour les personnes séronégatives au VIH.

[33] Afin de simplifier notre exemple, nous supposons ici que la proportion de femmes consultant en clinique prénatale en cas de grossesse est la même parmi les femmes VIH+ et parmi les femmes VIH-.

[34] Nous avons préféré le terme anglais, fréquent dans la littérature, plutôt que sa traduction française (rapport de cotes relatif) peu usitée.

[35] Pour plus de détails sur l'ORI, voir l'annexe 1.

2. les données collectées en population générale et en clinique prénatale vérifient la relation suivante (voir annexe 1 pour les détails) :

$$P_{toutes} = \frac{P_{enc}}{ORI - P_{enc} \cdot ORI + P_{enc}}.$$

Nos tests sont maintenant parfaitement définis. Pour chacun des énoncés (hypothèses) déduits de notre théorie, nous disposons bien d'une conditionnalité sur des concepts opératoires (indicateurs) nous permettant de corroborer ou d'infirmer nos propos.

2.6 Observations et énoncés d'observation

Nous venons de voir qu'avant de procéder à une opération de mesure, il importe de définir, au préalable, les concepts opératoires qui nous seront nécessaire pour infirmer ou vérifier nos différentes hypothèses. La définition de ces indicateurs déterminera en partie la manière dont nous observerons le phénomène que nous étudions. Nous avons vu, à titre d'exemple, que si nous souhaitons calculer un indice synthétique de fécondité, il sera nécessaire de dénombrer des naissances en fonction de l'âge de la mère.

Cependant, une fois nos indicateurs, nos concepts opératoires, définis, il n'y a pas une seule et unique manière possible de procéder à nos observations. Plusieurs techniques différentes peuvent être envisagées. L'observation résulte d'un *choix méthodologique* effectué par l'enquêteur. De nombreux critères peuvent influer ce choix, certains ne relevant pas de la question scientifique posée mais de considérations purement pragmatiques telles que le coût financier de la collecte de données ou les ressources temporelles et humaines à disposition de l'équipe. Par exemple, pour estimer la prévalence du VIH parmi les personnes ayant connu un épisode d'Infection Sexuellement Transmissible (IST), nous pouvons réaliser une enquête dans les services de santé prenant en charge les IST ou bien réaliser une enquête en population générale où nous demanderons aux individus s'ils ont été atteints d'une IST au cours des douze derniers mois. Il s'agit de deux manières d'observer le phénomène qui nous préoccupe, correspondant à deux choix distincts, selon deux *modes opératoires* donnés. Autrement dit, une opération de collecte vise à quantifier un concept opératoire donné selon un mode opératoire choisi.

Cependant, l'observation en elle-même, brute, ne quantifie pas directement l'indicateur qui nous préoccupe. Par exemple, dans le cadre de l'EDS 2004 réalisée au Cameroun, les données d'observation brutes correspondent aux résultats des tests de dépistage du VIH appliqués sur le sang des 9 900 personnes testées, âgées

de 15 à 49 ans, et qui se répartissent sur 466 zones d'enquêtes. Pour le moment, la valeur de l'indicateur "prévalence du VIH" n'est pas encore déterminée. À ce stade, il ne s'agit que d'une suite de "*négatif*"' et de "*positif*". Ou bien encore, il s'agit d'un tableau (forme sous laquelle peut-être représenté le contenu de la base de données). Pour transformer ces *données d'observation brutes* en un *énoncé d'observation*, il est nécessaire de procéder à un calcul mathématique[36], en appliquant la définition du concept opératoire considéré aux données. L'énoncé d'observation obtenu consistera alors en une valeur quantitative applicable à une population sur une superficie d'espace-temps donné. En l'occurrence, dans le cas de l'EDS 2004 réalisée au Cameroun, nous pourrons calculer l'observation numérique suivante : "prévalence de 5,5 %" qui sera associée à la population constituée par les 9 900 personnes enquêtées, sur la période de l'enquête, à savoir février-juillet 2004, et aux 466 zones enquêtées.

> **Postulat 2.5**
> *Un énoncé d'observation naît de l'application d'un concept opératoire donné à des données d'observations brutes obtenues selon un mode opératoire choisi.*

La notion d'énoncé d'observation présentée ici ne correspond pas forcément, dans un cadre quantitatif, à un nombre unique. Il peut également s'agir d'une suite de nombres, répartis dans le temps et/ou l'espace et/ou au sein d'un ensemble d'individus, obtenus par application d'un même concept opératoire à une série de sous-ensembles de données brutes (par exemple, des prévalences par âge ou bien une série de prévalences annuelles calculées sur un site sentinelle donné). Par ailleurs, nous pouvons remarquer, à partir de l'exemple de l'EDS 2004 du Cameroun, que le mode opératoire retenu détermine l'ensemble population-espace-temps que nous avons associé à notre énoncé d'observation[37]. En effet, si nous avions eu recours à un autre mode opératoire, les personnes sélectionnées pour l'enquête n'auraient pas été les mêmes et l'ensemble population-espace-temps de notre énoncé d'observation en aurait été affecté. C'est pourquoi, à tout énoncé d'observation, nous associerons également le mode opératoire retenu pour la collecte des données brutes.

---

[36] Il s'agit en l'occurrence d'un calcul mathématique dans la mesure où notre indicateur est quantitatif. Pour un indicateur qualitatif, le raisonnement est analogue. Par exemple, si notre indicateur est une catégorisation, nous déterminerons la catégorie d'appartenance d'un individu, en fonction de nos observations, selon la définition de chaque catégorie telle qu'elle est formulée par notre concept opératoire.

[37] Comme tout énoncé, il importe en effet de préciser l'ensemble population-espace-temps associé à un énoncé d'observation.

Enfin, comme tout énoncé, les énoncés d'observation peuvent également s'exprimer sous la forme d'une conditionnalité sur des concepts opératoires. Cette dernière, dans le cas des énoncés d'observation, est relativement simple. Il s'agit en effet d'une relation d'égalité associant une valeur donnée à un indicateur (ou une série de valeurs à une série d'indicateurs pour les énoncés d'observation plus complexes).

2.7 Domaine de validité d'un énoncé d'observation

L'application d'un concept opératoire, c'est-à-dire l'exécution du procédé répétable définissant un indicateur, permet de transformer des données d'observation brutes en un énoncé d'observation. Disposant dorénavant d'un énoncé, il est possible de procéder à des comparaisons ou à des raisonnements déductifs en vue de corroborer ou d'infirmer les énoncés que nous avons déduits de la théorie que nous cherchons à tester. Cependant, dans nombre de situations, l'énoncé d'observation obtenu est trop limité pour vérifier nos hypothèses. Cela est particulièrement vrai dès lors que les observations portent sur un sous-échantillon de la population visée. Revenons à notre exemple portant sur l'EDS réalisée en 2004 au Cameroun. Dans la pratique, nous cherchons le plus souvent à valider ou infirmer des énoncés portant sur l'ensemble de la population du pays. Par exemple, nous cherchons à savoir si "*la proportion d'adultes camerounais séropositifs était comprise entre 5 et 6 % en 2004*". Cet énoncé simple, que nous nommerons énoncé ET pour énoncé théorique, est caractérisé par une conditionnalité sur le concept de prévalence, ce dernier devant se situer entre 0,05 et 0,06, et par une population réduite à un seul élément, la population camerounaise, sur un espace-temps couvrant l'ensemble du Cameroun sur une période d'un an. En appliquant la définition du concept de prévalence aux données de l'EDS, nous obtenons un énoncé d'observation EO stipulant que "*la prévalence du VIH était de 5,5 % chez les 9 900 adultes camerounais enquêtés entre février et juillet 2004 sur les 466 zones retenues pour l'enquête*". Si la portion de l'espace-temps associée à cet énoncé est bien incluse dans la portion d'espace-temps de ET, il n'en est pas de même de la population associée à EO. En effet, cette dernière est un ensemble composé d'un seul élément, à savoir les 9 900 personnes testées. Or, nous avons montré dans la section 2.4 que, si ces 9 900 personnes en tant qu'ensemble constituent bien un sous-ensemble de la population camerounaise, en tant qu'élément d'un ensemble, la population associée à EO, elles ne font pas partie de la population associée à ET. Autrement dit, l'ensemble population-espace-temps de notre énoncé d'observation n'est pas inclus dans l'ensemble population-espace-temps de notre hypothèse, $PET(EO) \not\subset PET(ET)$ et l'intersection de ces deux ensembles population-espace-temps est vide, $PET(EO) \cap PET(ET) = \varnothing$. Il en résulte, qu'en vertu du Postulat 2.3, notre énoncé d'observation EO ne peut être un falsificateur de notre énoncé

théorique ET. Notre observation calculée sur les données de l'EDS ne nous informe en rien, pour le moment, sur ce qu'il en est dans la population générale camerounaise.

Or, dans la pratique, les observations effectuées sur les personnes enquêtées dans l'EDS seront appliquées à l'ensemble de la population camerounaise. Ainsi, le rapport final de l'enquête conclut que « *les résultats de l'EDSC-III*[38] *de 2004 montrent qu'au Cameroun, 5,5 % des adultes âgés de 15-49 ans sont séropositifs [au VIH]*[39] ». Implicitement, les auteurs sont passés d'un énoncé d'observation portant sur les 9 900 personnes testées à un énoncé portant sur l'ensemble de la population camerounaise. Cette extrapolation est courante et repose en l'occurrence sur un modèle probabiliste déterminé par la théorie des sondages. Ce point nous semble d'autant plus essentiel qu'il est le plus souvent occulté, considérant que cela relève de l'évidence. Or la question demeure de savoir si l'énoncé du rapport de l'enquête (énoncé R), déduit de l'énoncé d'observation EO, possède une signification objective.

Nous verrons dans la section suivante que le passage d'un énoncé d'observation à des énoncés d'information plus générale repose sur le recours à des modèles (probabilistes, déterministes ou autres) qui résultent du choix d'une ou de plusieurs *hypothèses anticipatrices*. Nous décidons de nommer *domaine de validité* de l'énoncé d'observation EO l'ensemble des énoncés ayant une signification objective qui peuvent être déduits, à l'aide de méthodologies données, de cette observation numérique. La définition du domaine de validité d'un énoncé d'observation recouvrera en fait ce que nous entendons, à la suite de MATHERON et d'autres auteurs, par *estimation*. Le domaine de validité d'un énoncé d'observation sera toujours *provisoire* car modifiable à tout moment lorsque de nouvelles informations seront disponibles.

2.8 Hypothèse anticipatrice et risque d'erreur radicale

Une estimation consiste, à partir d'informations fragmentaires sur lesquelles on aura appliqué un modèle, à extrapoler du connu à l'inconnu, c'est-à-dire à déduire du modèle, étalonné sur les seules données connues, des conclusions que l'on espère valables sur les parties non informées. Plus généralement, une estimation

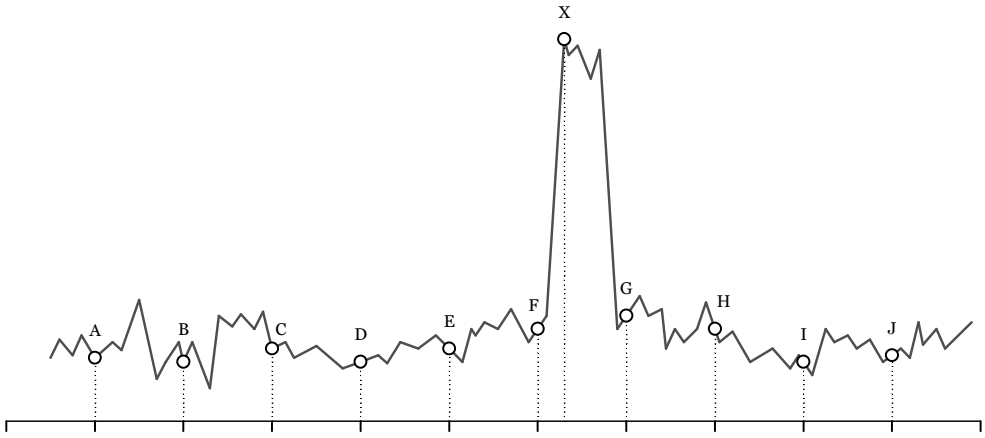
[38] EDSC-III signifie troisième EDS réalisée au Cameroun.

[39] INSTITUT NATIONAL DE LA STATISTIQUE et ORC MACRO, *Enquête Démographique et de Santé 2004 du Cameroun*, Calverton, Maryland (US), INS, ORC Macro, 2005, p. 301.

consiste à extraire de nos données plus d'informations que ce qui est initialement contenu. MATHERON pose la question en ces termes :

> *« À cette question décisive (très kantienne) "Comment l'estimation est-elle possible ?", il semble bien qu'on ne puisse apporter une réponse qu'en acceptant, sous une forme ou une autre, une hypothèse d'homogénéité statistique au moins locale, dont la stationnarité classique constitue une forme extrême, exagérément sévère : le phénomène devrait, en somme, se comporter, là où on ne le connaît pas, d'une manière raisonnablement analogue à ce que l'on peut observer sur les données disponibles au voisinage. Mais, si cette hypothèse est indispensable pour fonder la possibilité de l'estimation, il n'en résulte pas automatiquement qu'elle soit partout et toujours vérifiée. Il est facile de donner des contre-exemples. Sur la figure ci-après [Figure 2.2], où les points A, B, C… J représentent les données disponibles, on peut voir un cas de ce genre : rien ne permet de prévoir l'anomalie X et l'estimation n'est pas possible (quelle que soit la méthode). L'erreur radicale est, ici, inévitable. »*
>
> (MATHERON 1978, p. 7-8)

Figure 2.2
Exemple d'anomalie non prédictible



Source : (MATHERON 1978, Figure 1 page 7).

Quel que soit l'objet d'étude, dès lors que nous ne disposerons que d'informations fragmentaires, nous serons dans l'obligation de procéder à une *estimation* via le choix d'une *hypothèse anticipatrice*. Le plus souvent cette hypothèse anticipatrice correspondra au choix d'un modèle (mathématique, probabiliste, déterministe ou autre) qui nous permettra de réaliser une projection et de déduire de nouveaux énoncés à partir de notre énoncé d'observation initial. Cela ne signifie pas pour autant que n'importe quel modèle puisse être utilisé ou que l'ensemble des paramètres d'un modèle ait une signification. Nous avons montré précédemment (voir section 2.5) que l'objectivité d'un concept est fondée par son caractère

opératoire. Il importe alors de procéder à la *reconstruction opératoire* du modèle avant de le choisir en tant qu'hypothèse anticipatrice.

> « *Cela signifie qu'un concept (mathématiquement bien défini) intervenant dans un modèle (déterministe ou probabiliste) ne pourra être déclaré "objectif" avant d'avoir été entièrement redéfini, ou, au mieux, reconstruit en termes strictement opératoires : métamorphose radicale, ou refaçonnement en profondeur de sa personnalité, si l'on peut dire, qui, de son état initial de simple concept mathématique le fait accéder au statut de concept physique. Lors de l'examen critique d'un modèle probabiliste donné*[40]*, il sera très important de faire le tri : je veux dire distinguer soigneusement les concepts susceptibles d'être ainsi rendus opératoires, et les autres. Les premiers seuls, ainsi que les énoncés, les paramètres, etc. qui leur sont associés pourront être dits objectifs. Les autres (concepts, énoncés, paramètres) resteront purement conventionnels. Ils auront un sens (mathématique) bien défini dans le modèle, mais sans qu'il leur corresponde de contrepartie univoquement constatable dans le phénomène réel. Cela ne nous interdira nullement de les utiliser, mais uniquement à titre heuristique : pour nous suggérer des méthodes ou des algorithmes, auxquels nous n'aurions pas pensé autrement, non pour justifier nos conclusions. Plus précisément, les conclusions qu'ils auront suggérées devront être passées au crible de la critique, reformulées en termes opératoires et soumises à des tests objectifs avant d'être (provisoirement) adoptées… La règle, ici, consistera à s'assurer que toute trace de ces concepts ou paramètres conventionnels a disparu du résultat ultime.* »
>
> (MATHERON 1978, p. 44-45)

Tous les modèles ne sont pas soumis à la même exigence opératoire selon les objectifs visés. Les modèles physiques de la relativité générale ou de la mécanique quantique cherchent à expliquer l'ensemble des phénomènes physiques observables. En ce sens, ils sont *panscopiques*[41]. Il importe alors que l'ensemble des paramètres de ces modèles ait une contrepartie objective dans le monde réel. Par ailleurs, les relations entre ces différents paramètres sont censées traduire les interactions entre les différentes grandeurs physiques et doivent, elles-aussi, avoir une correspondance empirique. Dans d'autres circonstances, les modèles que nous utiliserons viseront uniquement à estimer la grandeur d'un ou de quelques paramètres uniquement. Il s'agit de modèles *polyscopiques* voir *monoscopiques*. Nous avons vu alors que la reconstruction opératoire du modèle employé devra permettre de donner un caractère objectif à chacun des paramètres que nous

[40] Nous étendrons, de notre point de vue, la remarque de MATHERON à tout type de modèles.

[41] Panscopique : tous les buts ; monoscopique : un seul but.

chercherons à estimer. Cependant, les autres caractéristiques du modèle et son fonctionnement interne ne seront pas obligés de répondre à cette même exigence. Le modèle utilisé n'en sera pas moins pertinent au regard de l'objectif visé, tant que nous nous limiterons à l'estimation des paramètres, ou concepts, précités. Par contre, et c'est là où nous devrons être vigilant, notre usage du modèle en question ne devra pas aller plus avant. Ce serait une erreur que d'interpréter le fonctionnement interne du modèle comme explicatif du phénomène étudié sans avoir procédé au préalable à la reconstruction opératoire adéquate.

Ayant choisi un modèle, il devient possible de produire une nouvelle information non contenue initialement dans nos données d'observation. Se faisant, nous prenons le *risque* que notre estimation soit fausse, le choix du modèle ne reposant que sur ce qui a été observé au moment où le chercheur est amené à poser une hypothèse. Cette hypothèse, quelle qu'elle soit, supposera toujours une certaine forme de *continuité* du phénomène étudié : les choses doivent, *a priori*, se passer de la même manière là où nous n'avons pas d'information. Mais il peut arriver que le phénomène change radicalement de nature pour des raisons que nous ne pouvions pas prévoir. Par exemple, dans le cadre de projections démographiques à moyen et long terme, le démographe posera plusieurs hypothèses de mortalité et de fécondité, le plus souvent une poursuite des tendances observées lors des années passées. Mais il ne pourra pas prévoir l'éventualité de l'apparition d'une nouvelle maladie mortelle touchant une large part de la population ou l'éclatement d'une troisième guerre mondiale dans quinze ans.

> « *Le statisticien a donc pris un risque (inévitable). Mais, s'il y a risque, c'est-à-dire si sa prévision peut être démentie par la suite, cela signifie que cette prévision était falsifiable, donc avait un sens objectif. Le statisticien a donc réellement avancé une hypothèse objective (falsifiable) : non pas exactement celle qu'il avait énoncée, qui ne concernait que le modèle mathématique*[42]*, mais, implicitement, une hypothèse anticipatrice relative à la validité du modèle physique, dont nous avons expliqué ci-dessus la construction opératoire. Il s'agit bien d'une hypothèse objective (puisqu'elle peut se révéler fausse après coup), et d'une anticipation (puisque les tests les plus rigoureux effectués sur les [données observées], à supposer qu'ils corroborent la validité du modèle pour [ces données-là], ne nous garantissent en aucune façon que le phénomène ne changera pas de nature par la suite). C'est parce qu'elle est objective (introduit une information supplémentaire, non contenue dans les [données observées]) que cette hypothèse nous permet de tirer de ces données plus qu'il n'y est contenu, et d'avancer une prévision.*

[42] Comme précédemment, nous étendons cette remarque de MATHERON à tout modèle, qu'il soit mathématique ou non.

> *C'est parce qu'elle est anticipatrice (adoptée avant que sa validité n'ait été contrôlée) qu'elle introduit un risque d'erreur radicale[43] : et ce risque est la contrepartie obligée du gain d'information qu'elle introduit.* »
>
> (MATHERON 1978, p. 49-50)

Les critères de validité d'un modèle ne sont pas les mêmes selon qu'il s'agit d'un modèle panscopique ou monoscopique. Le premier type de modèle, cherchant à expliquer une classe de phénomènes relativement universelle, s'apparente à une théorie scientifique. Ce type de modèle devra donc permettre de produire un grand nombre de prédictions différentes et il pourra être corroboré ou réfuté à l'aide de tests intersubjectifs. Le modèle monoscopique, quant à lui, est appliqué le plus souvent à une situation unique à partir de laquelle nous nous retrouvons néanmoins face à la nécessité de réaliser des estimations. Le modèle monoscopique est par ailleurs choisi, par définition, sur la base d'informations incomplètes, d'où l'obligation de poser une hypothèse anticipatrice, qui pourra éventuellement se révéler fausse *a posteriori*, mais qui ne peut être vérifiée en toute rigueur au moment où cette dernière est posée.

> « *Au moment où on le choisit, ce modèle monoscopique introduit une hypothèse anticipatrice dont la légitimité ne peut en aucune façon être garantie par sa compatibilité avec les données numériques disponibles : car cette hypothèse revient, en somme, justement, à admettre que les caractéristiques structurelles que nous avons induites à partir de ces données peuvent être extrapolées telles quelles aux parties inconnues du phénomène ; ou encore, si l'on veut, que le phénomène se comporte, là où on ne le connaît pas, d'une manière suffisamment analogue à ce que l'on a observé là où il est connu. Ceci implique deux conséquences : pour choisir une hypothèse de ce genre, il faut soigneusement tenir compte de toutes les sources d'information, numérique ou non, dont on dispose (connaissances générales sur la physique de ce phénomène, expérience acquise sur des cas analogues, etc.) ; d'autre part affaiblir au maximum cette hypothèse, de manière à la réduire au strict minimum indispensable pour permettre d'atteindre l'objectif visé par le modèle monoscopique : valorisation maximale de toutes les sources d'information, et principe d'économie stricte dans le choix des hypothèses anticipatrices.* » (MATHERON 1978, p. 54-55)

MATHERON pose ici deux règles importantes quant au choix de nos hypothèses anticipatrices. L'hypothèse anticipatrice devant prendre en compte l'information maximale disponible, il apparaît qu'elle dépend, en partie, du mode opératoire

[43] Note de MATHERON : « *Dont l'amplitude est d'un autre ordre de grandeur que la "fourchette d'erreur à 5 %" [...] que prévoit le modèle. C'est pourquoi je parle d'erreur radicale.* »

ayant produit les données d'observations brutes et qu'une analyse critique du procédé de collecte est nécessaire avant d'émettre ce type d'hypothèses.

> **Postulat 2.6**
> *L'hypothèse anticipatrice appliquée à un énoncé d'observation doit prendre en compte l'information maximale disponible au moment du choix de la dite hypothèse et doit être limitée au minimum requis pour permettre l'estimation du ou des concepts opératoires visés.*

La condition de répétabilité nécessaire à la définition de concepts opératoires semble ici faire défaut, d'où la difficulté de justifier la validité objective du modèle. Mais nous avons souligné précédemment qu'il n'était pas nécessaire que l'ensemble du modèle ait une signification objective du moment que, lors de la reconstruction opératoire du modèle, les paramètres sur lesquels nous porterons nos conclusions finales soient pourvus d'une signification objective, et donc falsifiables, dans le monde réel. Par ailleurs, s'il est difficile de déterminer la validité d'un modèle unique, nous avons le plus souvent recours aux mêmes types de modèles pour juger de situations analogues. Les hypothèses anticipatrices de ces modèles auront alors pu être vérifiées *a posteriori* et, si ces hypothèses se sont avérées pertinentes pour des cas semblables, il est probable qu'elles le resteront pour le cas qui nous préoccupe. Par exemple, les sondages réalisés à la sortie des urnes après chaque second tour des élections présidentielles peuvent ensuite être comparés aux résultats réels de l'élection. À la longue, dans la mesure où les méthodes employées s'avèrent relativement efficaces pour estimer, dès 20 h, quel candidat a été élu président de la République, le recours à ces mêmes méthodes lors de l'élection suivante s'avère justifié, bien que cela ne nous prémunisse jamais du risque d'erreur radicale mentionné plus haut.

> *« S'il n'est pas évident que l'on puisse porter un jugement sur chaque cas individuel, il n'en est certainement pas de même de la méthodologie générale que nous utilisons pour choisir des modèles monoscopiques dans chaque cas particulier : elle fera, à la longue, la preuve de sa plus ou moins grande efficacité. En effet, chaque situation, chaque problème est unique et fait l'objet d'un modèle monoscopique ad hoc. Mais il y a des classes de situations et de problèmes, non identiques, mais suffisamment analogues pour que les règles qui dictent le choix du modèle que nous adoptons dans chaque cas puissent être, au moins partiellement, formalisées, et finissent par constituer un système méthodologique : ce système sera soumis à la sanction de la pratique, et devra faire ses preuves ou être abandonné.*

> *En d'autres termes, il est exact que c'est, en définitive, la possibilité de répétition qui fonde l'objectivité (la "relation répétable" de J. ULLMO[44]). Mais cela ne signifie pas qu'il n'y ait pas de science possible de l'unique. D'abord, en effet, c'est toujours en un sens relatif que l'on parle de refaire la "même" expérience. À strictement parler, il n'y a pas deux expériences identiques : elles diffèrent toujours l'une de l'autre par quelques facteurs accessoires (mais c'est nous qui les jugeons tels), et par les conditions de lieu ou de temps. Tout ce que l'on peut dire, c'est que les facteurs qui nous paraissent importants, ont été rendus aussi semblables que nous le permettent nos moyens techniques, et les autres sont ce qu'ils sont. De même, on ne peut pas parler d'un phénomène qui se reproduit, mais seulement d'une classe de phénomènes que nous jugeons suffisamment proches les uns des autres pour les considérer comme équivalents. Cette proximité, ou ressemblance, peut d'ailleurs n'être que qualitative et structurelle, sans aller jusqu'à l'égalité des paramètres numériques descriptifs. En géologie ou en astronomie, par exemple, il n'y a pas deux objets identiques, ce qui n'empêche nullement de fonder l'objectivité sur la répétition du semblable. […]*
>
> *Ce critère externe d'objectivité revient donc, en somme, à examiner si l'on a, ou non, "raison en moyenne" d'utiliser telle méthodologie pour tenter de résoudre telle catégorie de problème.* »
>
> (MATHERON 1978, p. 55-57)

Ainsi, les hypothèses anticipatrices et modèles que nous utilisons ne sont pas totalement dénoués d'objectivité, d'autant plus qu'ils sont falsifiables et vérifiables éventuellement *a posteriori*. Cependant, vis-à-vis d'une situation particulière, nous ne pouvons jamais nous départir du risque d'erreur radicale que nous prenons dès lors que nous tentons de déduire de nouveaux énoncés, non contenus initialement dans nos données. Les *énoncés déduits* qui forment, selon la définition que nous avons posée à la section précédente, le domaine de validité de notre énoncé d'observation, n'auront donc qu'une valeur relative et temporaire, tant que nous n'aurons pu vérifier la validité de notre hypothèse anticipatrice. Autrement dit, toutes les tentatives de corroboration ou de réfutation que nous pourrons être amenés à effectuer à partir de nos énoncés déduits resteront donc, du point de vue épistémologique, en suspens, tant que l'hypothèse anticipatrice qui a permis de les générer n'aura pas été vérifiée ou réfutée. Il en résulte que nous devrons toujours accompagner nos énoncés déduits de l'hypothèse anticipatrice qui les a engendrés, la connaissance de cette dernière étant fondamentale pour un examen critique des dits énoncés.

[44] Voir section 2.5, page 87 et suivantes.

Postulat 2.7
Le domaine de validité d'un énoncé d'observation est composé de l'ensemble des énoncés déduits du dit énoncé d'observation par application d'une hypothèse anticipatrice à ce dernier. Les énoncés déduits ne seront valables temporairement que sous réserve de l'hypothèse anticipatrice qui les a engendrés et comporteront donc un risque d'erreur radicale.

Néanmoins, ceci n'empêchera pas de pouvoir prendre des décisions, même en l'absence de vérification, dès lors que le risque d'erreur sera accepté. Nos choix pourront alors être guidés entre autres par un examen critique de nos hypothèses anticipatrices. Le risque d'erreur radicale que nous encourrons, en acceptant, provisoirement, les résultats de nos estimations, ne peut être quantifié. Il s'agit d'un risque d'une toute autre espèce que celui que l'on peut calculer, par exemple, dans un modèle probabiliste, via un intervalle de confiance. Nous ne disposons *a priori* d'aucune information susceptible de nous guider pour estimer ce risque car, si nous disposions d'une telle information, nous aurions dû, en vertu du Postulat 2.6, en tenir compte dans le choix de notre hypothèse anticipatrice, et elle ne nous serait alors plus utile pour l'estimation de ce risque. Cependant, certains éléments peuvent nous fournir une indication qualitative pour l'évaluation de ce risque afin de faciliter nos prises de décisions. D'une part, si notre hypothèse anticipatrice a été vérifiée empiriquement pour des cas analogues, nous pourrons alors, en vertu de ce principe d'analogie, considérer qu'elle est justifiée pour le cas qui nous préoccupe. D'autre part, si plusieurs modèles ayant des présupposés différents nous amènent aux mêmes conclusions, la stabilité des résultats obtenus sera un facteur confortant nos décisions. Pour autant, le risque d'erreur radicale n'en demeure pas moins.

Postulat 2.8
Le risque d'erreur radicale pourra être considéré comme minimisé si l'hypothèse anticipatrice posée a été vérifiée expérimentalement pour des cas analogues et/ou si plusieurs hypothèses anticipatrices différentes, appliquées à un même énoncé d'observation, conduisent aux mêmes énoncés déduits.

Dans les exemples précédents, la nécessité d'avoir recours à une hypothèse anticipatrice est relativement évidente dès lors que les données d'observation portent sur un sous-échantillon de la population que le chercheur souhaite étudier. C'est le cas, entre autres, de toutes les études ayant recours à la technique des sondages. Mais le recours à une hypothèse anticipatrice se cache également dans des exemples *a priori* moins évidents. Revenons sur la tentative de réfutation de

COPERNIC par Tycho BRAHÉ (voir section 2.2). La théorie copernicienne impliquait l'existence d'une parallaxe des étoiles lointaines. Or, ses observations ne montrant pas l'existence d'une telle parallaxe, BRAHÉ conclut que COPERNIC s'était trompé. Mais, pour passer de l'énoncé d'observation "*aucune parallaxe des étoiles fixes n'a été détectée*" à l'énoncé déduit "*les étoiles fixes ne présentent pas de parallaxe*", il est nécessaire de poser une hypothèse anticipatrice telle que "*si les étoiles lointaines présentent une parallaxe, alors cette dernière sera suffisamment importante pour être détectable par les instruments à disposition de BRAHÉ*". Tycho BRAHÉ prenait dont un risque d'erreur radicale en infirmant l'héliocentrisme[45] de COPERNIC. Et d'ailleurs, par la suite, l'hypothèse anticipatrice posée par BRAHÉ a été réfutée. Prenons un dernier exemple, cité précédemment dans la section 2.3 : la confirmation expérimentale de la relativité générale d'Albert EINSTEIN par Sir Arthur EDDINGTON lors d'une éclipse de soleil le 29 mai 1919. La théorie prévoyait que les rayons lumineux en provenance d'une étoile lointaine seraient courbes à proximité du Soleil en raison des déformations de l'espace-temps induites par la masse de l'étoile. Suite à ses observations, EDDINGTON a conclu que les valeurs prédites par la théorie étaient conformes à ses observations. Cependant, même avec des instruments d'une très grande précision, il existe toujours un écart, si minime soit-il, entre la valeur prédite et la valeur mesurée. Autrement dit, la conclusion d'EDDINGTON nécessite que soit posée une hypothèse anticipatrice telle que "*les écarts entre les valeurs observées et les valeurs théoriques sont imputables uniquement aux imprécisions des instruments astronomiques utilisés*". Ces deux exemples illustrent bien le fait que les scientifiques font usage, très souvent implicitement, voire sans en être conscients, d'hypothèses anticipatrices afin de pouvoir procéder à des déductions à partir de leurs données d'observation.

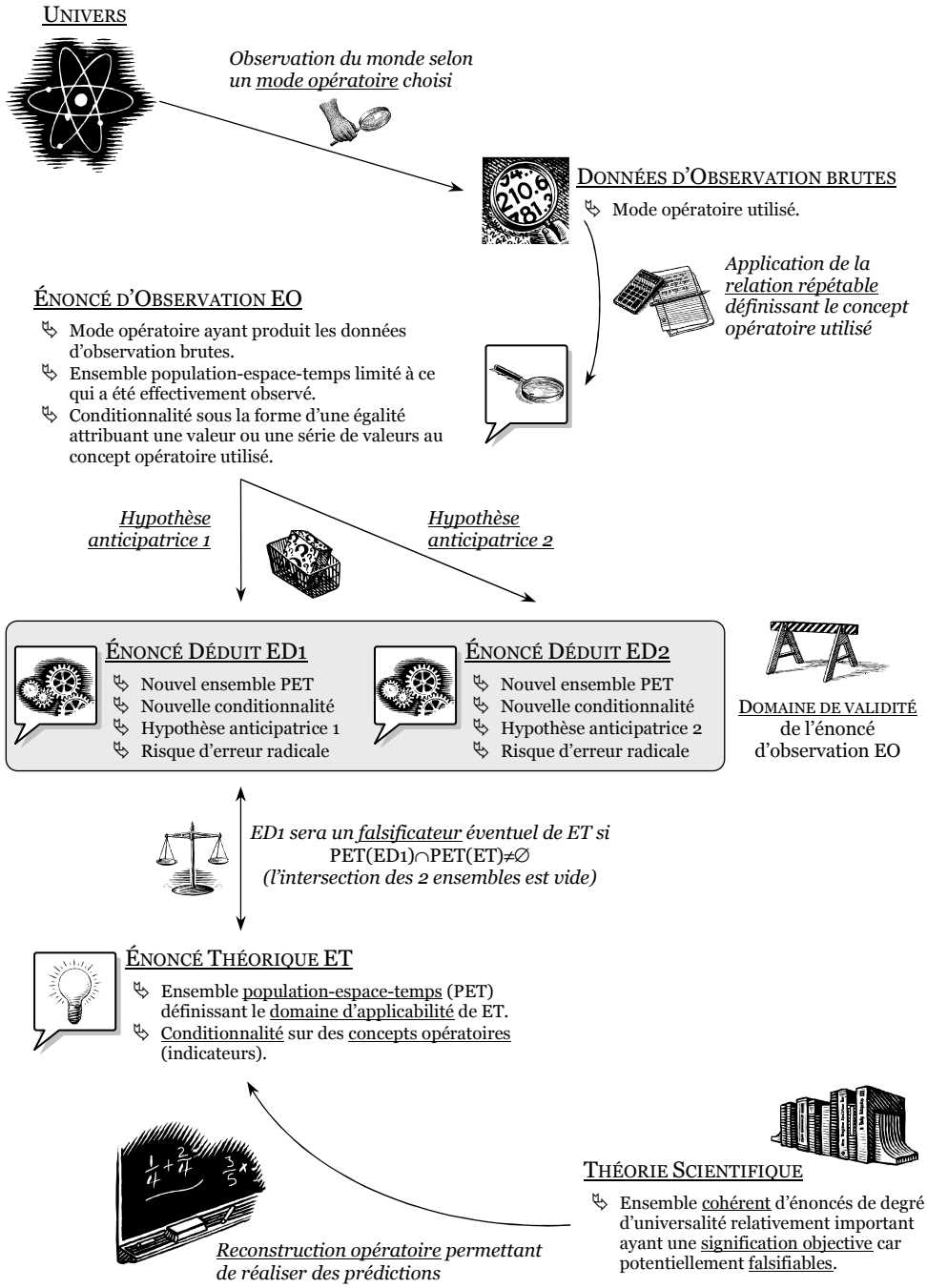
2.9 Synthèse du cadre conceptuel retenu

Nous avons retenu de Karl POPPER son critère de falsifiabilité pour déterminer si un énoncé possède ou non une signification objective. Ce sont donc les énoncés potentiellement falsifiables qui, selon nous, doivent préoccuper la science (Postulat 2.1). Cela ne signifie pas que de tels énoncés soient vrais. Simplement, il est *a priori* possible d'en déduire des prédictions qui pourront par la suite être testées empiriquement.

[45] Conception du monde plaçant le Soleil au centre de l'Univers.

Figure 2.3

Schéma synthétique des concepts épistémologiques retenus



LISTE DES POSTULATS POSÉS AU COURS DE CE CHAPITRE

Postulat 2.1
Un énoncé aura une signification objective s'il est potentiellement falsifiable.

Postulat 2.2
L'énoncé A sera plus universel que l'énoncé B si et seulement si $PET(B) \subset PET(A)$*, PET représentant l'ensemble population-espace-temps associé à tout énoncé et définissant son domaine d'applicabilité.*

Postulat 2.3
Un énoncé B ne pourra être un falsificateur de l'énoncé A qu'à la condition que $PET(B) \cap PET(A) \neq \varnothing$ *(l'intersection des ensembles population-espace-temps de A et B n'est pas vide).*

Postulat 2.4
Pour pouvoir être testé empiriquement, un énoncé doit être traduit en une conditionnalité sur des concepts opératoires ou indicateurs, la définition d'un concept opératoire devant comporter la définition d'un procédé régulier, d'une relation répétable, permettant de mesurer et d'identifier le concept ainsi défini.

Postulat 2.5
Un énoncé d'observation naît de l'application d'un concept opératoire donné à des données d'observations brutes obtenues selon un mode opératoire choisi.

Postulat 2.6
L'hypothèse anticipatrice appliquée à un énoncé d'observation doit prendre en compte l'information maximale disponible au moment du choix de la dite hypothèse et doit être limitée au minimum requis pour permettre l'estimation du ou des concepts opératoires visés.

Postulat 2.7
Le domaine de validité d'un énoncé d'observation est composé de l'ensemble des énoncés déduits du dit énoncé d'observation par application d'une hypothèse anticipatrice à ce dernier. Les énoncés déduits ne seront valables temporairement que sous réserve de l'hypothèse anticipatrice qui les a engendrés et comporteront donc un risque d'erreur radicale.

Postulat 2.8
Le risque d'erreur radicale pourra être considéré comme minimisé si l'hypothèse anticipatrice posée a été vérifiée expérimentalement pour des cas analogues et/ou si plusieurs hypothèses anticipatrices différentes, appliquées à un même énoncé d'observation, conduisent aux mêmes énoncés déduits.

Le terme "potentiellement" implique que nous pouvons considérer comme ayant une signification objective des énoncés dont les prédictions ne peuvent être, en raison de limites techniques ou historiques, vérifiées mais qui le seraient éventuellement ultérieurement si de nouvelles sources ou de nouveaux outils étaient découverts. Concernant les énoncés qui ne satisferaient pas à ce critère de falsifiabilité, nous n'en considérons pas moins qu'ils puissent être vrais ou qu'ils puissent être une source profitable de connaissances sur le monde. Nous ne posons aucun jugement de valeur sur les différentes manières de produire du savoir, ni ne considérons que la science serait la seule manière de faire. Nous posons simplement que les objets de la science, en tant que forme particulière d'acquisition de connaissances parmi d'autres, doivent avoir une signification objective telle que nous l'avons définie. Les perceptions subjectives ou l'intuition ne sont pour autant pas exclues de la pratique scientifique. Elles peuvent même la guider. Cependant, ces impressions devront être reconstruites de manière opératoire afin de devenir un objet de recherche scientifique.

À la suite de Jean Ullmo, nous avons posé que cette signification objective s'acquérait par le recours à des concepts opératoires, ces derniers étant caractérisés par un procédé régulier, une relation répétable, qui permet de « *repérer, mesurer, plus généralement atteindre et identifier le concept défini*[46] ». Si les exemples que nous avons développés ont été axés sur des concepts permettant une quantification numérique, cela n'implique pas que les concepts opératoires, usuellement appelés indicateurs en sciences sociales, ne puissent être que quantitatifs. En effet, toutes les catégorisations en sciences sociales sont des concepts opératoires, de même que les *idéal-type* de Weber, par exemple. En sciences sociales, nous parlerons plutôt de construction de l'objet de recherche et d'opérationnalisation des hypothèses plutôt que de construction opératoire, mais il s'agit néanmoins du même processus sous-jacent et chaque énoncé doit pouvoir être traduit sous la forme d'une conditionnalité sur des concepts opératoires, cette dernière pouvant être exprimée aussi bien quantitativement que qualitativement (Postulat 2.4).

Nous avons décidé de ne pas distinguer énoncés universels et énoncés singuliers. D'une part, cette distinction nous semble trop binaire, un énoncé ne pouvant être que l'un ou l'autre. D'autre part, aucune théorie scientifique ne peut se targuer d'être universelle. Si la recherche d'universalité peut être un des objectifs de la science, définir le domaine d'applicabilité d'une théorie s'avère plus opérationnel. Au-delà d'une vérité ultime sur l'essence même du monde qui nous entoure, nous préférons la notion de concordance avec le réel d'Albert Einstein. Nous concevons donc la possibilité d'un degré d'universalité plus ou moins important et posons qu'à chaque énoncé doit être associé un ensemble population-espace-temps définissant les objets sur lesquels porte le dit énoncé et contextualisant ce dernier via une

[46] Ullmo, *La Pensée scientifique moderne*, p. 25.

superficie d'espace-temps (une partie de l'espace et du temps) dans laquelle il s'inscrit. Nous avons montré qu'un énoncé pourra être "plus universel" qu'un autre si l'ensemble population-espace-temps du second est inclus dans celui du premier (Postulat 2.2). Par ailleurs, un énoncé ne pourra être un falsificateur d'un autre énoncé, sous réserve de validité du dit énoncé, qu'à la condition que l'intersection de leurs ensembles population-espace-temps soit non vide (Postulat 2.3), les assertions des deux énoncés ne pouvant être comparées qu'à cette intersection.

Pour mesurer[47] un concept opératoire donné, nous avons toujours à notre disposition plusieurs méthodes possibles, tant quantitatives (questionnaires, dispositif expérimental, sondes…) que qualitatives (observations, entretiens…). Nos données d'observation brutes seront donc obtenues selon un mode opératoire que nous aurons choisi et qui, d'une manière ou d'une autre, aura une influence sur ce que nous aurons observé. Ces données brutes ne permettent pas de mesurer directement un concept opératoire. Il faut leur appliquer la relation répétable définissant le dit concept pour obtenir un énoncé d'observation (Postulat 2.5). Comme tout énoncé, les énoncés d'observation sont définis par un ensemble population-espace-temps (correspondant en l'occurrence aux objets qui ont été effectivement observés et à la superficie d'espace-temps correspondant à la collecte) et une conditionnalité sur des concepts opératoires (sous la forme d'une égalité pour ce type d'énoncés).

Dans de nombreuses situations, en particulier dans les sciences humaines et sociales, les énoncés d'observation ne constituent pas des falsificateurs possibles des énoncés théoriques que nous cherchons à confirmer ou infirmer. Il devient alors nécessaire de procéder à une estimation, permettant de déduire, d'un énoncé d'observation, d'autres énoncés. Pour pouvoir tirer plus d'informations que celles initialement contenues dans les données d'observation, nous devons poser une hypothèse anticipatrice, en prenant en compte l'information maximale disponible et en la limitant au minimum requis pour permettre l'estimation (Postulat 2.6). Nous avons choisi d'appeler domaine de validité d'un énoncé d'observation l'ensemble des énoncés déduits en ayant recours à une hypothèse anticipatrice (Postulat 2.7). La validité de ces énoncés sera toujours temporaire, sous réserve de vérification de l'hypothèse anticipatrice. De ce fait, ces énoncés comporteront forcément un risque d'erreur radicale. Ce dernier ne peut être éliminé sans recours à d'autres données d'observations, mais peut être considéré comme minimisé dans certaines circonstances : hypothèse anticipatrice vérifiée expérimentalement pour des cas analogues, mêmes déductions obtenues à partir d'hypothèses différentes (Postulat 2.8).

[47] Nous utilisons ici le verbe "mesurer" dans un sens générique n'impliquant pas forcément une mesure quantitative mais incluant également les mesures qualitatives.

Cette présentation synthétique ainsi que le schéma récapitulatif proposé (Figure 2.3) pourraient laisser suggérer que le raisonnement scientifique est linéaire. Or, dans la pratique, il y a un va-et-vient permanent entre la construction théorique, le choix des indicateurs, la collecte des données, la construction de l'objet de recherche et l'interprétation des résultats. Simplement, le scientifique cherchera, au moment de sa démonstration, à réordonner ses réflexions pour les présenter dans un ordre logique, plus proche, en général, du schéma proposé que du déroulement chronologique effectif de sa pensée.

2.10 Espace non poppérien

L'ensemble des remarques développées ci-après mériterait des développements plus importants. Cependant, notre objectif premier ne vise pas l'élaboration d'une épistémologie complète mais à disposer de concepts opératoires pour appréhender la validité des différentes mesures de la prévalence du VIH. Il importera donc de considérer les points suivants comme des pistes de réflexion à approfondir.

Si nous avons retenu le critère de falsifiabilité, notre position n'est pas falsificationiste dans la mesure où nous avons montré la faillibilité des énoncés d'observations et la complexité des situations réelles de test. En outre, comme nous avons le plus souvent recours à des énoncés déduits pour prendre nos décisions, le risque d'erreur radicale n'est jamais écarté. Par ailleurs, plutôt que de rejeter un énoncé, il est parfois préférable d'en modifier le domaine d'applicabilité pour que ce dernier ne couvre plus l'énoncé falsifiant. Il n'y a pas de règle univoque quant aux décisions que le scientifique doit prendre. S'il peut être aisé d'arriver à un consensus concernant quelques énoncés simples, il est beaucoup plus complexe de déterminer si une théorie scientifique, dans son ensemble, doit être rejetée ou non. Les théories scientifiques ne sont pas figées. En tant qu'ensemble complexe d'énoncés, elles peuvent mettre des dizaines d'années avant d'être clairement formulées. Par ailleurs, les décisions prises par la communauté des scientifiques ne se font pas hors du temps, mais bien dans un contexte historique, social et politique. Plusieurs auteurs ont écrit sur les processus amenant au changement de paradigme scientifique : Imre LAKATOS et sa *méthodologie des programmes de recherche*, Gaston BACHELARD et la *formation de l'esprit scientifique*, Thomas KUHN et les *révolutions scientifiques* ou bien encore Paul FEYERABEND et sa *théorie anarchiste de la connaissance*[48].

[48] Pour reprendre une expression de CHALMERS A. F., *Qu'est-ce que la science ? Récents développements en philosophie des sciences : Popper, Kuhn, Lakatos, Feyerabend*, Paris (FR), La Découverte, 1976. Chapitre 12.

Les exemples qui émaillent ce chapitre ont souvent été empruntés à la physique. La majorité des écrits épistémologiques s'est focalisée sur cette discipline en particulier et il est difficile d'en faire totalement abstraction. Cependant, notre propos s'applique tout aussi bien aux sciences humaines et sociales, ce que nous avons voulu montrer au travers des autres exemples pris. Certains lecteurs auront peut-être été dérangés par le fait que nous avons illustré notre propos presque exclusivement à partir de concepts opératoires quantitatifs. Il s'agissait de ne pas trop nous écarter de notre objet de recherche. Cependant, aucun des concepts développés ici ne requiert le passage par un formalisme mathématique ou que le type de mesure soit absolument quantitatif. Il en résulte que le cadre proposé peut s'appliquer, *a priori,* aussi bien à des approches qualitatives que quantitatives.

À nos yeux, ces deux types d'approches ne diffèrent pas d'un point de vue épistémologique mais constituent simplement deux classes de modes opératoires possibles, deux formes linguistiques différentes utilisées pour exprimer nos assertions. D'ailleurs, ce serait une erreur que de croire que les approches dites quantitatives en sciences sociales pourraient être exprimées exclusivement dans un formalisme mathématique. De nombreux indicateurs relèvent d'une catégorisation qui n'est rien d'autre qu'un concept opératoire qualitatif. Il suffit de prendre l'exemple de la variable sexe. Alors que cette variable peut sembler évidente pour une majorité de personnes, il existe en réalité plusieurs concepts opératoires pour la définir et ne se recoupant pas exactement. C'est le cas par exemple entre les définitions de sexe biologique, sexe génétique et sexe social. Par ailleurs, le découpage de cette variable en deux modalités correspond à un choix arbitraire[49] : certains recensements de population, comme ceux du Népal, prévoient trois modalités. Le concept de rapport de masculinité (rapport entre le nombre d'hommes et le nombre de femmes), bien que mesuré quantitativement, repose également sur des définitions opératoires qualitatives (la notion de sexe en l'occurrence). Une majorité d'indicateurs numériques relèvent d'une taxinomie. Par ailleurs, si l'analyse quantitative a recours à des estimations numériques de ces différents indicateurs, elle procède également à une interprétation des résultats obtenus. Or, le processus interprétatif ne se limite pas à une simple conversion des chiffres en lettres, mais vise à donner une signification plus large des données chiffrées. Pour ce faire, le démographe ou le sociologue quantitativiste va supposer que les nombres qu'il a obtenus traduisent des processus, des démarches, des phénomènes décrits par ailleurs, le plus souvent qualitativement. Il a donc recours

[49] En ce sens, la théorie précède bien l'observation. Rien n'implique que la variable sexe ne puisse prendre que deux modalités, pas même des arguments prétendument naturels. L'hermaphrodisme va à l'encontre d'une définition morphologique en deux classes, tout comme la génétique. Il existe d'autres formes que XX ou XY, telles que XXY rencontrée dans le syndrome de KLINEFELTER. Si nous découpons cette variable en deux modalités, il s'agira alors d'un choix méthodologique et non d'une évidence qui s'imposerait naturellement à nous.

à des données additionnelles afin d'expliciter ces chiffres. Il ne pose ni plus ni moins qu'une nouvelle hypothèse anticipatrice. Ainsi, l'interprétation des données quantitatives en sciences sociales procède le plus souvent par l'adoption d'une première hypothèse anticipatrice quantitative puis l'adoption d'une seconde hypothèse anticipatrice, qualitative cette fois[50].

À la suite de George MATHERON, qui a largement inspiré nos choix, la position que nous adoptons ne peut se réduire à une position « *conciliatrice*[51] » entre objectivistes et subjectivistes. Si nous retenons des subjectivistes l'importance qu'ils accordent à l'information disponible "*a priori*" ou aux choix posés par le chercheur, nous ne pensons pas que toute conclusion découle de l'arbitraire de la subjectivité individuelle. Dans une lignée positiviste, nous pensons que le monde est constitué d'entités réelles qui existent au-delà de nos perceptions et qui possèdent des caractéristiques indépendantes des concepts que nous utilisons pour les décrire. Par contre, les observations que nous faisons du monde réel restent relatives et dépendantes des théories et des concepts que nous aurons définis au préalable. Cela ne signifie pas pour autant, selon nous, l'impossibilité de produire un savoir réel sur le monde qui nous entoure, puisque nous avons retenu d'EINSTEIN que la vérité d'une assertion provenait de sa concordance avec le réel. Mais, nous rejetons un positivisme naïf en ce sens que nous croyons vain de pouvoir donner un fondement certain et définitif à une connaissance universelle du monde. Nous prenons toujours le risque d'une erreur radicale. Cela ne nous empêche pas, néanmoins, de pouvoir prendre des décisions ni même de mettre à jour des connaissances solides. Simplement, nous devons garder à l'esprit que nos conclusions sont toujours temporaires et certaines de nos théories peuvent se révéler erronées ou bien encore plus limitées et plus approximatives que nous ne le pensions.

Nous avons évoqué un peu plus haut que, pour nous, il n'y avait pas de différence fondamentale, d'un point de vue épistémologique, entre les sciences humaines et sociales et les sciences de la nature telle que la Physique. Certains auteurs se sont attachés à montrer, au contraire, que des sciences telles que la sociologie relevait d'une méthode profondément distincte. Nous pensons en particulier à l'ouvrage de Jean-Claude PASSERON intitulé *Le Raisonnement sociologique : un espace non poppérien de l'argumentation*[52]. PASSERON distingue « *sciences expérimentales* »

[50] En effet, si nous avons décrit dans la section 2.8 le processus permettant de déduire d'un énoncé d'observation un nouvel énoncé par l'adoption d'une hypothèse anticipatrice, il reste possible d'obtenir un nouvel énoncé de cet énoncé déduit par l'adoption d'une seconde hypothèse, cette dernière venant s'ajouter à la première.

[51] MATHERON, *Estimer et Choisir*, p. 3.

[52] PASSERON J.-C., *Le Raisonnement sociologique, un espace non poppérien de l'argumentation*, Paris (FR), Albin Michel, 2006.

et *« sciences historiques* ». Les premières relèveraient de l'expérimentation au sens strict leur permettant de pouvoir corroborer ou falsifier leurs théories, tandis que pour les secondes, « *la mise à l'épreuve empirique d'une proposition théorique ne peut jamais revêtir en sociologie la forme logique de la réfutation ("falsification") au sens poppérien*[53] ». Dès lors, le raisonnement sociologie procède d'un raisonnement comparatif.

> « *Le raisonnement sociologique ne reste un raisonnement scientifique que dans la mesure où il s'astreint à ne transformer par l'interprétation conceptuelle ce qu'énonce un constat statistique qu'en s'interdisant la référence à d'autres constats que des constats empiriques. Mais le raisonnement sociologique se distingue du raisonnement expérimental en ce que, dans son argumentation multi-référentielle, il doit composer des constats qui ne sont pas cumulables entre eux, au sens strict de la combinatoire logique, mais qui restent sémantiquement assez "apparentés" pour que le raisonnement naturel puisse contrôler cette parenté. Le contrôle de la parenté des contextes est d'autant plus sûr qu'il s'appuie sur une méthodologie du raisonnement naturel qui n'est autre que celle du raisonnement comparatif, nécessaire à la construction des concepts typologiques.* » (PASSERON 2006, p. 206)

Les sciences historiques ont ceci de particulier, pour PASSERON, que leurs propos ne peuvent jamais être totalement extraits du contexte social et historique dans lequel leurs objets de recherche s'inscrivent. Le sociologue se doit alors de tenir compte de « *la forme contextuelle et circonstanciée de connaissance des faits sociaux qu'exclut – dans son principe même – l'obligation d'énumérer toutes les "conditions initiales" d'une "expérimentation proprement dite"* »[54]. « *Le contexte d'une mesure ou d'une observation portant sur le monde historique ne peut être épuisé par une série finie d'assertions qui énonceraient les traits pertinents du contexte pour la validité de la mesure ou de l'observation considérée.* »[55] Pour transformer une collection d'informations en un énoncé de base, le sociologue doit ses observations à raisonnement guidé par une problématique et recourir à des concepts pour produire des énoncés scientifiques ayant la forme logique d'un « *effet de connaissance* » (PASSERON 2006, p. 370).

> « *À ce niveau de formulation des propositions, les opérations accessibles à une discipline de l'observation historique sont logiquement les mêmes que celles que pratique une science expérimentale. Tant que l'énonciation n'entreprend pas de faire dire à de tels effets de*

53 PASSERON, *Le Raisonnement sociologique*, p. 542.

54 PASSERON, *Le Raisonnement sociologique*, p. 33.

55 PASSERON, *Le Raisonnement sociologique*, p. 558.

> *connaissance autre chose ou plus que ce qu'ils signifient descriptivement dans leur contexte, c'est-à-dire tant qu'elle n'entreprend pas de faire varier la généralité empirique dont elle crédite ses assertions en diminuant la précision des contraintes spatio-temporelles qui les indexent sur un contexte, constats, concepts et preuves ne revêtent pas en sociologie ou dans les sciences sociales une forme logique spécifique. La logique formelle des catégorisations, des opérations et des raisonnements, c'est-à-dire du calcul (logique ou quantitatif) est ici la même qu'ailleurs. C'est cette proximité entre une démarche accessible à toute science sociale et la démarche de base de la méthode expérimentale sur laquelle s'appuyait DURKHEIM pour définir la sociologie comme une "science expérimentale des faits sociaux"*[56]*. La spécificité de l'observation historique, qui sépare les sciences sociales des sciences expérimentales ou formelles, ne se fait sentir qu'à un niveau supérieur à celui des effets de connaissance, c'est-à-dire au niveau des effets d'intelligibilité. Ici seulement les opérations théoriques de la sociologie deviennent des opérations comparatives qui visent, en rapprochant des effets de connaissance de contextes différents, à formuler des généralités historiques dont la signification conceptuelle et la validation empirique cessent d'avoir le sens formellement univoque qu'elles ne pourraient tenir que de l'indexation sur un contexte constant ou analysable par une liste finie de variables : il devient alors un raisonnement de type expérimental mais sans expérimentation possible. […]*
>
> *L'effet d'intelligibilité que produit l'énonciation de vastes séries d'effets de connaissance dans une langue conceptuelle unifiée diffère, en effet, radicalement en sociologie, où il doit son ressort le plus significatif à la comparaison (reposant elle-même sur des conceptualisations analogiques), de ce qu'il est dans les sciences de l'expérimentation auxquelles l'universalité des "lois naturelles" et, mieux encore, l'articulation déductive de leurs énoncés universels dans un paradigme théorique procurent une définition épistémologiquement satisfaisante de la compréhension théorique du monde (même si elle se prête à des gloses philosophiques divergentes).* » (PASSERON 2006, p. 372-374)

L'interprétation, et donc « *l'énonciation sociologique* », requiert une formulation conceptuelle des données de base et la mise en correspondance de divers effets de connaissance pour obtenir des énoncés ayant une portée plus générale (PASSERON 2006, p. 211). En procédant ainsi, le sociologue fait un pari sur la possibilité de

[56] En référence au chapitre 6 de DURKHEIM É., *Les règles de la méthode sociologique*, Paris (FR), Félix Alcan, 1895.

généraliser ses effets de connaissance à des contextes plus vastes. En ce sens, nous pouvons considérer qu'il prend un risque.

> « *Seul le raisonnement sociologique peut restreindre le degré de liberté interprétative du pari fait sur la généralité de l'assertion en faisant appel à des comparaisons enfermant plus d'informations que les constats opérés dans le même contexte et les mêmes conditions d'enquête. […] C'est la conceptualisation historique qui autorise dans un raisonnement sociologique des rapprochements typologiques de plus grande ampleur et d'une plus grande richesse informative que la conjonction logique des énoncés. Le contrôle du rapprochement entre assertions conceptuellement "apparentés" ne relève pas d'une logique formelle mais d'une méthodologie de la comparaison.* »
>
> (PASSERON 2006, p. 213)

Nous sommes globalement en accord avec ce que PASSERON écrit à propos des sciences historiques. D'ailleurs, le cadre conceptuel que nous avons décrit dans ce chapitre correspond relativement bien à ce qu'il nomme le raisonnement sociologique. En effet, tous nos énoncés sont indissociablement liés à un ensemble population-espace-temps et, en ce sens, possèdent une dimension historique et contextuelle. Nous avons montré par ailleurs qu'il ne pouvait y avoir de réfutation formelle de ces derniers dans la mesure où les énoncés d'observation sont faillibles et où le chercheur prend toujours un risque d'erreur radicale. L'effet de connaissance de PASSERON qui permet, grâce au recours à des concepts, de produire des énoncés scientifiques correspond à ce que nous avons appelé l'application de la relation répétable définissant les concepts opératoires pour transformer des données brutes en énoncés d'observation, ces derniers pouvant être utilisés dans le cadre d'un raisonnement logique formel. Le « *pari* » que pose le sociologue pour réaliser des assertions plus générales en mettant en relation des énoncés de base s'apparente aux hypothèses anticipatrices permettant à partir d'un ou de plusieurs énoncés d'observation d'en extraire des énoncés déduits de portée plus générale. Le « *risque interprétatif* » de PASSERON renvoie, quant à lui, à notre risque d'erreur radicale, mais également à des choix inappropriés d'hypothèses anticipatrices.

En revanace, nous ne partageons pas la position de PASSERON vis-à-vis de ce qu'il nomme les sciences expérimentales, dans la mesure où il leur reconnaît la possibilité de mettre à jour des "lois universelles" et de pouvoir procéder au dénombrement exhaustif de l'ensemble des conditions initiales d'une expérience. Dans la section 2.3, nous nous sommes efforcés de montrer que toute théorie était forcément limitée et donc nous avons posé dans la section 2.4 que tout énoncé devait être contextualisé en lui associant un ensemble population-espace-temps. Nous reprenons par ailleurs à notre compte ces propos de MATHERON :

> « *C'est toujours en un sens relatif que l'on parle de refaire la "même" expérience. À strictement parler, il n'y a pas deux expériences*

> *identiques : elles diffèrent toujours l'une de l'autre par quelques facteurs accessoires (mais c'est nous qui les jugeons tels), et par les conditions de lieu ou de temps. Tout ce que l'on peut dire, c'est que les facteurs qui nous paraissent importants, ont été rendus aussi semblables que nous le permettent nos moyens techniques, et les autres sont ce qu'ils sont.* »
>
> (MATHERON 1978, p. 55-57)

Si la thèse des *Préalables épistémologiques*[57] peut se résumer selon PASSERON[58] par : « *la Sociologie est une science comme les autres, qui a seulement plus de difficultés que les autres à être une science comme les autres* », nous pouvons résumer la thèse du *Raisonnement sociologique* par "*les sciences historiques sont autant des sciences que les autres, mais leur méthode, le raisonnement comparatif, est spécifique*". Nous considérons, pour notre part, que "*la physique est une science historique comme toutes les autres*".

L'expérimentation au sens strict, d'un point de vue formel, demeure inatteignable, une expérience ne pouvant se répéter dans des conditions parfaitement identiques en tout point. L'expérimentation n'est donc qu'une classe de modes opératoires parmi d'autres. L'avantage de disciplines telles que la physique réside, selon nous, dans leur capacité à pouvoir user du principe de *négligeabibilité*. Dans la *Formation de l'esprit scientifique*, Gaston BACHELARD précise que « *le savant croit au réalisme de la mesure plus qu'à la réalité de l'objet* »[59]. Derrière cela, BACHELARD montre que pour appréhender toute mesure il importe de tenir compte de l'*ordre de grandeur* sur lequel le chercheur travaille ainsi que la *précision* des données collectées. Un excès de précision s'apparente à une meute de chiffres inutiles, une des différences entre les mathématiques et la physique étant de déterminer quelles décimales sont porteuses de sens. Dès que l'on tient compte de l'ordre de grandeur des observations effectuées, « *l'objet peut alors changer de nature quand on change le degré d'approximation* »[60]. Les objets de recherche de la physique appartiennent à des ordres de grandeur tellement différents que bon nombre de phénomènes perturbateurs deviennent invisibles. Si nous étudions le mouvement de deux boules de billards, nous devrions tenir compte de l'ensemble des conditions contextuelles qui président le phénomène que nous analysons. En toute rigueur, nous devrions considérer la couleur des boules de billard car, en reflétant la lumière différemment, l'impact des photons sur le mouvement de ces dernières ne sera pas le même. Cependant, l'effet induit par la lumière sur le

[57] BOURDIEU P., CHAMBOREDON J.-C. et PASSERON J.-C., *Le Métier de sociologue : préalables épistémologiques*, Paris-La Haye (FR), Mouton, 1973.

[58] PASSERON, *Le Raisonnement sociologique*, p. 65-66.

[59] BACHELARD G., *La Formation de l'esprit scientifique : contribution à une psychanalyse de la connaissance*, Paris (FR), J. Vrin, 1938, p. 254.

[60] *Idem*.

mouvement des boules de billard est tellement faible qu'il est quasiment indétectable. Le physicien pourra donc se permettre de négliger la couleur des boules dans son étude. Nous avons vu à la section 2.3 que la théorie newtonienne était parfaitement valable à nos échelles terrestres. Cela ne signifie pas pour autant que les phénomènes que nous observons à cette échelle ne subissent pas les fluctuations quantiques ou les déformations de l'espace-temps. Cependant, les manifestations de la courbure de l'espace-temps prédites par la Relativité Générale ne commencent à être vraiment sensibles que pour des vitesses proches de celle de la lumière (notée c). Quant aux fluctuations quantiques, elles ne deviennent manifestes qu'à des échelles microscopiques se rapprochant de la longueur de Planck (notée l_P). Or, ces deux grandeurs physiques se traduisent à notre échelle, c'est-à-dire dans nos unités de mesure usuelles, par des valeurs extrêmes : c vaut 299 792 458 mètres par seconde soit 1 079 252 848,8 kilomètres par heure ($1{,}07\times10^{9}$) et la valeur approchée de l_P est de $1{,}616\times10^{-35}$ mètre. Il a été possible de tester empiriquement la relativité générale et la mécanique quantique parce que, tandis que l'une s'applique aux très grandes distances, l'autre correspond au monde subatomique. À cette échelle, la force gravitationnelle peut être tout bonnement négligée dans la mesure où elle est 10^{44} fois plus faible que la force nucléaire forte. À l'échelle des galaxies, la force gravitationnelle prédomine, les forces nucléaires s'atténuant très rapidement à des échelles supérieures à la taille d'un atome[61]. C'est en raison de cette capacité de pouvoir *négliger* un grand nombre de variables que les expériences menées par le physicien semblent pouvoir être répétées à l'identique. L'expérimentation est également facilitée par la possibilité de *simplifier* les objets de recherche. Par là, nous entendons qu'il est possible de réduire de manière drastique le nombre d'objets en interaction ainsi que le nombre de variables influant sur ces dernières[62].

Le sociologue et l'historien, quant à eux, n'ont guère la possibilité de *simplifier* autant leurs objets de recherche ou bien de réduire significativement le nombre de variables ayant un effet observable sur les phénomènes qu'ils étudient. Certes, il est possible d'avoir recours à des modèles sociaux simples (y compris d'un point de vue quantitatif, c'est ce que sont les modèles démographiques par exemple) intégrant un nombre plus ou moins restreint de variables. Cependant, les modèles que l'on obtient ainsi s'avèrent toujours être relativement approximatifs, en tout cas largement plus que les modèles physiques, au point que leur imprécision soit observable en cas de vérification empirique. Par ailleurs, des raisons techniques,

[61] La force électromagnétique ne s'atténue pas à ces distances. Mais, comme la matière est globalement neutre, les interactions électromagnétiques se compensent aux grandes échelles.

[62] Cela ne signifie pas pour autant que les phénomènes étudiés soient simples, ni même que les équations rencontrées par le physicien ne soient pas d'une redoutable complexité mathématique. Mais l'objet de recherche du physicien peut être considéré comme relativement simple dans la mesure où il porte sur un nombre de variables et d'objets en interaction limité.

éthiques et pratiques interdisent de procéder à des expériences à l'échelle de sociétés. Nous ne pouvons pas déclencher une guerre entre la Russie et les États-Unis juste pour vérifier si nos théories géopolitiques sont exactes. Tout comme d'autres disciplines telles que l'astrologie, nous devons nous contenter d'observer les phénomènes tels qu'ils se présentent[63].

Là encore les sciences humaines et sociales auront un inconvénient sur d'autres disciplines. Les échelles, temporelles et géographiques, sur lesquelles elles travaillent, sont le plus souvent du même ordre de grandeur que nos propres échelles de vie. De ce fait, les phénomènes qui nous préoccupent s'avèrent évoluer au rythme même où nos effectuons nos recherches, si ce n'est plus rapidement encore. Lorsque les phénomènes étudiés se déroulent sur des échelles de temps longues, relativement à nous, il est alors possible de réitérer des observations qui, si elles ne seront pas parfaitement identiques, pourront être considérées comme relativement équivalentes. Il est ainsi possible, *a priori*, de réaliser aujourd'hui certaines observations astronomiques analogues à d'autres qui auraient été faites il y a cinquante ou cent ans, lorsqu'elles portent sur des processus qui évoluent suffisamment lentement pour présenter un caractère quasi-constant sur un siècle. Par contre, lorsque l'objet de recherche est une société, cette dernière évolue généralement trop rapidement pour que des observations réalisées aujourd'hui soient encore équivalentes à des observations plus anciennes. Le contexte social français en ce début de XXIe siècle n'est en rien analogue à celui qui précéda la Première Guerre Mondiale.

⁂

Au final, il n'y a, pour nous, aucune différence fondamentale, d'un point de vue épistémologique, entre les sciences dites expérimentales et les sciences historiques, les premières s'inscrivant également dans un contexte historique. Elles se distinguent quant aux ordres de grandeurs spatio-temporelles dans lesquels elles s'inscrivent, par la possibilité de simplifier leurs objets de recherche, celle d'appliquer le critère de négligeabilité et enfin par le degré d'universalité que leurs théories ont réussi à atteindre. Cela pourrait faire croire que certaines disciplines auraient accès à des lois naturelles universelles et à une connaissance exhaustive du contexte dans lequel s'effectuent leurs observations. Il n'en est rien.

[63] Mon propos est ici quelque peu simplificateur, dans la mesure où la recherche dans les sciences sociales ne se situe pas hors du monde. Les résultats produits par cette discipline font eux-mêmes partie des variables qui influent sur le devenir des sociétés. Il n'y a qu'à voir, à titre d'exemple, le nombre d'experts dans les milieux décisionnels et la manière dont certains résultats, dits scientifiques (qu'ils le soient ou non), sont utilisés pour influencer ou justifier des choix politiques.

Chapitre 3
Représentativité et biais

La prévalence du VIH peut se définir comme la proportion, au sein d'une population donnée, à un moment donné, de personnes infectées par le VIH. Cette définition a recours à deux concepts distincts : les notions de proportion et de personne infectée. Cela implique deux étapes dans toute opération de quantification de la prévalence : identifier la population étudiée et sélectionner les personnes à tester, d'une part, et déterminer pour chacune d'elles, à partir d'un prélèvement sanguin ou salivaire, si elle est infectée par le VIH. La qualité de notre mesure finale dépendra de ces deux étapes.

Dans un premier temps, nous aborderons les limites propres aux tests de dépistage utilisés pour déterminer le statut des personnes vis-à-vis de l'infection (section 3.1). Puis, nous développerons la notion de représentativité (section3.2) à laquelle nous avons implicitement recours dès lors que, à défaut d'un recrutement exhaustif de la population étudiée pour des raisons pratiques, financières ou éthiques, nous procédons à une observation partielle (en réalisant un échantillonnage par exemple). C'est parce que nous supposons que nos échantillons sont représentatifs que nous pouvons alors poser des hypothèses anticipatrices relevant de la théorie des sondages. Nous verrons alors que la représentativité est fortement liée à la notion de biais. Nous analyserons les Enquêtes Démographiques et de Santé et les AIDS Impact Surveys (section 3.3). Il s'agit d'enquêtes auprès des ménages et, de ce fait, une petite partie de la population générale n'est pas sélectionnée. Par ailleurs, les taux de personnes non testées pour le VIH (refus ou absence) y sont non négligeables. Dans quelle mesure ce biais influe-t-il sur les estimations nationales ?

Nous aborderons ensuite la représentativité de la surveillance sentinelle des femmes enceintes consultant en clinique prénatale en deux temps (section 3.4). À un niveau local, ces dernières sont-elles représentatives de l'ensemble des femmes et plus largement de la population générale ? À l'échelle nationale, les cliniques sentinelles retenues sont-elles représentatives de l'ensemble de la population ? Cela nous amènera à comparer la représentativité de ces deux sources de données et à soulever les limites d'une comparaison directe entre leurs estimations réciproques (section 3.4).

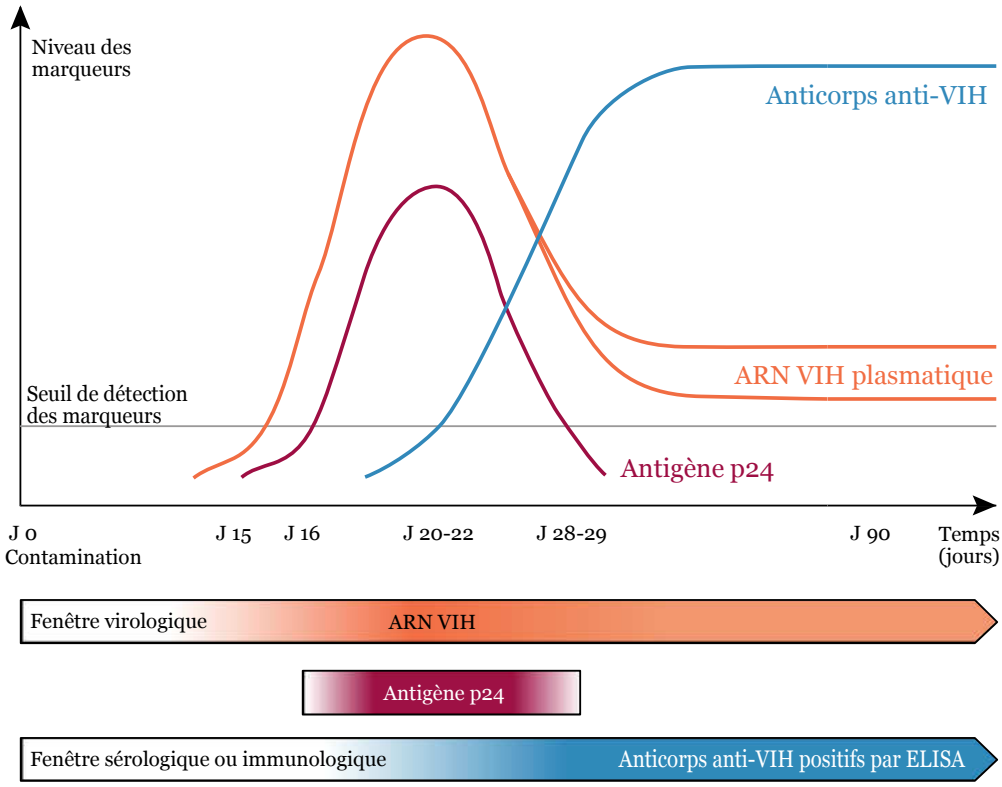
3.1 Tests de dépistage

3.1.1 Infection à VIH et statut sérologique

Dans le langage courant, nous employons régulièrement comme des synonymes les expressions "personne infectée par le VIH" et "personne séropositive au VIH". La sérologie désigne en fait une procédure visant à déterminer la réponse immunitaire d'un individu, face à un agent pathogène, en recherchant les anticorps spécifiques de ce pathogène développés par l'organisme. En règle générale, cette réponse immunitaire permet d'éradiquer l'infection en question. Cependant, dans le cas du VIH, elle est insuffisante pour supprimer totalement le virus. La majorité des tests de dépistage utilisés vise à détecter les anticorps anti-VIH développés par l'organisme en réponse à l'infection. Une fois produits par la réponse immune, les anticorps anti-VIH persisteront toute la vie de la personne infectée (PLANTIER 2002). Cependant, suite à une contamination, dans la période appelée *primo-infection*, il faut un certain temps avant que ne se développent les anticorps anti-VIH (voir Figure 3.1) et ne se produise la *séroconversion* de la personne, c'est-à-dire le passage d'une sérologie négative à une sérologie positive. Ce laps de temps induit l'existence d'une *fenêtre sérologique* pendant laquelle, bien que l'individu soit infecté, il ne sera pas dépisté séropositif au VIH. Il n'y a donc pas concordance exacte entre infection à VIH et sérologie au VIH.

> « *Les premiers anticorps sont détectables en moyenne vers le 21e jour mais le délai d'apparition des anticorps après le contact infectant peut varier de 3 semaines à 3 mois. Cette cinétique peut varier en fonction de chaque patient et aussi de la souche infectante. […] Actuellement, les tests de dépistage sont le plus souvent capables de détecter, en plus des anticorps, simultanément, la fraction "antigène p24". L'utilisation de ces tests raccourcit donc la période de "silence" sérologique […] lors de la primo-infection.* »
>
> (CRITON 2007, p. 9)

Figure 3.1
Évolution des marqueurs de la contamination du VIH



Source : (BAROUILLET 2005).

D'autres techniques de dépistage existent. Certaines reposent sur la détection du matériel génétique du virus et, plus précisément, sur la présence d'ARN plasmique. La quantité d'ARN augmente avec la multiplication du virus au cours de la primo-infection. Comme pour les anticorps, il existe une *fenêtre virologique* pendant laquelle la quantité d'ARN est insuffisante pour être détectée. Cependant, cette fenêtre virologique est plus courte que la fenêtre sérologique. Mais la détection de l'ARN présente d'autres inconvénients. D'une part, cette dernière nécessite un plateau technique plus important, pas toujours disponible dans les laboratoires africains, et, d'autre part, son coût est plus élevé que la recherche d'anticorps. Par ailleurs, chez les personnes sous traitement, l'efficacité de ce dernier induit une diminution du taux d'ARN telle que la présence d'ARN devient indétectable par nos outils actuels. Les techniques reposant sur la recherche d'ARN plasmatique s'avèrent donc inadéquates dans une optique de surveillance de la population. Un autre marqueur, l'antigène p24, s'avère pertinent pour détecter une primo-infection mais, comme nous le montre la Figure 3.1, ce dernier n'est détectable que pendant une courte période qui suit l'infection. Il n'est donc pas à retenir dans une optique de surveillance, à moins d'être intégré à un test de recherche d'anticorps.

Les enquêtes épidémiologiques ont donc généralement recours à des techniques de dépistage portant sur la recherche d'anticorps anti-VIH, ce marqueur s'avérant le plus approprié pour ce type d'études tant que ces dernières portent sur la population adulte. En effet, concernant les enfants infectés par leur mère au cours de la grossesse, de l'accouchement ou de l'allaitement, le dépistage des anticorps anti-VIH est inapproprié car les anticorps de la mère peuvent rester présents dans le sang de l'enfant pendant 18 mois (OMS 2004).

3.1.2 Fenêtre sérologique

Dans le cas du VIH, les anticorps apparaissent dans une période de un à trois mois suivant l'exposition et l'infection à VIH (PARK 2005). La variabilité de la période de séroconversion implique que, dans une optique de diagnostic, la sérologie négative d'un individu ne sera confirmée qu'après une période de trois mois après l'épisode à risques (BAROUILLET 2005). Cependant, dans le cadre d'une surveillance épidémiologique, il n'est pas possible de procéder à un second prélèvement (salivaire ou sanguin) quelques semaines après le premier[1]. L'existence de cette fenêtre sérologique implique donc qu'une partie des personnes infectées ne seront pas détectées positives.

Des progrès considérables ont été accomplis pour améliorer les performances des tests depuis l'introduction des premiers tests de dépistage des anticorps en 1985 (BARRETT 1986). Plusieurs améliorations ont été apportées au fil des années en passant des protéines virales (1e génération) aux peptides et protéines recombinantes (2e génération), de l'ELISA[2] indirect à l'ELISA sandwich (3e génération) (ZAAIJER 1992). Cependant, ces tests de troisième génération restaient limités par la durée d'apparition des anticorps de trois semaines environ après la contamination (LAPERCHE 2002). Des tests de quatrième génération, basés sur la détection simultanée des anticorps anti-VIH1, anti-VIH2 et de l'antigène p24 du VIH1, ont permis de réduire encore la fenêtre où l'infection à VIH n'est pas détectable (GURTLER 1998, WEBER 1998). Entre 1985 et 1990, la durée moyenne de la fenêtre sérologique était estimée à 45 jours avec un intervalle de confiance à 95 % situé entre 34 et 55 jours (PETERSEN 1994). Au milieu des années 1990, cette fenêtre a été réduite à une moyenne de 25 jours avec un intervalle de confiance à 95 % entre 9 et 41 jours (BUSCH 1995). Les tests de quatrième génération permettent de réduire encore cette fenêtre de 4 à 8 jours (GURTLER 1998, WEBER 1998).

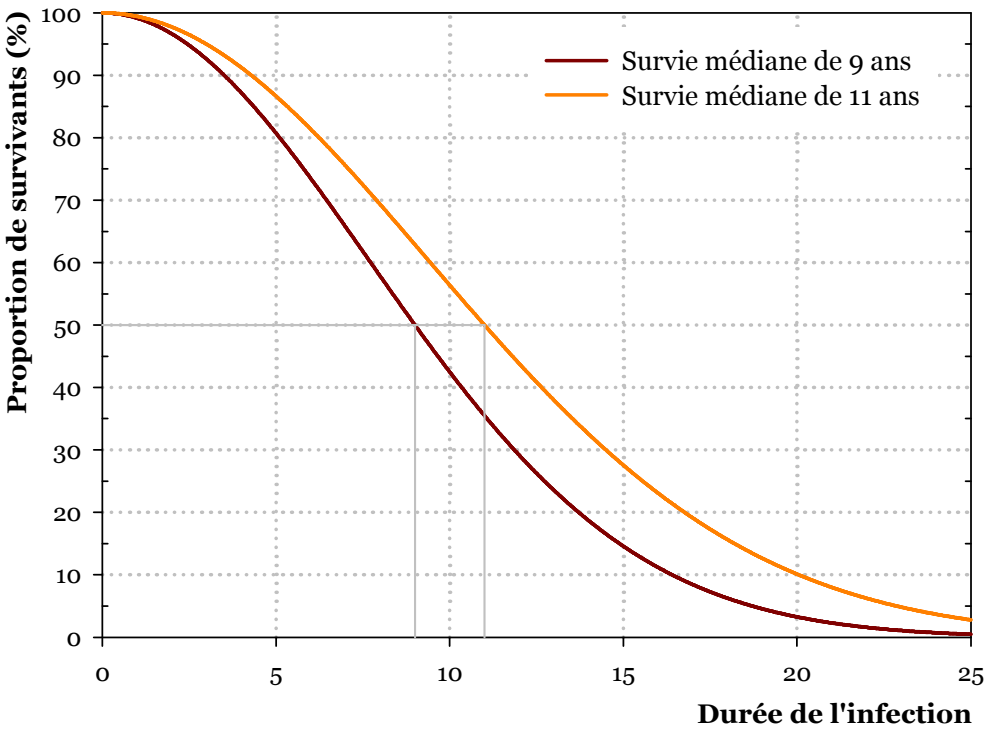
[1] Ce qui sera par contre recommandé dans le cadre d'une démarche de diagnostique.

[2] *Enzyme linked immunosorbent assay*, technique biochimique de détection des anticorps.

Nous allons essayer d'estimer la proportion de personnes, infectées par le VIH, testées négatives car situées dans la fenêtre sérologique. Pour cela, nous allons supposer que nous nous situons dans le cadre d'une épidémie stationnaire, ce qui signifie une incidence et une mortalité constantes. Par ailleurs, nous allons modéliser la mortalité des personnes infectées, en fonction de la durée d'infection, par une courbe de Weibull, conformément aux recommandations du Groupe de Référence d'ONUSIDA (THE UNAIDS REFERENCE GROUP ON ESTIMATES MODELLING AND PROJECTIONS 2002). En 2002, le Groupe préconisait d'utiliser une durée de survie médiane de 9 années, avec une marge d'incertitude allant de 8 à 11 ans. Fin 2006, suite à la publication de nouvelles données portant sur l'Afrique du Sud (GLYNN 2007), la Tanzanie (ZABA 2006) et le Rwanda (PETERS 2006), le Groupe de Référence a recommandé de prendre en compte une durée médiane de 11 ans (THE UNAIDS REFERENCE GROUP ON ESTIMATES MODELLING AND PROJECTIONS 2006a). Ces deux hypothèses sont représentées sur la figure ci-dessous (voir l'annexe 2 pour plus de détails).

Figure 3.2

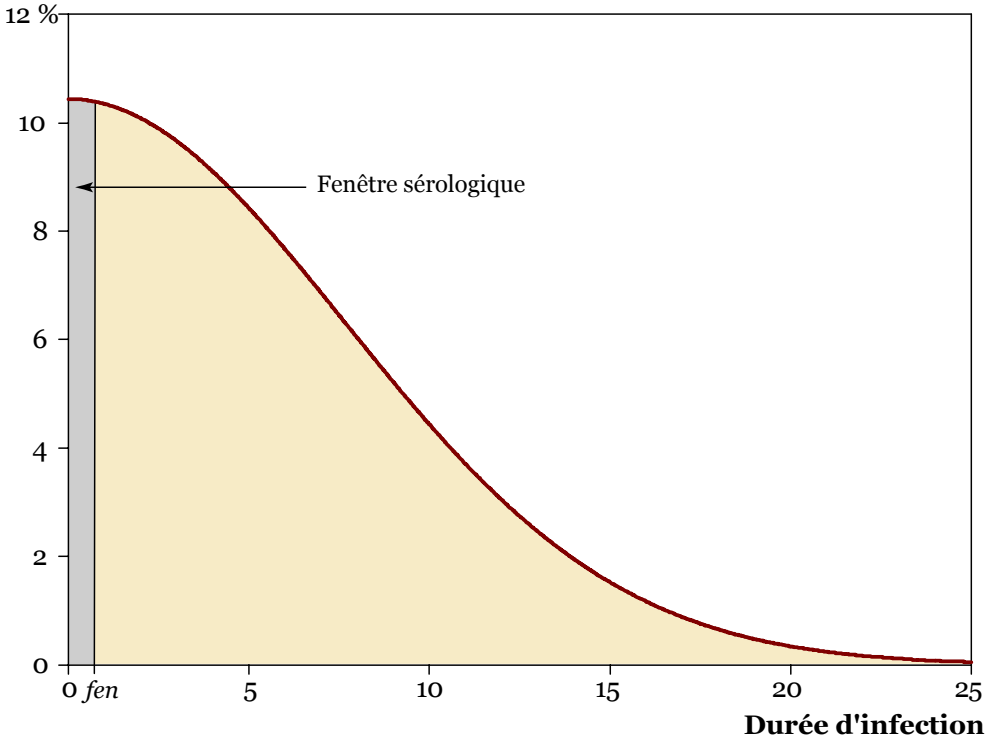
Courbes de survie en fonction de la durée d'infection en années selon deux hypothèses de durée médiane de survie



Sous l'hypothèse d'une incidence constante, la distribution des personnes infectées en fonction de la durée d'infection est égale à la répartition des décès, moyennant un facteur d'échelle afin que la superficie de la surface située sous la courbe soit égale à 1. Les personnes infectées par le VIH se répartissent alors selon la courbe de

la Figure 3.3. La proportion de personnes infectées non observables, car situées dans la fenêtre sérologique, correspond alors à la superficie de la partie grisée, située sous la courbe (pour le détail de la méthode de calcul, voir annexe 2).

Figure 3.3
Fenêtre sérologique et distribution des personnes infectées en fonction de la durée d'infection sous l'hypothèse d'une incidence constante.



Pour le calcul du risque résiduel dans le domaine du don de sang, la majorité des études utilise, pour la durée de la fenêtre sérologique, la valeur de 22 jours (BUSCH 1995, SCHREIBER 1996, BUSCH 2000), avec des marges d'incertitudes entre 6 et 38 jours. C'est le cas d'analyses menées à la fois sur des pays occidentaux comme la France (PILLONEL 2002), l'Italie (VELATI 2002), l'Espagne (ALVAREZ 2002), l'Australie ou les États-Unis (GLYNN 2002), mais également sur des pays africains comme la Guinée (LOUA 2004), la Côte d'Ivoire (OUATTARA 2006) ou l'Afrique du Sud (FANG 2003), ou bien encore en Chine (SHANG 2007). Concernant les tests de recherche d'anticorps anti-VIH de quatrième génération (incluant la recherche de l'antigène p24), la fenêtre sérologique est estimée à 17 jours (BUSCH 2005).

Nous retiendrons donc comme hypothèse centrale de durée de la fenêtre sérologique la valeur de 22 jours. Notre hypothèse haute portera sur 38 jours et notre hypothèse basse sur 17 jours. Les résultats obtenus ont été portés dans le Tableau 3.1.

Tableau 3.1
Proportion de personnes infectées situées dans la fenêtre sérologique selon plusieurs hypothèses de survie et de durée de la fenêtre sérologique

Fenêtre sérologique	Durée médiane de survie	
	9 ans	11 ans
17 jours	0,49 %	0,40 %
22 jours	0,63 %	0,51 %
38 jours	1,09 %	0,89 %

Ce modèle simple présuppose, pour être vérifié, une incidence constante depuis au moins 25 ans. Or, les différentes épidémies africaines ont débuté vraisemblablement entre 1979 et 1982. De ce fait, le temps que l'épidémie se développe, les nouvelles infections ont été plus nombreuses au cours des dix dernières années que pendant la décennie 1980. Le modèle simple représente donc une estimation *a minima* de la proportion de personnes infectées non observables égale à la valeur vers laquelle une épidémie stabilisée devrait tendre.

Au démarrage d'une épidémie, cette proportion devrait *a priori* diminuer dans la mesure où se constitue une population de personnes anciennement infectées. Par la suite, une fois l'épidémie développée et les premières générations de personnes infectées décédées en partie, une incidence croissante devrait induire, à mortalité constante, une augmentation de la proportion d'individus infectés non observables. À incidence constante, une diminution de la mortalité (ce qui correspond à une augmentation de la durée médiane de survie) induit une réduction de cette proportion.

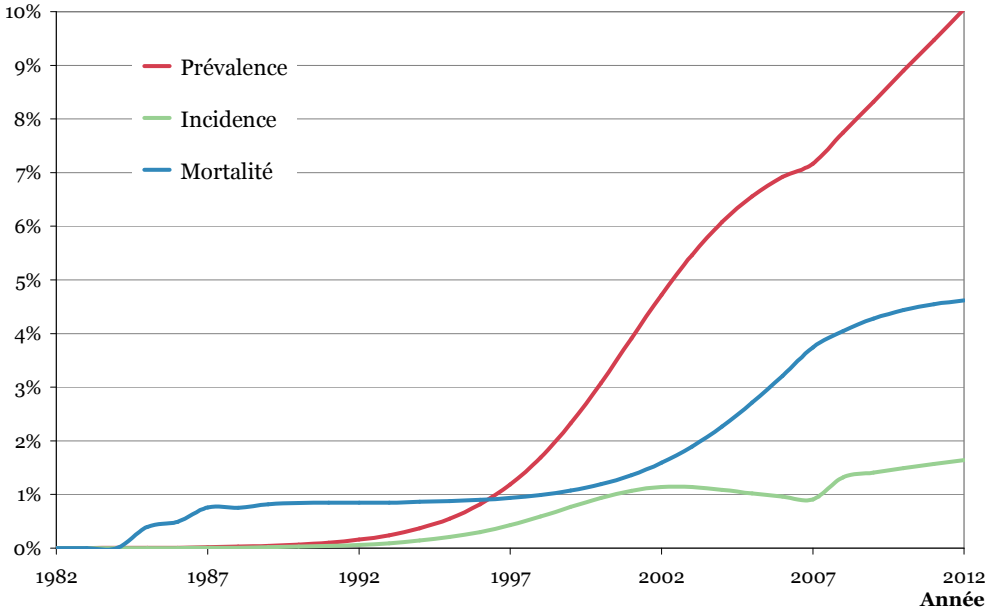
Pour mieux appréhender l'impact sur les estimations de prévalence, il nous faut tenir compte de la dynamique d'une épidémie au sein d'une population donnée. Pour des raisons financières, humaines et éthiques, nous ne disposons pas d'une population nationale pour laquelle aurait été observé depuis le début de l'épidémie l'ensemble des nouvelles infections et des décès de personnes infectées. Cependant, le recours à une modélisation mathématique peut nous fournir une assez bonne approximation de la dynamique d'une épidémie à VIH au sein d'une population nationale. Pour cela, nous allons utiliser le logiciel de projection démographique et épidémiologique Spectrum, développé par Futures Institute, sur un financement USAID, et utilisé par ONUSIDA pour ses estimations. La version 3.13 Beta 6[3] d'août 2007 est livrée avec une projection exemple correspondant à un pays africain typique présentant une épidémie généralisée qui a débuté en 1981 et qui a connu une croissance continue de la prévalence du VIH parmi les 15-49 ans pour

[3] Téléchargeable sur http://www.unaids.org.

atteindre environ 7 % en 2007 (voir Figure 3.4). Cette projection a été réalisée à partir des paramètres standards, pour les pays d'Afrique subsaharienne, préconisés par le Groupe de Référence d'ONUSIDA sur les estimations, la modélisation et les projections.

Figure 3.4

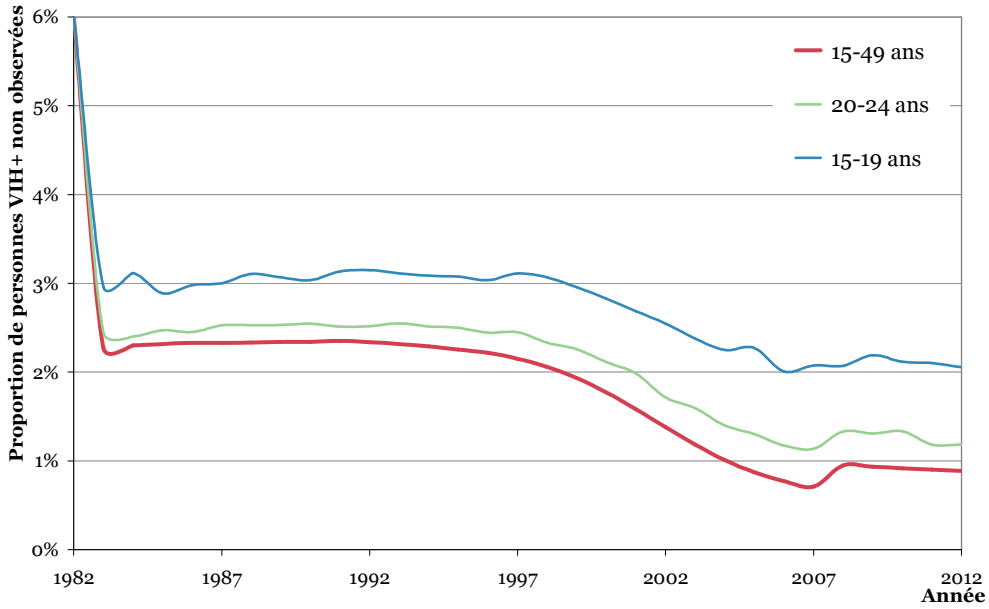
Prévalence, incidence et mortalité liée au VIH des 15-49 ans, de 1982 à 2012, de la projection exemple du logiciel Spectrum



Spectrum nous fournit, pour chaque année entre 1982 et 2012 et par groupe d'âges, le nombre de nouvelles infections dans l'année et le nombre de personnes infectées (en prenant en compte la mortalité spécifique des personnes VIH+ et les évolutions démographiques de la population). En considérant que les nouvelles infections annuelles se répartissent uniformément au cours de l'année[4], nous pouvons calculer pour chaque année la proportion de personnes infectées non observables (voir l'annexe 2 pour le détail des calculs). Nous avons retenu l'hypothèse centrale concernant la durée de la fenêtre sérologique, à savoir 22 jours. Nous avons détaillé les calculs pour l'ensemble des 15-49 ans ainsi que pour les deux groupes d'âges les plus jeunes, 15-19 ans et 20-24 ans (voir Figure 3.5).

[4] Ce qui constitue une hypothèse moins contraignante que celle du modèle simple, puisque nous supposons l'incidence constante uniquement sur une durée d'un an.

Figure 3.5
Proportion de personnes infectées non observables, sous l'hypothèse d'une fenêtre sérologique de 22 jours, pour trois groupes d'âges, selon la projection exemple du logiciel Spectrum



Pendant les quinze premières années de l'épidémie, tandis que celle-ci se développe et se généralise dans la population, environ 2,3 % des personnes infectées se situent dans la fenêtre sérologique. Puis, au fur et à mesure que les premières générations de personnes infectées décèdent (*cf.* l'augmentation de la courbe de la mortalité liée au VIH sur la Figure 3.4), cette proportion diminue pour atteindre 0,7 % en 2007. Avec la remontée de l'incidence en 2008/2009, selon ce modèle, la proportion de personnes non observables remonte à 0,9 % avant de diminuer à nouveau.

Pour les groupes d'âges les plus jeunes, cette proportion est plus importante que dans le reste de la population du fait que, les individus de ces groupes d'âges n'ont pu être infectés qu'il y a moins de 5 ou 10 ans car ils n'ont commencé leur sexualité que récemment (l'âge médian au premier rapport sexuel se situant entre 15 et 22 ans en Afrique subsaharienne, voir Tableau annexe 2.5). La structure par durée d'infection des personnes infectées est donc profondément différente pour ces groupes d'âge. Ainsi, la proportion de personnes VIH+ non observables parmi les 15-19 ans (respectivement 20-24 ans), se situe aux alentours de 3 % (respectivement 2,5 %) pendant la phase de développement initial de l'épidémie, puis diminue pour atteindre 2 à 2,5 % (respectivement 1,1 à 1,4 %).

Nous pourrons ainsi retenir que, pour une épidémie généralisée de plus de 20 ans, la proportion de personnes infectées non observables est de l'ordre de 1 % pour la population adulte et qu'elle se situe entre 2 et 2,5 % pour les 15-19 ans et entre 1 et 1,5 % pour les 20-24 ans.

3.1.3 Sensibilité et spécificité d'un test

Les épidémiologistes distinguent deux caractéristiques principales des tests de dépistage : leur *sensibilité* et leur *spécificité*. Un test est considéré comme sensible s'il renvoie majoritairement un résultat positif quand une personne est infectée et il sera considéré comme spécifique s'il renvoie un résultat négatif quand une personne n'est pas infectée. Plus précisément, la sensibilité est la proportion de personnes dépistées positives parmi les personnes infectées et la spécificité la proportion de résultats négatifs parmi les individus non infectés (voir Tableau 3.2).

Tableau 3.2
Sensibilité, spécificité et valeurs prédictives d'un test

		Situation réelle		
		infecté	*non infecté*	*Total*
Résultat du test	*positif*	**a vrais positifs**	b faux positifs	*a+b*
	négatif	c faux négatifs	**d vrais négatifs**	*c+d*
	Total	*a+c*	*b+d*	

Définitions *:*

- Sensibilité (sb) : a/(a+c)
- Spécificité (sp) : d/(b+d)
- Valeur prédictive positive (vpp) : a/(a+b)
- Valeur prédictive négative (vpn) : d/(c+d)

Selon les objectifs d'un test de dépistage, on privilégiera différemment la sensibilité ou la spécificité d'un test. L'ONUSIDA et l'OMS recommandent néanmoins comme norme minimale une sensibilité supérieure à 99 % et une spécificité supérieure à 95 % (ONUSIDA/OMS 1997b, p. 82). Dans une optique de sécurité transfusionnelle, par exemple, il importe avant tout d'écarter les échantillons infectés et l'on privilégiera les tests les plus sensibles. Dans une optique de surveillance de l'épidémie, nous chercherons à ce que la prévalence observée p_o soit la meilleure estimation possible de la prévalence réelle p_r de l'épidémie. La prévalence observée correspond à la proportion de résultats positifs, à savoir l'addition de la proportion de vrais positifs et de celle de faux positifs. Il en résulte l'Équation 3.1.

Équation 3.1

Lien entre prévalence observée et prévalence réelle selon la spécificité et la sensibilité du test

$$p_o = p_r \times sb + (1 - p_r) \times (1 - sp)$$

p_o : *prévalence observée*

p_r : *prévalence réelle*

sb : *sensibilité*

sp : *spécificité*

La meilleure estimation possible sera obtenue dans le cas où la prévalence observée sera égale à la prévalence réelle, c'est-à-dire que leur ratio sera égal à 1. Nous obtenons alors l'Équation 3.2.

Équation 3.2

Condition pour que la prévalence observée soit égale à la prévalence estimée

$$\frac{p_o}{p_r} = 1$$

$$\Leftrightarrow \frac{p_r \times sb + (1 - p_r)(1 - sp)}{p_r} = 1$$

$$\Leftrightarrow \frac{(1 - sp)(1 - p_r)}{p_r} = 1 - \frac{p_r \times sb}{p_r}$$

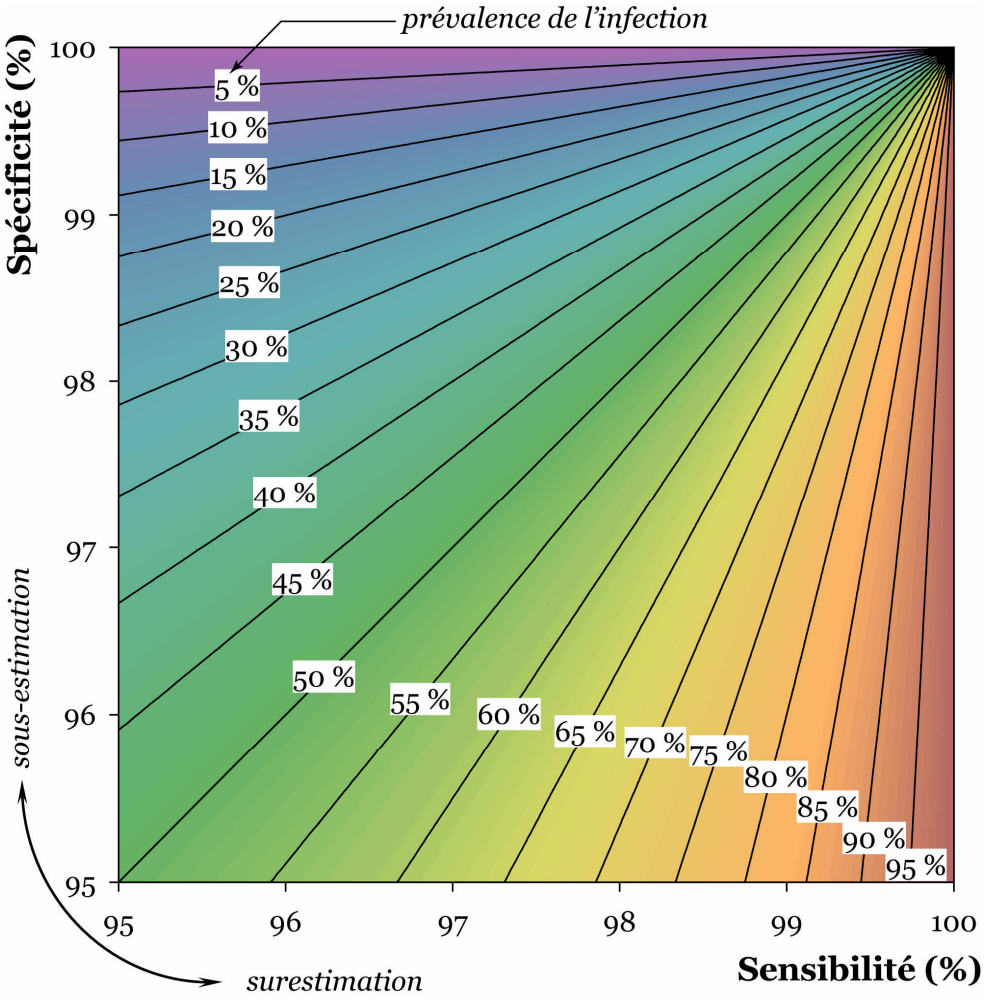
$$\Leftrightarrow (1 - sp)(1 - p_r) = (1 - sb)p_r$$

$$\Leftrightarrow \textit{proportion de faux positifs} = \textit{proportion de faux négatifs}$$

$$\Leftrightarrow sp = \frac{p_r}{1 - p_r} sb + \frac{1 - 2p_r}{1 - p_r}$$

Cette condition est obtenue quand le nombre de faux positifs est compensé par le nombre de faux négatifs. Il apparaît par ailleurs que ces deux quantités sont non seulement tributaires de la sensibilité et de la spécificité du test mais également du niveau de la prévalence. Lorsque faux positifs et faux négatifs se compensent exactement, il existe alors une relation linéaire, dépendantes du niveau de prévalence, entre sensibilité et spécificité. La Figure 3.6 met en évidence, pour différents niveaux de prévalence, ces « lignes de compensation ». C'est un graphique à trois entrées. Connaissant la sensibilité et la spécificité d'un test ainsi que la prévalence de l'infection que l'on étudie, si l'on se situe sur la ligne correspondant au niveau de prévalence, alors les faux positifs et les faux négatifs se compenseront. Si l'on se situe au-dessus de cette ligne, alors le test employé étant plus spécifique, les faux positifs seront moins nombreux que les faux négatifs et la prévalence observée sous-estimera la prévalence réelle. À l'inverse, si l'on se situe à la droite de la ligne de compensation, la sensibilité accrue du test induira une augmentation des faux positifs et donc une surestimation de la prévalence.

Figure 3.6
Lignes de compensation entre sensibilité et spécificité pour certaines valeurs de prévalence



Lecture du graphique :
Les lignes en noir indiquent les valeurs de la sensibilité et de la spécificité d'un test telles que, pour un niveau de prévalence donné, indiqué dans les rectangles blancs, les nombres de faux positifs et de faux négatifs se compensent.
Si, pour un niveau de prévalence donné, on se situe à la droite de la ligne de compensation, la prévalence observée *surestimera* la prévalence réelle. Si l'on se situe à gauche de la ligne de compensation, la prévalence observée *sous-estimera* la prévalence réelle.

Nous pouvons remarquer sur cette figure que, lorsque la prévalence de l'infection est inférieure à 50 %, ce qui est le cas pour le VIH, la compensation des faux positifs par les faux négatifs est obtenue pour des spécificités supérieures aux sensibilités, et cela d'autant plus que le niveau de prévalence est faible.

Il n'est pas toujours simple de mesurer avec précision la sensibilité et la spécificité des différents tests dans la mesure où ces dernières varient en fonction du contexte. En effet, chaque test repose sur une recherche particulière de certains antigènes du

VIH. Or, les études génétiques montrent une très grande variabilité du VIH (MCCUTCHAN 2000, KORBER 2001, TATT 2001). Outre la distinction entre deux virus VIH-1 et VIH-2, chaque type de virus comporte lui-même des différences. Le VIH-1 est divisé en trois groupes : M (majoritaire), O (outlier) et N (non M, non O). Le groupe M est lui-même subdivisé en sous-types dénommés par des lettres allant de A à J. La connaissance de la variabilité génétique est importante pour la mise au point des tests de dépistage. En effet, un test peut être bon dans un pays et moins efficace dans un autre s'il ne détecte pas tous les sous-types circulant dans cet autre pays. Cette variabilité doit donc être prise en compte dans le choix des tests diagnostiques et elle influe sur la sensibilité et la spécificité de chaque test dans une population donnée.

Il n'est pas possible de présenter ici la spécificité et la sensibilité de l'ensemble des tests existants. Pour avoir une idée des valeurs de ces dernières, nous présenterons une seule étude, réalisée par Nicolas MÉDA et ses collaborateurs au Burkina Faso et publiée en 1999 (MÉDA 1999). Dans la mesure où, depuis 8 ans, les différents kits de dépistage se sont améliorés, nous pourrons considérer les résultats de cette étude comme des valeurs *a minima* de la sensibilité et de la spécificité des tests actuels. Ils ont tout d'abord analysé, individuellement, 8 tests : 5 tests ELISA mixtes (VIH-1 et VIH-2) ainsi que trois tests rapides (Tableau 3.3).

La majorité des tests présentent une excellente sensibilité (100 %), certains tests étant à la fois sensibles et spécifiques à 100 % (ICE HIV-1.O.2).

Tableau 3.3
Sensibilité et spécificité de 8 tests de recherche d'anticorps anti-VIH mesurées sur un échantillon de 733 sérums

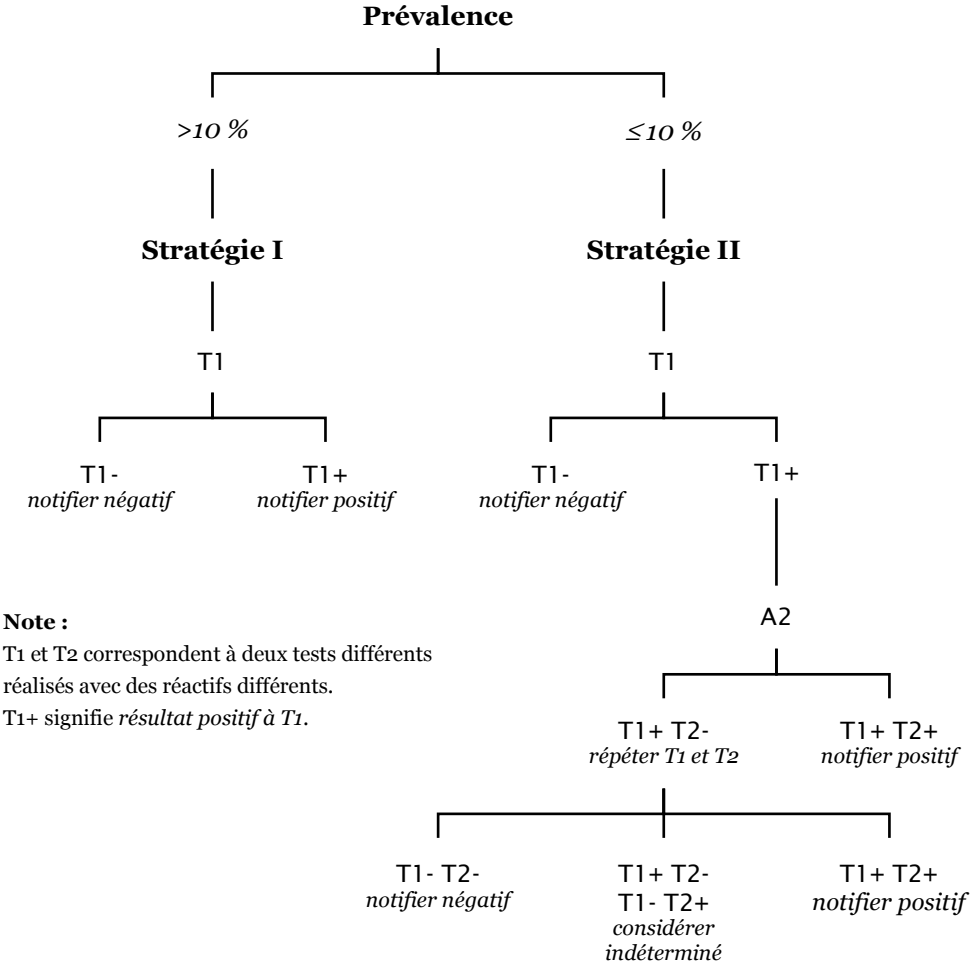
Test	Sensibilité (en %)	IC 95 %	Spécificité (en %)	IC 95 %
Tests ELISA Mixtes				
Genelavia Mixt® (GLA)	100,0	98,3-100,0	93,4	90,7-95,5
Enzygnost HIV 1/2 Plus® (ZYG)	100,0	98,3-100,0	99,6	98,3-99,3
Vironostika Uni-Form II® (VKU)	100,0	98,3-100,0	96,5	94,3-97,9
Murex HIV 1 + 2® (REX)	100,0	98,3-100,0	81,4	77,5-84,8
ICE HIV-1.O-2® (ICE)	100,0	98,3-100,0	100,0	99,0-100,0
Tests rapides				
Multispot HIV-1/HIV-2® (MTT)	99,3	97,1-99,9	98,7	97,0-99,5
CombAIDS-RS® (COM)	100,0	98,3-100,0	99,6	98,3-99,9
HIV Spot® (SPO)	98,2	98,2-99,3	100,0	99,0-100,0

Source : (MÉDA 1999).
IC 95 % : Intervalle de Confiance à 95 %.

3.1.4 Algorithme de dépistage

Les stratégies de dépistage se contentent rarement d'avoir recours à un seul test et des algorithmes combinant plusieurs tests différents sont utilisés. Pour la surveillance de l'épidémie, l'ONUSIDA et l'OMS recommandent deux stratégies : la stratégie I pour les pays présentant une prévalence supérieure à 10 % et la stratégie II pour les pays présentant une prévalence inférieure à 10 % (ONUSIDA/OMS 1997b).

Figure 3.7
Stratégies ONUSIDA et OMS pour le dépistage du VIH en matière de surveillance



Note :
T1 et T2 correspondent à deux tests différents réalisés avec des réactifs différents.
T1+ signifie *résultat positif à T1*.

Source : (ONUSIDA/OMS 1997b).

La stratégie II repose sur la combinaison de deux tests : un premier test très sensible puis un second test, très spécifique, afin d'éliminer les faux positifs. Si un tel algorithme ne permet pas d'augmenter la sensibilité globale du dépistage[5], il permet au final d'obtenir une spécificité plus importante que celle des deux tests pris individuellement.

Le nombre total de personnes notifiées positives correspond au nombre de personnes ayant eu deux tests concordants au premier passage, ou bien, après répétition des tests en cas de tests discordants au premier passage. Le nombre de personnes notifiées négatives correspond majoritairement aux personnes notifiées négatives après T1, plus les personnes dont les deux tests étaient négatifs après répétition suite un résultat T1+ T2- au premier passage.

Dans le cadre d'une opération de surveillance du VIH, il n'est pas possible de recontacter les individus pour procéder à un nouveau prélèvement en cas de résultat indéterminé. Ainsi, dans le calcul d'une prévalence, les indéterminés (personnes ayant eu des résultats discordants aux deux passages) seront retirés du dénominateur.

Si nous supposons que chaque test est totalement indépendant[6] des autres, nous pouvons calculer *sb* et *sp*, la sensibilité et la spécificité globale de l'algorithme, à l'aide des formules ci-dessous où sb_1 et sp_1 correspondent à la sensibilité et la spécificité du premier test T1, et sb_2 et sp_2 à celles du second test T2.

Équation 3.3

Sensibilité et spécificité globale d'une stratégie de type II

$$sb = \frac{sb_1 sb_2 + sb_1(1-sb_2)sb_1 sb_2}{1-\left[\left((1-sb_1)sb_2 + sb_1(1-sb_2)\right)sb_1(1-sb_2)\right]}$$

$$sp = \frac{sp_1 + (1-sp_1)sp_2 sp_1 sp_2}{1-\left[(1-sp_1)sp_2\left((1-sp_1)sp_2 + sp_1(1-sp_2)\right)\right]}$$

L'étude de MÉDA *et al.* recommande trois stratégies de type II pour lesquelles ont été mesurées une spécificité et une sensibilité de 100 %. Il s'agit de ICE / COM, ICE / ZYG et ICE / VKU. Nous allons essayer de maximiser les biais induits par ces

[5] Plus précisément, la sensibilité globale de l'algorithme est légèrement inférieure à celle du premier test, puisqu'aux faux négatifs produits par le premier test, il y a également un nombre de faux négatifs, extrêmement faible, lorsque la répétition des deux tests s'est avérée négative après un second test discordant.

[6] Ce qui n'est pas tout à fait le cas car, même si les deux tests portent sur des antigènes différents, il peut y avoir interférence entre les deux. De plus, lorsque l'on répète un même test sur un même échantillon, bien que les résultats puissent être discordants, les deux mesures ne sont jamais totalement indépendantes.

combinaisons de tests en estimant, à l'aide des formules précédentes, la sensibilité et la spécificité générale de ces algorithmes à partir des bornes inférieures des intervalles de confiance des mesures de sensibilité et de spécificité de ces quatre tests pris individuellement.

Tous les quatre présentent, selon le Tableau 3.3, une sensibilité de 100 % avec une borne inférieure de 98,3 %. Cela implique une sensibilité globale de l'algorithme, selon l'Équation 3.3, de 98,2986 %. Comme annoncé plus haut, la sensibilité globale de l'algorithme est légèrement inférieure à celle du premier test. Cependant, cette diminution est très légère. Si la sensibilité des deux tests est estimée à 99,5 %, alors cette réduction de la sensibilité globale devient infime ($-3{,}719\times10^{-7}$).

Concernant la spécificité globale de l'algorithme, en tenant compte des bornes inférieures des intervalles de confiance, nous obtenons pour ICE / VKU une spécificité de 99,9424 % et pour les combinaisons ICE / ZYG et ICE / COM 99,9828 %. Nous voyons donc que l'algorithme de type II permet d'augmenter significativement la spécificité globale de l'algorithme par rapport aux spécificités des deux tests initiaux. Et si les deux tests présentent chacun une spécificité de 99,5 %, la spécificité de l'algorithme atteint 99,9975 %.

Dans la mesure où la majorité des tests est plus sensible que spécifique, les stratégies de type II permettent d'augmenter très significativement la spécificité de l'ensemble du processus de dépistage tout en diminuant à peine la sensibilité de l'ensemble. Au final, ce type de stratégie induit une spécificité supérieure à la sensibilité.

Dans le cadre des enquêtes de surveillance sérologique, une stratégie *a minima* de type II est utilisée bien que le plus souvent les algorithmes employés soient plus complexes, notamment dans le cadre des Enquêtes Démographiques et de Santé (EDS). Les algorithmes employés sont variables d'une enquête à l'autre mais présentent cependant des similitudes. Tous reposent sur un premier test très sensible dont un résultat négatif induit une notification négative. Le second test est souvent remplacé par une série de deux tests, l'un spécifique du VIH-1 et l'autre spécifique du VIH-2, ou par un test de confirmation et un test permettant de discriminer en VIH-1 ou VIH-2. Par ailleurs, certains algorithmes prévoient des procédures pour pouvoir distinguer les infections récentes. Les résultats douteux ou indéterminés sont le plus souvent confirmés en dernier recours par la technique Western Blot, souvent considérée comme la méthode de référence, mais peu utilisée en raison de son coût prohibitif. Par ailleurs, la majorité des EDS inclut une étude de validation en testant à l'aide d'un second réactif une partie (5 % à 10 %) des échantillons notifiés négatifs suite au premier test.

Bien que ces algorithmes soient largement plus complexes que la stratégie de type II, nos remarques précédentes restent valables : la sensibilité globale de l'algorithme peut être assimilée à la sensibilité du premier test et être considérée comme atteignant quasiment les 100 % (d'autant plus que les études de validation n'ont pas détecté de faux négatifs parmi les échantillons notifiés négatifs suite au premier test). La spécificité globale de l'algorithme, quant à elle, est supérieure à la spécificité de chaque test pris individuellement et, même quand les tests affichent des spécificités comprises entre 95 et 99 %, la spécificité globale de l'algorithme frôle les 100 %.

Cependant, s'il apparaît que nous pouvons considérer que les algorithmes utilisés atteignent des sensibilités et des spécificités de l'ordre de 100,0 %, il y a un biais que nous avons évoqué à la section 3.1.2 qui n'est pas pris en compte dans les études de sensibilité et de spécificité : la fenêtre sérologique. Cette dernière n'a d'impact que sur la sensibilité. Nous avons estimé qu'il restait une proportion irréductible de personnes non observables de l'ordre de 0,5 à 1 %. Autrement dit, la sensibilité des algorithmes employés, même dans le meilleur des cas, sera de l'ordre de 99 à 99,5 %. Reste à estimer l'impact de ce biais sur les mesures de prévalence du VIH.

3.1.5 Impact du biais

À partir de l'Équation 3.1, il est possible de calculer la prévalence observée en fonction de la sensibilité et de la spécificité de l'algorithme employé ainsi que du niveau de prévalence. Nous avons représenté sur la Figure 3.8 les écarts entre la prévalence observée et la prévalence réelle, selon le niveau de prévalence, pour différentes valeurs de sensibilité et de spécificité. Un écart négatif indique une sous-estimation. Par ailleurs, pour mieux apprécier l'ampleur du biais induit, nous avons également représenté les intervalles de confiance à 95 % correspondant à des échantillons de 5 000, 10 000 et 20 000 individus. Pour des échantillons plus restreints, les intervalles de confiance sont beaucoup plus larges[7]. Les enquêtes sur des populations spécifiques ou bien la surveillance sentinelle des femmes enceintes portent sur des échantillons le plus souvent inférieurs à 500. Les enquêtes en population générale, en particulier les EDS et les AIS, portent sur des échantillons situés entre 5 000 et 17 000 individus (voir Tableau 3.5 page 144).

Conformément à ce que nous avons montré précédemment, les stratégies de test de type II[8] présentent une spécificité de l'ordre de 100 % et une sensibilité de 99 à

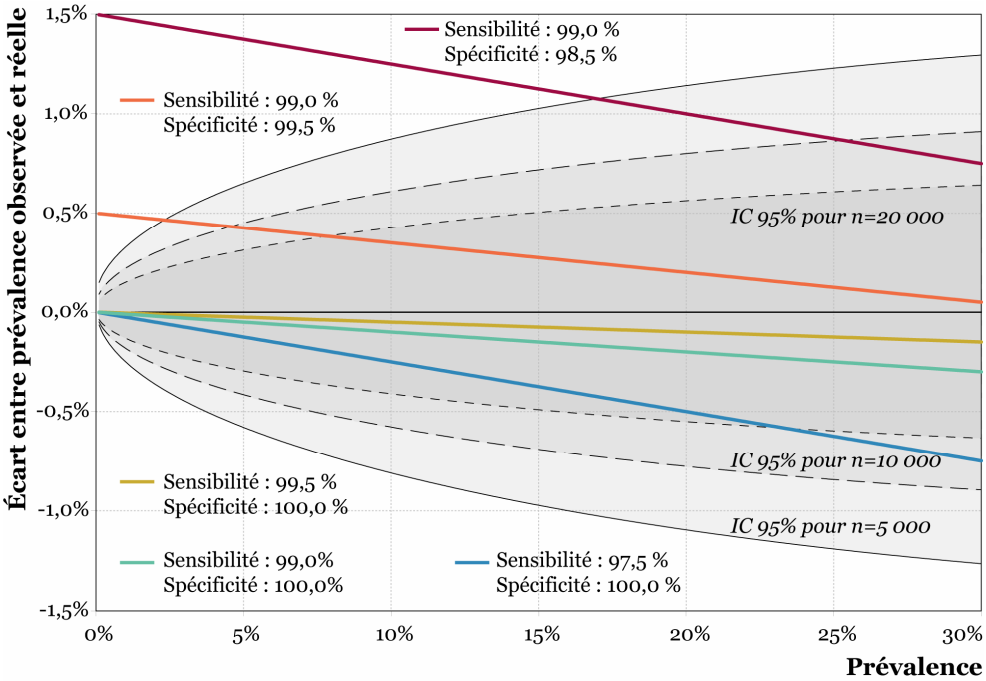
[7] Si l'on divise la taille de l'échantillon par deux, la largeur de l'intervalle est approximativement multipliée par 4.

[8] Ainsi que les algorithmes plus complexes.

99,5 % en raison de la proportion d'individus non observables due à la fenêtre sérologique (courbes verte et jaune). Une sensibilité de 99 % et une spécificité de 100 % induisent une sous-estimation de la prévalence réelle particulièrement faible : environ un centième de la prévalence réelle (soit un écart de -0,05 points pour une prévalence de 5 %). Cela constitue un biais plus faible que l'incertitude liée aux échantillonnages. En effet, cet écart se situe au sein de l'intervalle de confiance à 95 % pour un effectif de 20 000 individus. Pour que l'imprécision liée aux tests de dépistage soit supérieure à celle de l'échantillonnage[9], il faut atteindre des tailles d'échantillons de l'ordre de 500 000 à 1 000 000 selon le niveau de prévalence.

Figure 3.8

Écarts entre prévalence observée et prévalence réelle, pour différentes valeurs de sensibilité et de spécificité, et intervalles de confiance à 95 %, pour différentes tailles d'échantillons, selon le niveau de prévalence



IC 95 % : intervalle de confiance à 95 % d'une proportion, calculé selon la méthode de WILSON avec correction de continuité (voir Annexe 3). Les intervalles de confiance sont représentés par les aires grisées et les limites des intervalles par les courbes noires.

Les méthodes de dépistage s'améliorant, il est possible de réduire la taille de la fenêtre sérologique à 17 jours induisant une proportion de non observables de l'ordre de 0,5 à 0,75 % (courbe jaune). Cependant, pour certains groupes d'âges, en

[9] Nous discuterons plus précisément dans la suite de ce chapitre des notions de représentativité et de biais statistique ainsi que du concept d'intervalle de confiance (section 3.2).

particulier les 15-19 ans, la proportion de non observables peut atteindre 2 à 2,5 %. Même ainsi, le biais lié à la fenêtre sérologique reste relativement inférieur aux imprécisions statistiques (courbe bleue).

Par contre, si les biais sont relativement faibles pour des variations de sensibilité lorsque la spécificité avoisine les 100 %, la Figure 3.8 montre que les écarts sont beaucoup plus importants dès que la spécificité diminue (courbes rouge et orange).

L'étude de MÉDA et de ses collaborateurs (MÉDA 1999) recommandait, pour la stratégie de type I de l'ONUSIDA et de l'OMS (ONUSIDA/OMS 1997b), d'utiliser ZYG pour les opérations de surveillance dans les pays présentant une prévalence supérieure à 10 %. La même étude mesurait une spécificité de 99,6 % pour ce test. Or, lorsque la spécificité est de 99,5 % pour une sensibilité de 99 % (avec prise en compte de la fenêtre sérologique), les faux négatifs ne sont plus suffisants pour compenser les faux positifs et la prévalence observée surestime cette fois-ci la prévalence réelle. Si cette surestimation diminue au fur et à mesure que la prévalence se rapproche de 50 %, les écarts sont beaucoup plus importants et, à des niveaux de prévalence inférieurs à 5 %, sont supérieurs à l'imprécision de l'échantillonnage. Pour des prévalences supérieures à 10 %, le biais reste inférieur à l'imprécision statistique. Par contre, il augmente rapidement quand la spécificité diminue (voir courbe rouge pour une spécificité de 98,5 %), de même quand la sensibilité augmente. Plus précisément, la spécificité du test détermine la surestimation observée aux très faibles prévalences et la sensibilité celle aux très fortes prévalences (proches de 100 %) et donc, pour une même valeur de spécificité, la pente de la droite.

*
**

En conclusion, si les stratégies de type II permettent de réduire les biais d'observations dus aux tests de dépistage à des niveaux largement inférieurs aux imprécisions statistiques, il importe, dans le cadre d'une stratégie de type I, d'équilibrer correctement sensibilité et spécificité. Favoriser un test très sensible et peu spécifique induit alors, dans une optique de surveillance, un biais non négligeable.

3.2 Représentativité, erreur aléatoire et erreur systématique

Nos résultats précédents mettent en évidence que les biais induits par les tests biologiques de dépistage sont largement inférieurs à l'imprécision statistique quand des algorithmes de type II sont utilisés. Nous avons eu recours au concept d'intervalle de confiance pour quantifier cette imprécision, mais il nous faut la définir plus clairement.

3.2.1 Représentativité d'un échantillon

Les techniques d'enquêtes visent à être *représentatives* de la population étudiée. Cette dernière peut rarement être mesurée de manière exhaustive. Nous avons alors recours à une sélection d'un sous-échantillon de cette population-mère, appelée généralement univers en théorie des sondages. Dès lors, si l'échantillon est dit représentatif, il est possible de déduire les caractéristiques de la population-mère à partir du sous-échantillon étudié. Plusieurs définitions de la représentativité existent (GALVANI 1951). Selon KIAËR, un échantillon doit être une miniature aussi exacte que possible de la population-mère (KIAËR 1895), tandis que MARCH estime qu'un relevé représentatif « *devrait fournir une espèce de réduction de toutes les parties de l'ensemble*[10] » (MARCH 1925). Si cette définition est intuitive, elle n'est cependant pas suffisamment précise. Usuellement, un échantillon est considéré comme représentatif lorsqu'il a été obtenu selon une méthode probabiliste, c'est-à-dire que tous les éléments de l'échantillon avaient la même probabilité, *a priori*, d'être sélectionné, dans le cadre d'un échantillonnage aléatoire simple, ou plus généralement que cette probabilité d'être sélectionné, *a priori*, est connue. Le recours à des modèles probabilistes permet alors de déterminer avec une certaine marge d'erreur l'écart entre les caractéristiques de l'échantillon observé et celles de la population-mère étudiée. En pratique, un échantillonnage n'est jamais parfaitement aléatoire.

Nous retiendrons, pour notre part, que *la représentativité est la possibilité d'extrapoler les résultats obtenus sur un échantillon à l'ensemble de la population dont il est extrait*. Cette définition implique que, pour déclarer un échantillon représentatif, il faut également préciser de quelle *population* il est représentatif,

[10] Cité par GALVANI L., « Révision critique de certains points de la méthode représentative », *Revue de l'Institut International de Statistique*, n°19(1), 1951.

pour quelles *variables* et quelles *caractéristiques* (moyenne, variance, distribution…) de ces variables.

Dans toute observation partielle, il y a toujours un écart entre la valeur de la variable étudiée au sein de la population-cible (celle que l'on veut étudier) et cette même valeur au sein de l'échantillon observé. Cet écart peut être subdivisé entre deux types d'erreurs : l'*erreur aléatoire* et l'*erreur systématique*. Il importe de bien distinguer ces deux types d'erreurs de l'*erreur radicale* que nous évoquions au chapitre 2 (section 2.8). Une mesure "idéale" sera donc atteinte si à la fois l'erreur aléatoire et l'erreur systématique sont nulles, ce qui n'est jamais atteint en pratique. Dans une analyse portant sur 70 études empiriques, COTE et BURKLEY estimaient que seulement 42 % de la variance reflétaient des variations de la caractéristique étudiée, 32 % de la variance provenaient de l'erreur aléatoire et 26 % étaient imputables à la méthode utilisée (COTE 1987).

3.2.2 L'erreur aléatoire

Les méthodes d'échantillonnage probabilistes s'appuient sur des techniques de sélection des individus où seul le "*hasard statistique*" intervient. Il est inutile ici d'avoir recours à une notion métaphysique de ce qu'est le Hasard. Dans le cadre d'un modèle probabiliste, l'idée de hasard implique que chaque élément de la population-mère ait une même probabilité d'être sélectionné[11]. Il suffit pour cela d'avoir recours à des dispositifs opératoires tels que « *des conditions initiales inséparables expérimentalement sont suivies ultérieurement d'une séparation manifeste des phénomènes observés* » (ULLMO 1967, p. 649). Un exemple classique est fourni par la sélection de boules numérotées au sein d'une urne. En connaissant avec une extrême précision la position de chaque boule ainsi que le mouvement réalisé par la personne procédant au tirage, nous serions en capacité, théoriquement, de déterminer quelle boule sera sélectionnée. Cependant, dans la pratique, nous ne connaissons pas ces détails avec précision et une infime variation des conditions initiales induira la sélection d'une autre boule. Dès lors, nous pouvons considérer que les différentes boules sont bien tirées au hasard.

Cette condition de hasard est nécessaire pour que nous puissions appliquer les modèles statistiques probabilistes. Ces derniers nous permettent de quantifier avec précision l'erreur aléatoire que nous prenons. Si nous tirons 100 boules dans une urne qui en contient 5 000 rouges et 5 000 bleues, il est rare que nous observions exactement 50 boules rouges parmi les 100 tirées au sort. Il va donc y avoir un

[11] Ou que cette probabilité d'être sélectionné *a priori* soit parfaitement connue lorsque l'on procède à des sélections selon des probabilités non proportionnelles comme dans le cas de sondages en grappes ou en strates. Néanmoins, cela ne modifie pas notre propos sur la notion de Hasard.

écart entre la proportion de boules rouges que nous observons et la proportion réelle de boules rouges dans l'urne. Cet écart entre la proportion observée et la proportion réelle constitue justement ce que nous avons appelé *erreur aléatoire*. Elle peut être parfaitement décrite dans le cadre d'un modèle probabiliste. Dans cet exemple, il est possible de montrer qu'il y a 19 chances sur 20 pour que le nombre de boules rouges tirées au sort soit compris entre 40 et 60. Notre échantillon de boules sera considéré comme représentatif de l'urne dans la mesure où la proportion de boules rouges observées sera un bon estimateur de la proportion réelle de boules rouges dans l'urne. Reste à préciser les propriétés que l'on demande à un bon estimateur, à savoir d'être sans biais et convergent. Un estimateur sera considéré *sans biais* s'il est égal, en moyenne, au paramètre qu'il estime[12]. Par ailleurs, on cherche à ce que notre estimateur soit d'autant plus précis que notre échantillon sera grand. Un estimateur sera donc *convergent* si l'erreur aléatoire tend à devenir nulle quand le nombre d'observations augmente[13].

L'erreur aléatoire est donc inhérente au procédé utilisé pour la sélection des individus composant l'échantillon. Une enquête idéale d'un point de vue probabiliste est une enquête où l'erreur systématique est nulle. Il ne subsiste alors que l'erreur aléatoire. Dans ce cas, les modèles probabilistes permettent de fournir une idée de l'amplitude de cette erreur, par le calcul d'intervalles de confiance, et de procéder à des tests statistiques.

Un *intervalle de confiance* consiste à calculer un intervalle au sein duquel la caractéristique qui nous intéresse se situe, connaissant cette même caractéristique au sein de l'échantillon, avec une certaine confiance, usuellement 95 %. Même dans le cadre d'un échantillon parfaitement représentatif, le risque d'erreur radical subsiste, bien que la probabilité, *a priori*, que la valeur réelle du paramètre étudié soit située en dehors de l'intervalle de confiance puisse être réduite autant que l'on veut[14].

Le principe d'un *test statistique* vise, quant à lui, à faciliter le choix entre deux hypothèses appelées usuellement hypothèse nulle et hypothèse alternative. Dans un premier temps, en supposant que l'hypothèse nulle est vraie, est calculée la probabilité d'observer ce qui a été effectivement observé. Si cette probabilité s'avère être inférieure à un certain seuil, par exemple 5 %, alors l'hypothèse nulle sera rejetée pour accepter l'hypothèse alternative. Par exemple, nous pouvons chercher

[12] T_n est un estimateur sans biais du paramètre θ si et seulement si $E(T_n)= \theta$. D'après BOUZITAT C., BOUZITAT P. et PAGÈS G., *Statistique, Probabilités, Estimation ponctuelle : cours et exercice d'application*, Paris (FR), Cujas, 1990, p. 163.

[13] T_n est un estimateur convergent du paramètre θ si et seulement si $\forall\varepsilon>0$, $P(|T_n-\theta|>\varepsilon)\rightarrow 0$ quand $n\rightarrow\infty$. D'après BOUZITAT, BOUZITAT et PAGÈS, *Statistique, Probabilités, Estimation ponctuelle*, p. 165.

[14] Il suffit pour cela d'augmenter le niveau de confiance de notre intervalle.

à savoir si la proportion de personnes séropositives diffère significativement entre les hommes et les femmes. L'hypothèse nulle supposera que cette prévalence est la même dans ces groupes et que les écarts observés ne seront donc dus qu'à des effets d'échantillonnage. Si sous l'hypothèse nulle nous calculons une probabilité d'observer les prévalences effectivement observées de 0,001 %, alors, cette probabilité étant très faible, nous ferons le choix de rejeter l'hypothèse nulle et de considérer que la prévalence du VIH est *significativement* différente entre les hommes et les femmes. À l'inverse, si nous avions obtenu une probabilité de 50 %, nous n'aurions pas rejeté l'hypothèse nulle et aurions conclu qu'il n'y avait pas de *différence significative* entre les hommes et les femmes. Dans le premier cas, en rejetant l'hypothèse nulle, le risque que nous prenions de nous tromper était très faible (0,001 %) mais non nul. Dans la seconde situation, cela ne signifie pas pour autant qu'il n'y a pas de différences entre hommes et femmes. Simplement, que la différence observée pourrait parfaitement n'être due qu'à des variations aléatoires liées à l'échantillonnage. Dans les deux cas, nous avons fait un *choix* quant au rejet ou non de l'hypothèse nulle et, de ce fait, avons posé une hypothèse anticipatrice.

Le fait que l'erreur aléatoire puisse être mathématiquement connue avec précision ne nous prémunit donc pas du risque d'erreur radicale évoqué précédemment. Par contre, cela permet de pouvoir évaluer, *a priori* en l'absence d'autre information, l'ampleur de ce risque.

3.2.3 L'erreur systématique

Lors de la réalisation d'une enquête, de nombreuses autres sources d'erreurs sont possibles et elles sont, selon les cas, plus ou moins évaluables et quantifiables. Ces sources d'erreurs correspondent à ce que l'on nomme usuellement *biais* et elles induisent que certains aspects du phénomène seront sur ou sous-représentés dans l'échantillon final. À la section 3.1, nous avons justement analysé un type d'erreur systématique lié aux techniques de dépistage utilisées. Suivant la sensibilité ou la spécificité de l'algorithme utilisé, la prévalence observée pouvait être sur ou sous-estimée. D'autres facteurs peuvent également contribuer à l'erreur systématique :

- base de sondage présentant des écarts avec la population-cible (ancienneté, couverture partielle, etc.) ;
- technique d'échantillonnage (type de sondages, processus aléatoire, etc.) ;
- effets de sélection dans le recrutement sur le terrain (personnes absentes, zones inaccessibles, etc.) ;
- biais liés aux enquêteurs (interprétation du questionnaire, affinité vis-à-vis de certains enquêtés, etc.) ;
- biais liés aux enquêtés (compréhension des questions, oublis, refus de répondre, etc.) ;

- situations dans lesquelles s'effectue la collecte des données (questionnaire auto-administré, face-à-face, confidentialité du lieu de passation, présence de tiers, etc.) ;
- erreurs de transcription, de codage ou de saisie ;
- limites techniques des outils de mesure (questionnaire adapté, tests de dépistage, récepteur GPS, pèse-personne, etc.)…

L'ensemble des sources d'erreur systématique ne peut être déterminé ou mesuré à l'aide des modèles probabilistes. Il importe donc d'analyser qualitativement la manière et le contexte dans lesquels ont été produites les données numériques analysées.

L'impact d'une erreur systématique peut jouer à différents niveaux quant à la représentativité de l'échantillon. Prenons, à titre d'exemple, le biais induit par des refus de la part des enquêtés à une question concernant leurs revenus. Il se peut que ces refus concernent à la fois des personnes jeunes et des individus plus âgés mais pour des raisons différentes. Parmi les personnes jeunes, les refus seraient, dans notre exemple, essentiellement le fait de ceux ayant les plus faibles revenus tandis que chez les individus âgés les refus seraient plus fréquents parmi ceux ayant des revenus élevés. Les refus des jeunes compenseraient ceux de leurs aînés et ainsi le revenu moyen observé dans l'échantillon correspondrait bien au revenu moyen réel de l'ensemble de la population. Pour qu'un effet de sélection induise un biais sur la caractéristique étudiée, il importe que les personnes non enquêtées présentent un différentiel, selon la caractéristique étudiée, par rapport à ceux qui ont été interrogés avec succès. Dans notre exemple, la représentativité de notre échantillon n'est pas remise en cause, concernant le revenu moyen, dans la mesure où l'erreur systématique induit par les jeunes était compensée par celle des plus âgés. Par contre, notre échantillon ne sera plus représentatif du revenu moyen si nous réalisons une analyse par groupe d'âges. Cette fois-ci, le revenu moyen sera sous-estimé aux jeunes âges et surestimé pour les autres. D'autre part, la distribution et la variance des revenus au sein de l'ensemble de la population seront également affectées. Sur l'ensemble des revenus, les revenus extrêmes seront sous-représentés par rapport aux revenus proches de la moyenne et la variance du phénomène sera sous-estimée.

Autrement dit, il importe, lorsqu'un échantillon est dit représentatif, de préciser la population-cible qu'il représente et cela pour quelles variables et selon quelles caractéristiques de ces dernières. Lorsque plusieurs biais se compensent, nous sommes en droit que considérer que l'échantillon est toujours représentatif de notre population-cible concernant certaines caractéristiques mais nous devrons rester prudent et ne pas le considérer comme valable pour d'autres caractéristiques.

⁂

Nous pouvons maintenant préciser ce que nous avons montré à la section précédente (3.1) : l'imprécision due à l'erreur systématique induite par les limites techniques des tests de dépistage (sensibilité, spécificité, fenêtre sérologique et algorithme employé) s'avère être inférieure, dans les contextes relevant de la Figure 3.8, à l'imprécision due à l'erreur aléatoire de l'échantillonnage des enquêtes. Nous allons maintenant regarder la possibilité d'autres sources d'erreurs systématiques dans le cadre des enquêtes en population générale type EDS et AIS (section 3.3) et dans celui de la surveillance sentinelle des femmes enceintes (section 3.4).

3.3 Représentativité des EDS

Les Enquêtes Démographiques et de Santé, et les enquêtes apparentées comme les AIDS Impact Surveys, les HIV/AIDS Sero-Behavioural Survey ou les Enquêtes sur les Indicateurs du SIDA, sont échantillonnées sur des ménages. Les personnes vivant hors ménage n'y sont dons pas inclues, ce qui peut constituer une première source de biais, analysée à la section 3.3.1, dans la détermination de la prévalence de l'ensemble de la population adulte d'un pays.

Par ailleurs, la sélection des ménages se fait en deux temps (voir Encadré 1.5 page 49) pour un rappel de l'échantillonnage des EDS). Au premier degré, des grappes ou clusters sont tirés au sort avec une probabilité proportionnelle au nombre de ménages de chaque grappe. Pour cela, les enquêteurs ont recours à une base de sondage listant l'ensemble des zones ainsi que leur taille en nombre de ménages. La base de sondage correspond au dernier recensement de la population réalisé dans le pays ou, parfois, à une base de sondage dérivée de celui-ci. Or, entre la réalisation du recensement de la population et celle de l'enquête (Tableau 3.4), la population a évolué. L'ancienneté de la base de sondage constitue donc une seconde source de biais (section 3.3.2).

Une fois les zones d'enquête déterminées, un recensement exhaustif des ménages de chaque grappe est réalisé afin de constituer la base de sondage du tirage au second degré. À l'exception des quelques déménagements qui ont lieu entre le recensement des ménages et la passation du questionnaire ménages, nous pouvons considérer que les bases de sondage pour le tirage au second degré sont représentatives de l'ensemble des ménages de chaque grappe. Néanmoins, certains ménages ne peuvent être enquêtés lors du retour sur le terrain pour différentes raisons (déménagement, absence, refus, etc.) ce qui peut être potentiellement un biais (section 3.3.3).

L'enquête ménages permet d'identifier l'ensemble des individus les composant. Parmi les ménages sélectionnés, seule une partie (la moitié, le tiers ou la totalité selon l'enquête, voir Tableau 3.4) est enquêtée à la fois pour le questionnaire hommes et pour le dépistage du VIH. L'ensemble des adultes de ces ménages sont éligibles pour un prélèvement, sanguin ou salivaire selon l'enquête, après consentement de leur part. Certains individus peuvent refuser ce prélèvement. Par ailleurs, malgré des passages répétés, d'autres peuvent être absents de leur domicile au moment de l'enquête. Il en résulte qu'une partie des personnes sélectionnées pour l'enquête VIH ne sont pas testées (absence ou refus), ce qui constitue une quatrième source de biais (section 3.3.4).

Tableau 3.4

Résumé de l'échantillonnage de 17 enquêtes nationales récentes, en population générale, avec dépistage du VIH

Pays	Année de l'enquête	Type	Année de la base de sondage	Strates	Taux de sondage* pour VIH	Disponibilité des données individuelles	Disponibilité des données VIH	Disponibilité des données GPS
Burkina Faso	2003	EDS	1996	25	1/3	Disponibles	Disponibles	Disponibles
Cameroun	2004	EDS	2002-03	22	1/2	Disponibles	Disponibles	Pas encore†
Côte d'Ivoire	2005	EIS	1998	21	1/1	Données préliminaires	Pas encore	Non collectée
Éthiopie	2005	EDS	1994	20	1/2	Données préliminaires	Données préliminaires	Pas encore
Ghana	2003	EDS	2000	20	1/1	Disponibles	Disponibles	Disponibles
Guinée	2005	EDS	1996	15	1/2	Disponibles	Disponibles	Pas encore
Kenya	2003	EDS	1999§	16	1/2	Disponibles	Disponibles	Disponibles
Lesotho	2004	EDS	1996	20	1/2	Disponibles	Disponibles	Disponibles
Malawi	2004	EDS	1998	n.d.	1/3	Disponibles	Disponibles	Pas encore
Mali	2001	EDS	1998	14	1/3	Disponibles	Disponibles	Non liables aux résultats VIH‡
Niger	2006	EDS	2001	15	1/2	Pas encore	Pas encore	Pas encore
Ouganda	2004	HSBS	2002	17	1/1	Données en accès restreint	Données en accès restreint	Données en accès restreint
Rwanda	2005	EDS	2002	23	1/2	Disponibles	Disponibles	Non collectée
Sénégal	2005	EDS	2002	22	1/3	Données préliminaires	Données préliminaires	Non collectée
Tanzanie	2003	AIS	2002	42	1/1	Données préliminaires	Disponibles	Disponibles
Zambie	2001-02	EDS	2000	18	1/3	Disponibles	Disponibles	Non liables aux résultats VIH‡
Zimbabwe	2005-06	EDS	2002	34	1/1	Données préliminaires	Données préliminaires	Pas encore

Sources: http://www.measuredhs.com et rapport final de chaque enquête..

EDS : Enquête démographique et de Santé ; AIS : AIDS Impact Survey ; HSBS : HIV/AIDS Sero-Behavioural Survey ;

EIS : Enquête sur les Indicateurs du SIDA ; GPS : Global Positioning System.

n. d. : information non disponible.

Disponibilité en téléchargement sur www.measuredhs.com au 12 septembre 2007.

* Il s'agit de la proportion, parmi les ménages sélectionnés pour l'enquête individuelle, éligibles pour le questionnaire hommes et le dépistage du VIH.

† Les coordonnées GPS de l'EDS 2004 du Cameroun étaient disponibles début 2006 et ont été retirées plusieurs mois plus tard.

‡ Ces deux premières enquêtes sont sorties avant que Macro Internationale ne mette en place une procédure pour protéger l'anonymat des personnes. Les résultats des tests de dépistage VIH ne peuvent donc être liées aux données individuelles ni aux coordonnées géographiques des zones enquêtées.

§ La base de sondage utilisée au Kenya est fondée sur le recensement réalisé en 1999 mais a connu des mises à jours partielles en 2002.

Tableau 3.5
Résumé de l'échantillonnage de 17 enquêtes nationales récentes, en population générale, avec dépistage du VIH, suite

Pays	Année	Type	Grappes	Ménages éligibles*	Personnes testées 15-49 ans	Nombre moyen de pers. testées par grappe	Prévalence du VIH (%) 15-49 ans	IC 95 %
Burkina Faso	2003	EDS	400	3179	7151	17,9	1,8	1,5 - 2,1
Cameroun	2004	EDS	466	5319	9900	21,2	5,5	5,1 - 6,0
Côte d'Ivoire	2005	EIS	249	4368	8436	33,9	4,7	4,3 - 5,2
Éthiopie	2005	EDS	540	6689	10540	19,5	1,4	1,2 - 1,6
Ghana	2003	EDS	412	6251	9144	22,2	2,2	1,9 - 2,5
Guinée	2005	EDS	297	3126	6388	21,5	1,5	1,2 - 1,8
Kenya	2003	EDS	400	4234	6001	15,0	6,7	6,1 - 7,4
Lesotho	2004	EDS	405	4185	5043	12,5	23,5	22,3 – 24,7
Malawi	2004	EDS	522	4580	5150	9,9	11,8	10,9 - 12,7
Mali	2001	EDS	403	4087	6475	16,1	1,7	1,4 - 2,1
Niger	2006	EDS	345	3815	7262	21,0	0,7	0,5 - 0,9
Ouganda	2004	HSBS	417	9529	16906	40,5	6,4	6,0 - 6,8
Rwanda	2005	EDS	462	5136	10016	21,7	3,0	2,7 - 3,4
Sénégal	2005	EDS	377	2453	7503	19,9	0,7	0,5 - 0,9
Tanzanie	2003	AIS	345	6499	10747	31,2	7,0	6,8 - 7,2
Zambie	2001-02	EDS	320	2368	3807	11,9	15,6	14,5 - 16,8
Zimbabwe	2005-06	EDS	400	9285	12796	32,0	18,1	17,4 - 18,8

Sources: http://www.measuredhs.com et rapport final de chaque enquête..

EDS : Enquête démographique et de Santé ; AIS : AIDS Impact Survey ; HSBS : HIV/AIDS Sero-Behavioural Survey ; EIS : Enquête sur les Indicateurs du SIDA.

IC 95 % : Intervalle de Confiance à 95 %, calculé selon la méthode de Wilson avec correction de continuité.

* Il s'agit du nombre de ménages éligibles pour le dépistage du VIH et effectivement enquêtés.

Nous essaierons enfin d'estimer dans quelle mesure l'ensemble des ces biais ont un impact sur les estimations de la prévalence nationale du VIH (section 3.3.5). Les exemples pris tout au long de cette partie pourront concerner plusieurs pays d'Afrique subsaharienne pour lesquelles une enquête nationale en population générale avec dépistage du VIH a été réalisée. Cependant, nous nous intéresserons plus en détails au Burkina Faso, au Cameroun et au Kenya.

3.3.1 Populations hors ménage

Si la majorité de la population d'un pays vit au sein d'un ménage, une petite partie vit au sein d'institutions et est usuellement prise en compte différemment lors des recensements de population. Ainsi, au Burkina Faso, il s'agit principalement de personnes résidant dans « *des établissements hospitaliers, des établissements pénitentiaires, des campus universitaires, des établissements touristiques (hôtels*

et auberges), des internats privés ou publics scolaires, des couvents et monastères, des regroupements de travailleurs logés dans des baraquements sur leur chantier de travail et tout autre type d'établissement similaire » (INSTITUT NATIONAL DE LA STATISTIQUE ET DE LA DÉMOGRAPHIE 2000a). Lors du dernier recensement du Burkina Faso réalisé en 1996, 18 392 personnes ont été recensées au sein d'une institution contre 10 312 609 en ménage ordinaire, soit moins de 0,18 % de la population burkinabé totale et, si nous considérons que ces 18 392 individus ont entre 15 et 49 ans révolus, moins de 0,43 % des 15-49 ans.

Au Kenya, lors du recensement de 1999, 28 686 607 individus ont été recensés[15] dont 28 159 922 résidant en ménage ordinaire[16], ce qui induit une population hors ménage de 526 685 personnes, soit 1,84 % de la population. Sous l'hypothèse que l'ensemble des personnes hors ménage sont âgées de 15 à 49 ans, elles représenteraient 3,91 % de cette tranche d'âges[17].

Au Cameroun, le dernier recensement disponible date de 1987. Un nouveau recensement de la population est prévu depuis plusieurs années mais il a connu de nombreuses difficultés, notamment concernant son financement. L'EDS de 2004 a utilisé comme base de sondage la liste des zones de dénombrement établie lors des opérations de cartographie du troisième Recensement Général de la Population et de l'Habitat, menées par le BUCREP entre juin 2002 et avril 2003 (INSTITUT NATIONAL DE LA STATISTIQUE 2005, p. 12). Cependant, les opérations de dénombrement ont pris du retard et n'ont pas été effectuées avant fin 2005. Les résultats du troisième recensement ne sont donc pas encore disponibles. En 1987, 9 239 712 individus en ménage privé ont été recensés[18], tandis que la population cumulée des ménages privés et collectifs étaient de 9 312 429 personnes[19] (dont 4 003 637 âgés de 15 à 49 ans). Nous avons donc une population hors ménage de 72 717 individus, soit 0,78 % de la population nationale et 1,81 % des 15-49 ans sous l'hypothèse que l'ensemble des personnes hors ménage appartiennent à cette tranche d'âge.

---

[15] CENTRAL BUREAU OF STATISTICS, *Kenya 1999 Population and Housing Census - Volume I Population distribution by administrative areas and urban centers*, Nairobi (KE), Ministry of planning and national development, Central bureau of statistics, 2001, p. xxvii.

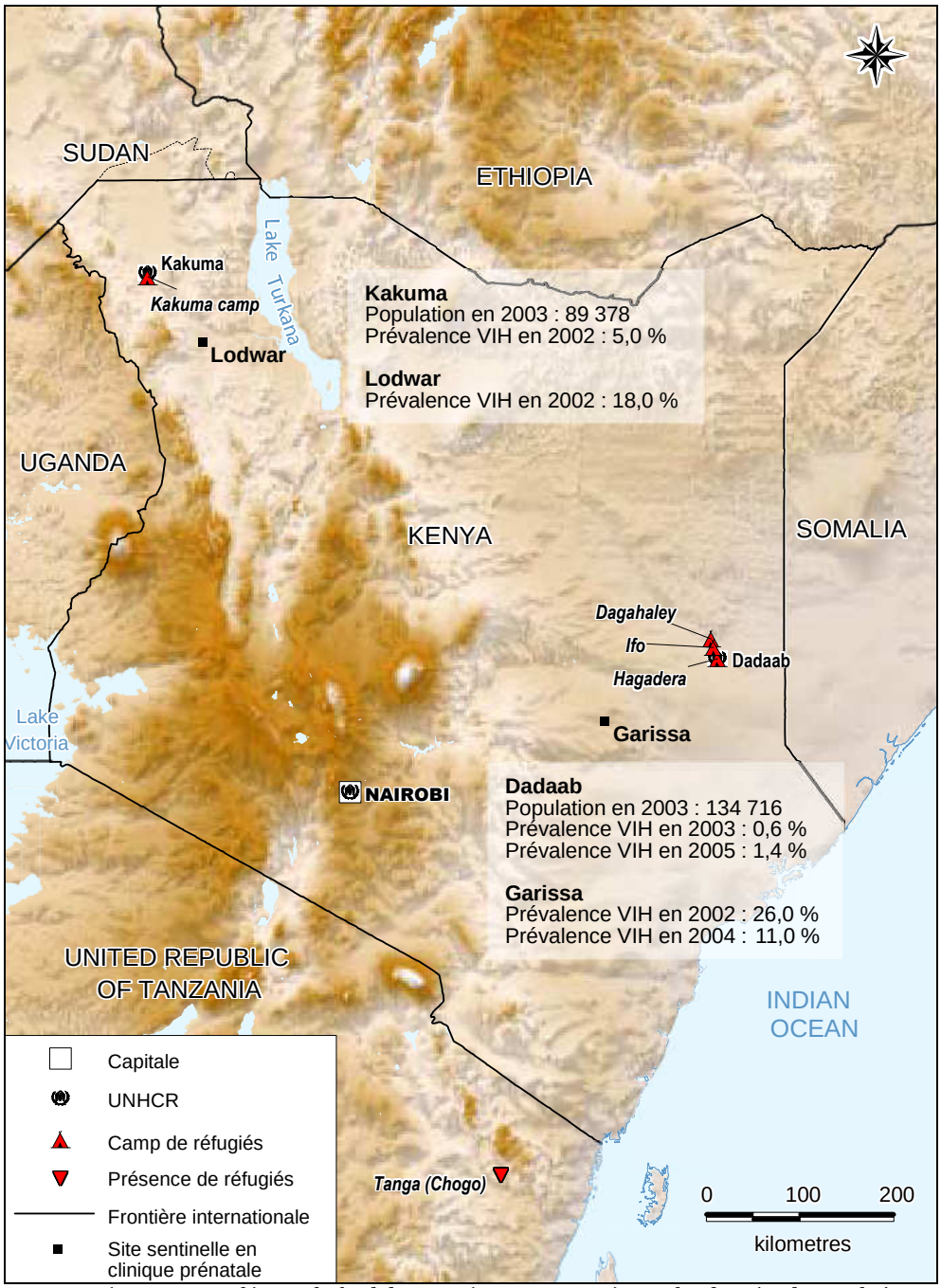
[16] CENTRAL BUREAU OF STATISTICS, *Kenya 1999 Population and Housing Census - Volume X Analytical Report on Housing Conditions and Household amenities*, Nairobi (KE), Ministry of planning and national development, Central bureau of statistics, 2002b, p. 15.

[17] On dénombre 13 472 852 individus de 15-49 ans. CENTRAL BUREAU OF STATISTICS, *Kenya 1999 Census - Vol. I Population distribution*, p. 2-1.

[18] DIRECTION NATIONALE DU DEUXIÈME RECENSEMENT DE LA POPULATION ET DE L'HABITAT, *Deuxième Recensement Général de la Population et de l'Habitat du Cameroun 1987 - Volume I Résultats bruts - Tome 1 République du Cameroun*, Yaoundé (CM), 1992, p. 597.

[19] DIRECTION NATIONALE DU DEUXIÈME RECENSEMENT DE LA POPULATION ET DE L'HABITAT, *RGPH 1987 Cameroun - Vol. I Tome 1*, p. 22-24.

Figure 3.9
Prévalence du VIH dans les camps de réfugiés et les sites sentinelles proches au Kenya (2002-2005)



Sources : (UNHCR 2004b) pour le fond de carte, (UNHCR 2004a) pour les données de population et (SPIEGEL 2007) pour les données de prévalence.

Si la proportion de personnes hors ménage reste en général relativement faible, la présence de camps de réfugiés dans un pays peut l'affecter grandement. Les effectifs de populations réfugiées sont relativement bien documentés par le Haut

Commissariat des Nations Unies pour les réfugiés (UNHCR). Fin 2003, l'UNHCR recensait 965 réfugiés au Burkina Faso, résidant en ville, 64 770 au Cameroun (47 787 en milieu urbain et 16 983 dispersés) mais aucun résidant en camps dans ces deux pays[20]. Au Kenya, l'UNHCR dénombrait 242 372 réfugiés dont 224 094 répartis en quatre camps : l'un près de Kakuma et les trois autres à proximité de Dadaab (voir Figure 3.9). La proportion de personnes âgées de 18 à 59 ans variait de 46 à 49 % dans les quatre camps (UNHCR 2004a, table 11) ce qui induit qu'environ la moitié de ces populations est âgée entre 15 et 49 ans révolus.

Il est souvent avancé que les conflits alimentent l'épidémie de VIH et induisent, conséquemment, une prévalence élevée parmi les réfugiés et les populations déplacées (CARBALLO 2001, SALAMA 2001, HANKINS 2002). Cependant, plusieurs études questionnent cette hypothèse et mettent en évidence que la dynamiques des épidémies de VIH dans ce type de contextes est plus complexe (MOCK 2004, SPIEGEL 2004, ALLEN 2006). Outre le fait que les conflits s'inscrivent dans des dynamiques épidémiques déjà en place, les conflits et les déplacements de populations produisent à la fois des facteurs favorisant et réduisant la propagation du VIH.

Dans une étude portant sur sept pays (République Démocratique du Congo, Soudan, Rwanda, Ouganda, Sierra Leone, Somalie et Burundi), à partir de données provenant de 65 articles publiés, Paul SPIEGEL et ses collaborateurs n'ont pu mettre en évidence une augmentation de la prévalence du VIH pendant les périodes de conflits. Sur les 12 groupes de camps de réfugiés étudiés, neuf présentent une prévalence plus faible, deux une prévalence de même ordre et un une prévalence plus élevée que celle mesurée parmi les populations d'accueil (SPIEGEL 2007). Au Kenya, les camps de réfugiés de Dadaab présentent une prévalence entre 0,6 et 1,4 % tandis que la site sentinelle de Garissa, à proximité, a enregistré une prévalence de 11 à 26 %. En 2002, la population du camp de Kakuma présentait une prévalence de 5 % contre les 18 % observés parmi les femmes enceintes de Lodwar (Figure 3.9).

Suivant les conflits, la population des camps de réfugiés peut être amenée à fluctuer de manière importante. Entre 1999 et 2003, ce ne fut pas le cas au Kenya puisque la population des camps de Dadaab était estimée à 124 600 fin 1999, et celle du camp de Kakuma à 86 500, les 15-49 ans représentant toujours environ la moitié de l'effectif[21]. Si l'on retire la population des camps de réfugiés de la population

[20] UNHCR, *2003 Global Refugee Trends: Overview of refugee populations, new arrivals, durable solutions, ayslum-seekers and other persons of concern to UNHCR*, Genève (CH), UNHCR, 2004a. Table 12.

[21] UNHCR, *Refugees and others of concern to UNHCR - 1999 statiscal overview*, Genève (CH), UNHCR, 2000. Table III.4.

hors ménage, cette dernière s'élève alors à 315 585, soit 1,10 % de la population générale et 2,34 % des 15-49 ans sous l'hypothèse que la totalité de la population hors ménage appartient à ce groupe d'âges. En 2003, la population kenyane de 15-49 ans était estimée à 15 782 110[22]. Nous pouvons considérer que les 15-49 ans étaient, en 2003, de 44 689 à Kakuma et de 67 358 à Dadaab soit en tout 112 047 personnes ou encore 0,71 % de l'ensemble de la population adulte kenyane. La prévalence du VIH parmi la population adulte des camps de réfugiés était alors en 2003 de 2,35 % ([5,0 % × 44 689 + 0,6 % × 67 358] / [44 689 + 67 358]).

Pour que la population hors ménage constitue un biais dans l'estimation de la prévalence nationale, il faudrait que sa prévalence diffère de celle des personnes vivant en ménage. Il est difficile voir impossible d'estimer la prévalence du VIH parmi les personnes hors ménages. Si, selon les pays, nous disposons d'enquêtes de prévalence menées auprès de populations spécifiques, telles que les militaires ou les patients atteints d'IST, ces populations ne recoupent que partiellement les personnes hors ménages. Les militaires dormant dans des dortoirs au sein d'une caserne seront comptabilisés parmi les personnes hors ménages, tandis que ceux habitant dans des logements individuels constitueront des ménages ordinaires. De même, les personnes atteintes d'IST ne représentent qu'une partie de l'ensemble des personnes hospitalisées.

Pour estimer l'ampleur du biais imputable aux individus vivant hors ménage, nous poserons deux hypothèses extrêmes en considérant que ces personnes présentent une prévalence égale à la moitié ou au double de celle observée parmi le reste de la population. Une prévalence corrigée sera ensuite calculée à partir de l'Équation 3.4.

Équation 3.4

Correction de la prévalence du VIH pour prendre en compte la population hors ménage et celle des camps de réfugiés

$$P_{corr} = F_{hm} \cdot PR_{hm} \cdot P_{mén} + P_{cr} \cdot F_{cr} + P_{mén} \cdot (1 - F_{hm} - F_{cr})$$

P_{corr} : *prévalence corrigée*

$P_{mén}$: *prévalence de la population en ménage ordinaire*

P_{cr} : *prévalence observée dans les camps de réfugiés*

F_{hm} : *proportion d'individus hors ménages (camps de réfugiés exclus)*

F_{cr} : *proportion d'individus en camps de réfugiés*

PR_{hm} : *prévalence relative des personnes hors ménage*

[22] CENTRAL BUREAU OF STATISTICS, *Kenya 1999 Population and Housing Census - Volume VII Analytical Report on Population Projections*, Nairobi (KE), Ministry of planning and national development, Central bureau of statistics, 2002a, p. 30. Table 4.3

Tableau 3.6
Correction de la prévalence du VIH, 15-49 ans, observée dans les EDS, selon la population hors ménage, au Burkina Faso, Cameroun et Kenya

	Burkina Faso 2003	Cameroun 2004	Kenya 2003
Prévalence EDS (%)	1,77	5,44	6,88*
Proportion d'adultes hors ménage (%)†	0,43	1,81	2,34
Proportion d'adultes en camps de réfugiés (%)	0	0	0,71
Prévalence des adultes en camps de réfugiés (%)	-	-	2,35
Prévalence relative, hypothèse basse‡	½	½	½
Prévalence relative, hypothèse haute‡	2	2	2
Prévalence corrigée hypothèse basse (%)	1,77	5,39	6,77
Prévalence corrigée hypothèse haute (%)	1,78	5,54	7,01

* Le Tableau 3.5 présente une prévalence de 6,7 % telle que publiée dans le rapport final de l'enquête. Nous affichons ici une prévalence de 6,88 % telle que calculée, à partir des bases de données, lorsque l'on utilise les taux de pondération du fichier ménage. Avec les taux de pondération fournis avec le fichier VIH, nous retrouvons la valeur de 6,7 %. Il semble donc que les taux de pondération de ce fichier prennent en compte d'autres correctifs que ceux mentionnés dans l'Encadré 3.1. Nous avons préféré, pour notre part, travailler avec les taux de pondération du fichier ménage, d'une part afin d'appliquer nos différents correctifs les uns après les autres et d'autre part parce que ce sont les seuls taux de pondération disponibles pour les personnes non testées que nous aborderons à la section 3.3.4).

† Camps de réfugiés exclus, sous l'hypothèse que l'ensemble des individus hors ménage font partie des 15-49 ans.

‡ Prévalence relative de la population hors ménage (camps de réfugiés exclus) par rapport à la population vivant en ménages.

En raison de la faible proportion que représentent les individus hors ménage au sein de la population totale, leur impact sur les estimations nationales de la prévalence du VIH reste limité (Tableau 3.6). Par ailleurs, le biais qu'ils induisent est inférieur à l'estimation que nous en donnons dans la mesure où nos différentes hypothèses maximisent ce biais. En effet, les personnes vivant hors ménage n'ont pas toutes entre 15 et 49 ans. Leur part dans la population adulte totale est donc plus faible que ce que nous avons estimé. Par ailleurs, il est fort probable qu'il n'y ait pas autant d'écarts de prévalence entre population hors ménage et population en ménages ordinaire. Les prévalences relatives devraient donc se situer plus certainement entre 0,75 et 1,5.

Encadré 3.1
Calcul des taux de pondération des EDS

Nous présentons le mode de calcul des taux de pondération utilisés pour l'EDS du Burkina Faso. Ce mode de calcul est relativement standard mais certaines enquêtes peuvent présenter de légères variations.

Deux probabilités de sondage sont calculées :

- P_{1hi} : probabilité de sondage au premier degré de la grappe *i* de la strate *h*.
- P_{2hi} : probabilité de sondage au second degré des ménages de la grappe *i* de la strate *h*.

Nous avons alors :

$$P_{1hi} = \frac{a_h \times M_{hi}}{\sum_i M_{hi}}$$

$$P_{2hi} = \frac{b_{hi}}{L_{hi}}$$

a_h : *nombre de grappes tirées dans la strate h*
M_{hi} : *nombre de ménages de la grappe i selon le recensement*
L_{hi} : *nombre de ménages de la grappe i après redénombrement de la grappe i*
b_{hi} : *nombre de ménages tirés au sort dans la grappe i*

Pour chaque strate *h* d'échantillonnage, le taux de pondération W_{hi} d'un individu des ménages de la grappe *i* est alors de :

$$W_{hi} = \frac{1}{P_{1hi} \times P_{2hi}}$$

Si le taux de sondage diffère selon les strates, un dernier correctif est alors appliqué pour rendre l'échantillon représentatif au niveau national.

3.3.2 Ancienneté de la base de sondage

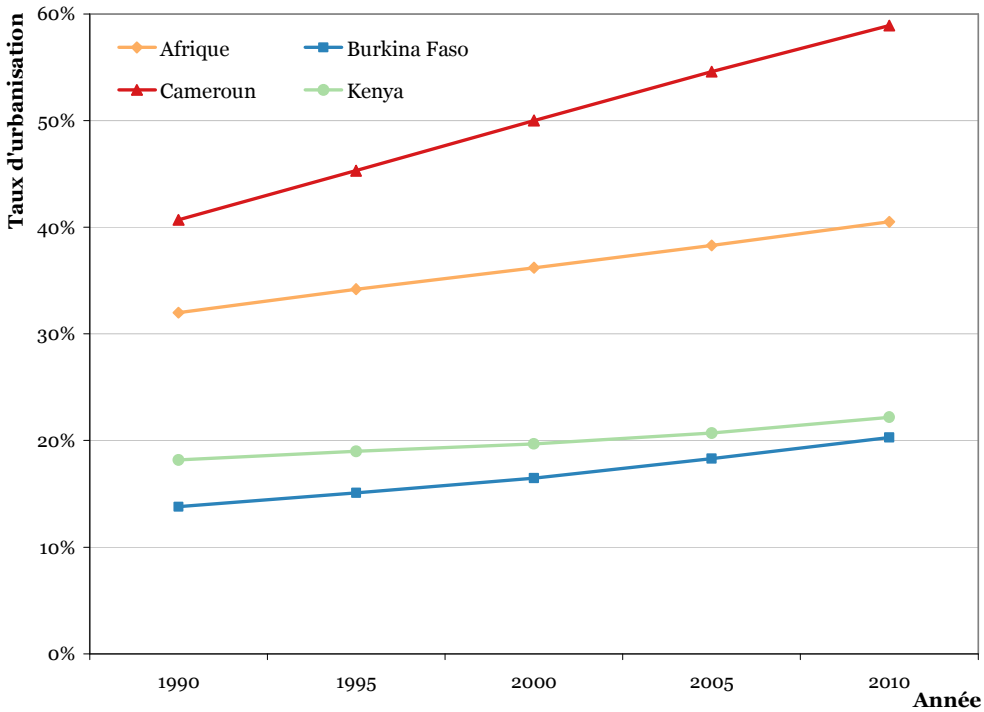
La base de sondage utilisée pour le tirage au premier degré correspond le plus souvent au dernier recensement de la population effectué dans le pays. Comme le montre le Tableau 3.4, ce recensement peut être parfois relativement ancien par rapport à la période de l'enquête. Au Burkina Faso, il s'est ainsi écoulé 7 ans entre le recensement de 1996 et l'EDS de 2003. L'écart est moins marqué au Cameroun (respectivement 2002-2003 et 2004). Quant au Kenya, l'EDS de 2003 a utilisé la fourth National Sample Survey and Evaluation Programme (NASSEP IV). Cette base de sondage a été construire à partir des zones de dénombrement du recensement de 1999. La liste des ménages de chaque zone a été réactualisée en

2002. Seules 50 grappes sur les 400 tirées au sort pour l'EDS 2003 ont fait l'objet d'un nouveau recensement des ménages en 2003[23].

Chaque grappe est tirée au sort avec une probabilité proportionnelle à son nombre de ménages. Or le nombre de ménages utilisé pour calculer ces probabilités correspond au nombre de ménages ordinaires recensés lors du recensement utilisé (à l'exception du Kenya dans le cas présent). Cette probabilité entre par la suite dans le calcul des coefficients de pondération utilisés pour rendre la base de données représentative (Encadré 3.1).

Or, la population de chaque pays évolue continuellement et il y a donc un écart entre la répartition des ménages au moment du recensement utilisé comme base de sondage et la répartition des ménages au moment de l'enquête. Si cette évolution est inégale selon les zones, les zones ayant connu la moins forte croissance seront surreprésentées tandis que les zones présentant la plus forte progression seront sous-représentées.

Figure 3.10
Évolution du taux d'urbanisation en Afrique, au Burkina Faso, au Cameroun et au Kenya (1990-2010)

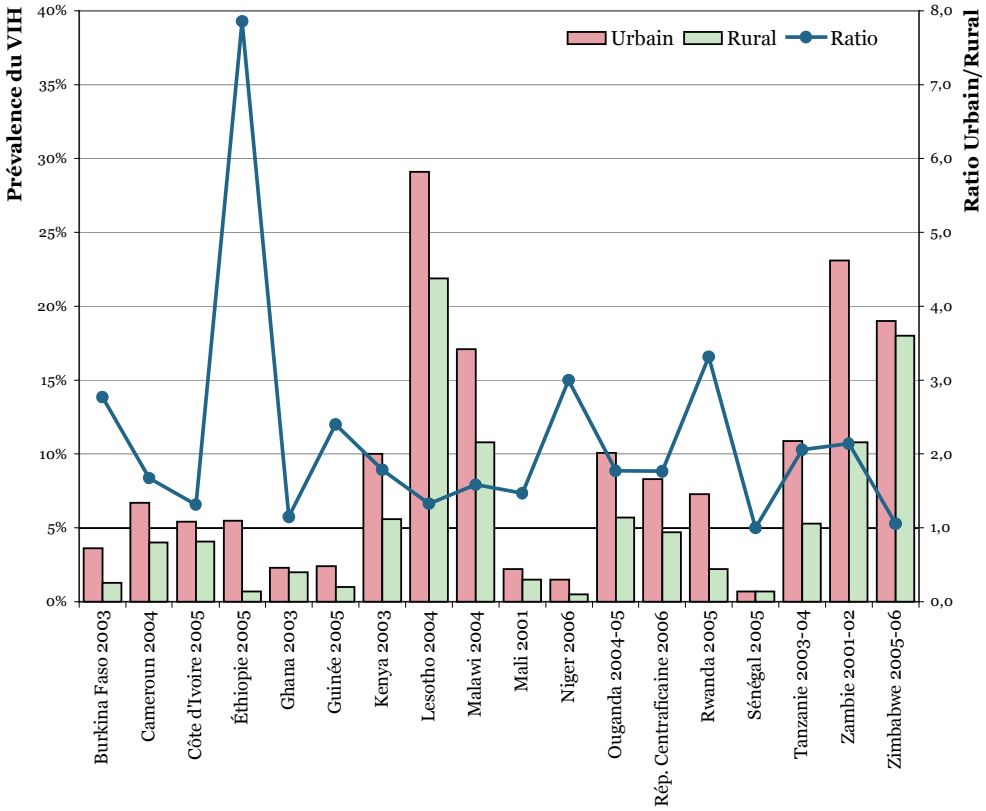


Sources : World Population Prospects et World Urbanization Prospects (UNITED NATIONS POPULATION DIVISION 2007).

[23] CENTRAL BUREAU OF STATISTICS, MINISTRY OF HEALTH et ORC MACRO, *Kenya Demographic and Health Survey 2003*, Calverton, Maryland (US), CBS, MOH, ORC Macro, 2004.

Figure 3.11

Prévalence nationale du VIH à 15-49 ans selon le milieu de résidence et ratio prévalence urbaine sur prévalence rurale, pour 18 pays d'Afrique subsaharienne



Sources : rapport final de chaque enquête disponible sur http://www.measuredhs.com. Il s'agit d'Enquêtes Démographiques et de Santé (EDS), à l'exception de la Côte d'Ivoire (Enquête sur les Indicateurs du SIDA – EIS), de l'Ouganda (HIV/AIDS Sero-Behavioural Survey – HSBS), de la République Centrafricaine (Enquête de Sérologie VIH – ESV) et de la Tanzanie (AIDS Impact Survey – AIS). Ce sont toutes des enquêtes nationales réalisées en population générale.

La majorité des pays d'Afrique connaissent depuis plusieurs années une urbanisation croissante (Figure 3.10). Par ailleurs, la prévalence du VIH diffère significativement, pour une majorité de pays, entre milieu urbain et milieu rural (Figure 3.11), le milieu urbain présentant des prévalences plus élevées. De ce fait, les estimations nationales de la prévalence du VIH selon les EDS devraient être légèrement sous-estimées dans la mesure où, le taux d'urbanisation ayant augmenté entre la base de sondage et le moment de l'enquête, le milieu urbain devrait être sous-représenté.

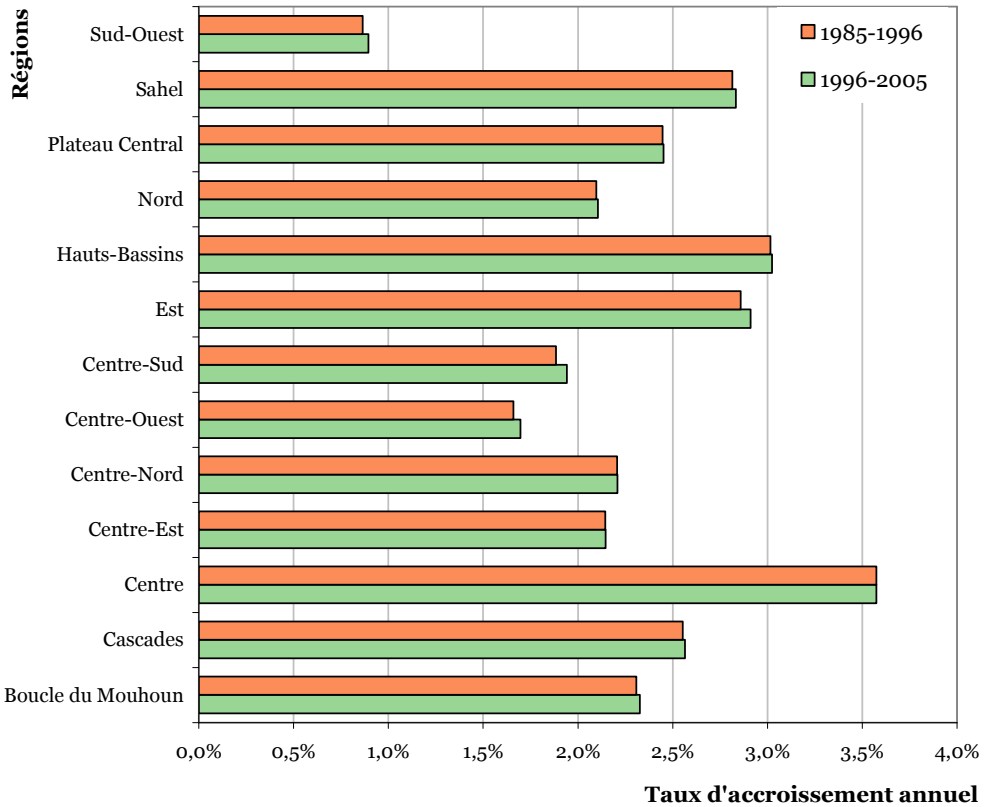
La situation est équivalente en ce qui concerne les régions d'un pays. D'une part, elles connaissent une croissance de population inégale (Figure 3.12). D'autre part, la prévalence du VIH varie fortement d'une région à l'autre (Figure 3.13). Les variations régionales recoupent par ailleurs une bonne partie des variations selon le milieu de résidence, les régions qui croissent le plus étant en général celles connaissant une urbanisation importante. Par ailleurs, les migrations, internes ou

internationales, sont le plus souvent différenciées selon le sexe et la prévalence du VIH varie également fortement selon le sexe (voir Figure 3.19 page 184).

Pour corriger ces effets de sur ou sous représentation de certaines zones géographiques, nous pouvons ajuster les prévalences observées par la structure de la population au moment de l'enquête. L'Institut National de la Statistique et de la Démographie (INSD) du Burkina Faso a effectué des projections de population à partir des données des recensements de 1985 et 1996 (INSTITUT NATIONAL DE LA STATISTIQUE ET DE LA DÉMOGRAPHIE 2004). Les résultats publiés sont suffisamment détaillées pour calculer la structure par sexe et région de la population des adultes de 15 à 49 ans en 2003 (Tableau 3.7). Le Central Bureau of Statistics du Kenya a réalisé un travail similaire permettant également de calculer la structure par région et sexe de la population adulte (CENTRAL BUREAU OF STATISTICS 2002a). Au Cameroun, en raison de l'ancienneté du dernier recensement disponible (1987), nous ne disposons pas de projections de population aussi détaillées. Néanmoins, l'Institut National de la Statistique (INS) fournit la structure par région de la population en 2004 (INSTITUT NATIONAL DE LA STATISTIQUE 2006).

Figure 3.12

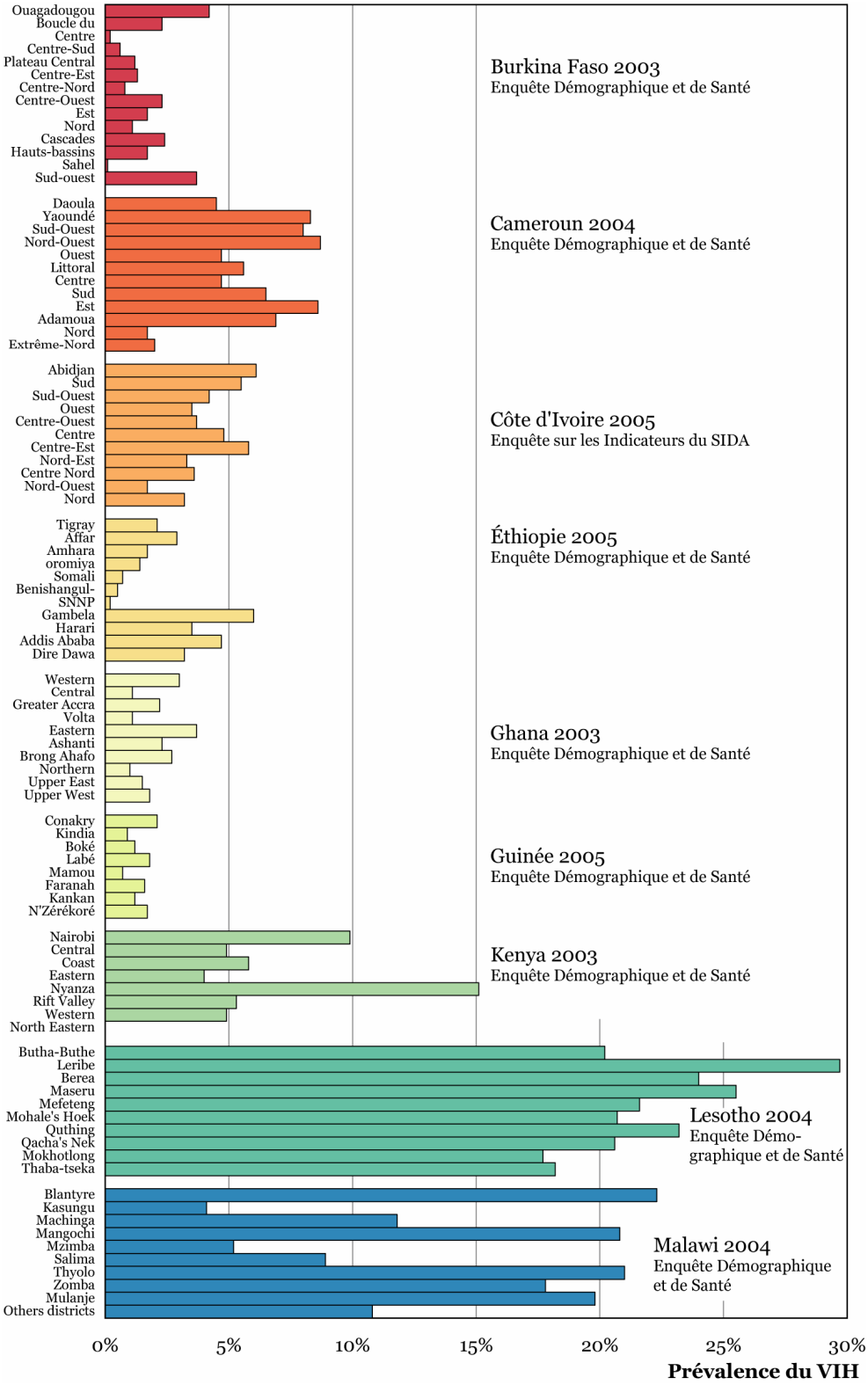
Taux d'accroissement naturel de la population, par région, au Burkina Faso, sur la période 1985-2005

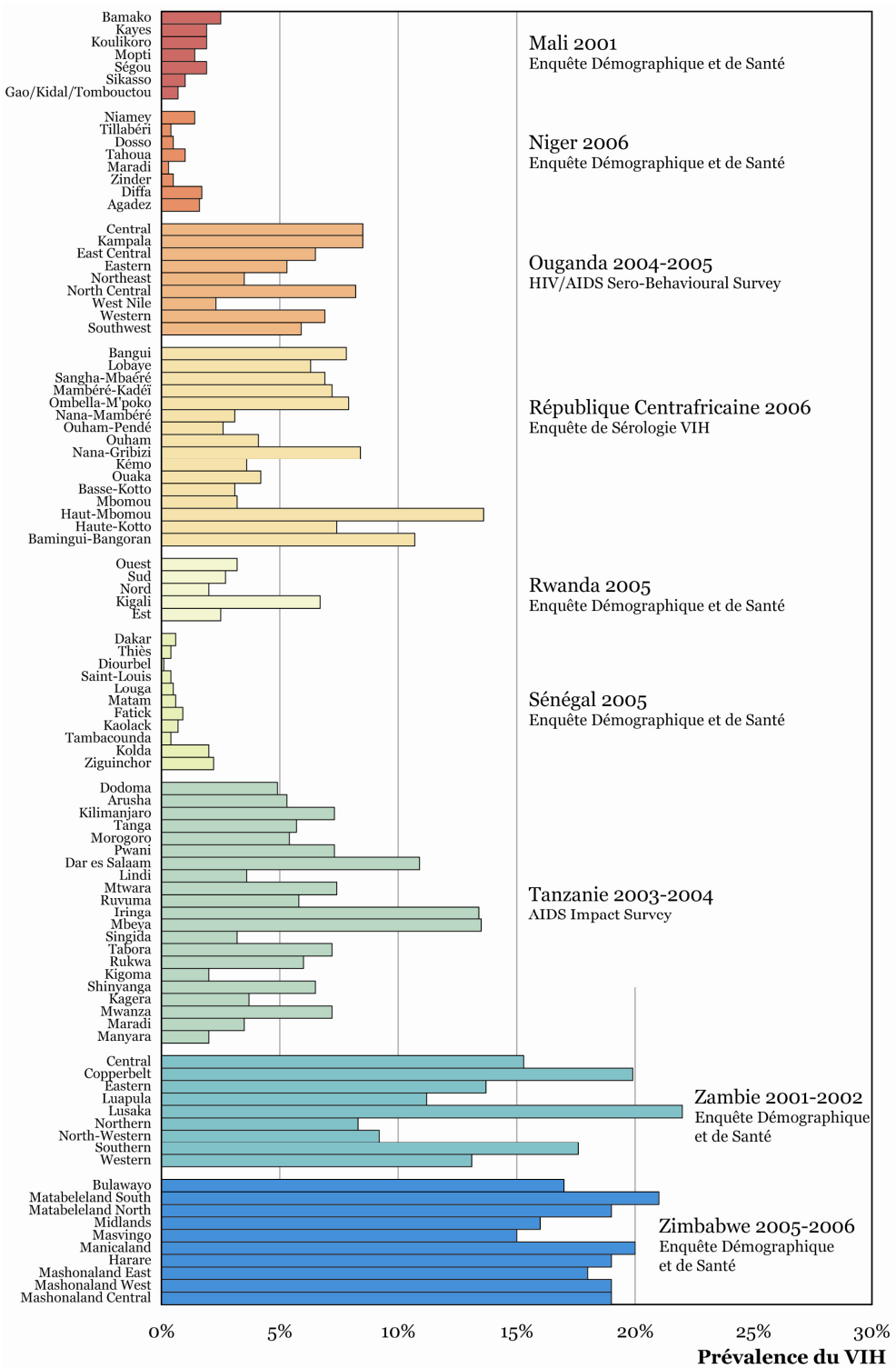


Sources : RGPH 1985 et 1996, projections pour 2005, (INSTITUT NATIONAL DE LA STATISTIQUE ET DE LA DÉMOGRAPHIE 2004), nos calculs.

Figure 3.13

Prévalence du VIH, 15-49 ans, par région, pour 18 pays d'Afrique subsaharienne





Sources : rapport final de chaque enquête et/ou *HIV Fact Sheets* disponibles sur http://www.measuredhs.com.

Tableau 3.7

Structure par sexe et région des 15-49 ans du Burkina Faso et du Kenya en 2003 et structure par région de la population du Cameroun en 2004

Région	Effectif hommes	Effectif femmes	Proportion hommes	Proportion femmes
Burkina Faso 2003				
Boucle du Mouhoun	256 822	283 185	5,2 %	5,7 %
Centre	306 364	294 653	6,2 %	6,0 %
Centre-Sud	98 013	133 407	2,0 %	2,7 %
Plateau Central	107 886	152 707	2,2 %	3,1 %
Centre-Est	163 624	218 955	3,3 %	4,4 %
Centre-Nord	181 609	243 096	3,7 %	4,9 %
Centre-Ouest	164 355	238 861	3,3 %	4,8 %
Est	194 598	219 891	3,9 %	4,5 %
Nord	176 436	235 852	3,6 %	4,8 %
Cascades	72 255	87 337	1,5 %	1,8 %
Hauts-Bassins	262 214	284 034	5,3 %	5,8 %
Sahel	173 163	186 432	3,5 %	3,8 %
Sud-Ouest	86 231	113 784	1,7 %	2,3 %
Ensemble	*2 243 570*	*2 692 194*	*45,5 %*	*54,5 %*
TOTAL	*4 935 764*		*100 %*	
Cameroun 2004				
Adamoua	782 000		4,6 %	
Centre	2 703 000		15,9 %	
Est	816 000		4,8 %	
Extrême-Nord	2 941 000		17,3 %	
Littoral	2 380 000		14,0 %	
Nord	1 326 000		7,8 %	
Nord-Ouest	1 989 000		11,7 %	
Ouest	2 142 000		12,6 %	
Sud	578 000		3,4 %	
Sud-Ouest	1 343 000		7,9 %	
TOTAL	*17 000 000*		*100,0 %*	
Kenya 2003				
Nairobi	851 350	752 524	5,4 %	4,8 %
Central	976 129	1 088 321	6,2 %	6,9 %
Coast	619 772	787 245	3,9 %	5,0 %
Eastern	1 111 344	1 228 796	7,1 %	7,8 %
Nyanza	1 117 641	1 230 766	7,1 %	7,8 %
Rift Valley	1 839 149	1 898 218	11,7 %	12,1 %
Western	779 839	870 976	5,0 %	5,5 %
North Eastern	280 261	311 116	1,8 %	2,0 %
Ensemble	*7 575 485*	*8 167 962*	*48,1 %*	*51,9 %*
TOTAL	*15 743 447**		*6,88 %*	

* Le total présenté ici diffère de celui utilisé dans la section 3.3.1 et qui était de 15 782 110. Cet écart est dû au fait que la projection nationale est réalisée séparément des projections par régions. Cependant, cet écart n'influe pas significativement sur le calcul de la proportion d'adultes vivant dans un camp de réfugié.

Sources : nos calculs à partir de (CENTRAL BUREAU OF STATISTICS 2002a, INSTITUT NATIONAL DE LA STATISTIQUE ET DE LA DÉMOGRAPHIE 2004, INSTITUT NATIONAL DE LA STATISTIQUE 2006).

Tableau 3.8

Prévalence du VIH, 15-49 ans, par région, selon les EDS 2003 du Burkina Faso et du Kenya et l'EDS 2004 du Cameroun

Région	Effectif hommes	Effectif femmes	Prévalence hommes	Prévalence femmes
Burkina Faso 2003				
Boucle du Mouhoun	229	315	2,39 %	2,16 %
Centre	390	509	3,06 %	4,00 %
Centre-Sud	188	244	0,42 %	0,73 %
Plateau Central	146	224	0,94 %	1,35 %
Centre-Est	218	359	0,93 %	1,55 %
Centre-Nord	227	365	1,13 %	0,63 %
Centre-Ouest	203	323	3,11 %	1,80 %
Est	238	307	3,17 %	0,55 %
Nord	264	404	1,27 %	0,92 %
Cascades	92	133	1,81 %	2,77 %
Hauts-Bassins	452	563	0,93 %	2,10 %
Sahel	237	293	0,00 %	0,25 %
Sud-Ouest	149	202	4,51 %	3,69 %
Ensemble	*3 033*	*4 242*	*1,78 %*	*1,76 %*
TOTAL	*7 275*		*1,77%*	
Cameroun 2004				
Adamoua	435		6,93%	
Centre	1 709		6,66%	
Est	491		8,37%	
Extrême-Nord	1 462		1,97%	
Littoral	1 560		4,90%	
Nord	890		1,61%	
Nord-Ouest	1 062		8,61%	
Ouest	968		4,66%	
Sud	463		6,44%	
Sud-Ouest	807		7,97%	
TOTAL	*9 848*		5,44%	
Kenya 2003				
Nairobi	224	251	7,78 %	11,93 %
Central	378	448	1,99 %	7,63 %
Coast	186	270	4,51 %	7,05 %
Eastern	480	537	1,99 %	5,93 %
Nyanza	470	558	11,36 %	17,93 %
Rift Valley	705	810	3,46 %	6,80 %
Western	372	446	3,71 %	5,64 %
North Eastern	49	61	0,00 %	0,00 %
Ensemble	*2 864*	*3 380*	4,70 %	8,74 %
TOTAL	*6 244*		*6,88 %*	

* Ouagadougou a été réintégrée à la région Centre.

† Yaoundé a réintégrée à la région Centre.

‡ Douala a été réintégrée à la région Littoral.

Sources : nos calculs à partir des bases de données des EDS disponibles sur www.measuredhs.com.

Pour aucun de ces pays, nous ne disposons d'informations permettant d'avoir une structure à la fois par région et par milieu de résidence. Cependant, nous pouvons considérer que la majorité de l'évolution de l'urbanisation du pays sera rendue par l'évolution de la population de chaque région, dans la mesure où les évolutions du taux d'urbanisation et du poids de chaque région se recoupent en partie.

Une fois les prévalences du VIH calculées par région et par sexe à partir des données des EDS (Tableau 3.8), l'ajustement de la prévalence nationale se calcule ainsi :

Équation 3.5

Ajustement de la prévalence selon la structure par région et par sexe

$$P_{ajustée} = \sum_{sexe} \sum_{régions} F_{sr} \cdot P_{sr}$$

F_{sr} : *proportion de la population de sexe s appartenant à la région r*

P_{sr} : *prévalence des individus de sexe s de la région r*

Les résultats (Tableau 3.9) montrent que l'ajustement selon la structure par région et par sexe ne fait que peu varier la prévalence nationale du VIH. En effet, pour l'EDS du Burkina Faso, la prévalence ajustée est plus élevée que la prévalence observée d'à peine 1,13 %, tandis que les prévalences ajustées du Cameroun et du Kenya sont inférieures respectivement de 1,29 % et 2,62 % aux prévalences observées.

Tableau 3.9

Ajustement de la prévalence du VIH, 15-49 ans, des EDS 2003 et 2004 du Burkina Faso, du Cameroun et du Kenya à partir de la structure par région et par sexe

	Burkina Faso 2003	Cameroun 2004	Kenya 2003
Prévalence des femmes dans l'EDS (%)	1,76	-	8,74
Prévalence des femmes ajustée par régions (%)	1,73	*	8,57
Prévalence des hommes dans l'EDS (%)	1,78	-	4,70
Prévalence des hommes ajustée par régions (%)	1,86	*	4,69
Prévalence nationale dans l'EDS (%)	1,77	5,44	6,88
Prévalence ajustée par région et par sexe (%)	1,79	5,37	6,70

* Pour le Cameroun, nous ne disposions que d'une structure par région. Seul un ajustement national a donc été calculé.

3.3.3 Ménages non enquêtés

Après sélection des différentes zones d'enquêtes, un recensement exhaustif de l'ensemble des ménages de chaque grappe est effectué afin de constituer la base de sondage du tirage au second degré. Un nombre prédéterminé de ménages est alors tiré au sort dans chaque grappe. Cependant, parmi les ménages sélectionnés pour l'enquête, une certaine partie n'est pas identifiée lors du retour sur le terrain pour la passation du questionnaire ménages. Par ailleurs, certains ménages ne peuvent être enquêtés pour différentes raisons : absence lors des différents passages des enquêteurs, refus de participer à l'enquête, etc. Le Tableau 3.10 présente les taux de réponse à l'enquête ménages pour dix-sept enquêtes nationales en population générale. Les taux de réponses varient de 94,6 à 99,6 % ce qui est considéré usuellement comme étant de bons taux de réponses. Par ailleurs, une majorité de pays (11 sur 17) présentent une proportion de ménages non enquêtés inférieure à 2,5 %.

Tableau 3.10

Couverture des enquêtes ménages de 17 enquêtes nationales récentes, en population générale, avec dépistage du VIH

Pays	Année	Type	Ménages sélectionnés*	Ménages identifiés	Ménages enquêtés	Taux de réponse (%)
Burkina Faso	2003	EDS	3 297	3 203	3 179	99,3
Cameroun	2004	EDS	5 884	5 481	5 319	97,0
Côte d'Ivoire	2005	EIS	4 498	4 573	4 368	95,5
Éthiopie	2005	EDS	7 160	6 787	6 689	98,6
Ghana	2003	EDS	6 628	6 333	6 251	98,7
Guinée	2005	EDS	3 240	3 157	3 126	99,0
Kenya	2003	EDS	4 868	4 396	4 234	96,3
Lesotho	2004	EDS	4 863	4 426	4 185	94,6
Malawi	2004	EDS	5 029	4 690	4 580	97,7
Mali	2001	EDS	4 541	4 188	4 087	97,6
Niger	2006	EDS	4 210	3 893	3 815	98,0
Ouganda	2004	HSBS	10 437	9 842	9 529	96,8
Rwanda	2005	EDS	5 322	5 156	5 136	99,6
Sénégal	2005	EDS	2 614	2 499	2 453	98,2
Tanzanie	2003	AIS	6 901	6 595	6 499	98,5
Zambie	2001-02	EDS	2 658	2 408	2 368	98,3
Zimbabwe	2005-06	EDS	10 752	9 978	9 285	95,0

Sources: http://www.measuredhs.com et rapport final de chaque enquête.

EDS : Enquête démographique et de Santé ; AIS : AIDS Impact Survey ; HSBS : HIV/AIDS Sero-Behavioural Survey ; EIS : Enquête sur les Indicateurs du SIDA.

* Il s'agit du nombre de ménages sélectionnés pour le questionnaire hommes et le dépistage du VIH.

Nous ne disposons d'aucune information concernant ces ménages. Il est donc difficile d'estimer quelle peut être la prévalence du VIH parmi les adultes composant ces ménages. Nous ne savons pas si ces ménages comptent plus ou moins d'individus, en moyenne, que les ménages effectivement enquêtés. Cependant, il est raisonnable de considérer que la taille moyenne des ménages non enquêtés est équivalente à celle des ménages enquêtés, auquel cas la proportion d'individus non enquêtés car vivant dans ces ménages est égale au taux de non réponse des ménages.

Équation 3.6

Ajustement de la prévalence du VIH en tenant compte du taux de non réponse des ménages

$$P_{aj.} = P_{obs} \times (1 - T_{NRM}) + P_{obs} \times PR \times T_{NRM}$$

$P_{aj.}$: *prévalence ajustée*

P_{obs} : *prévalence observée*

T_{NRM} : *taux de non réponse des ménages*

PR : *prévalence relative des ménages non enquêtés*

Tout comme pour les individus vivant hors ménage (section 3.3.1), nous pouvons seulement tenter d'encadrer le biais induit en posant deux hypothèses extrêmes, c'est-à-dire en considérant que la prévalence du VIH parmi ces personnes est égale au double ou à la moitié de la prévalence observée parmi les ménages effectivement enquêtés. Nous pouvons alors calculer une prévalence ajustée à partir de l'Équation 3.6. Nous procédons à un ajustement national et non région par région dans la mesure où la majorité rapports finaux ne fournit pas le détail des taux de réponse par région. Les résultats sont présentés dans le Tableau 3.11.

Pour la plupart des pays, le biais induit par la proportion d'individus hors ménage reste mineur, les prévalences ajustées se situant au sein de l'intervalle de confiance à 95 % de la prévalence observée. L'imprécision due aux ménages non enquêtés est alors inférieure à l'imprécision liée à l'erreur aléatoire de l'échantillonnage. Pour deux pays, le Lesotho et le Zimbabwe, la prévalence ajustée dans le cadre de notre hypothèse haute (les individus des ménages non enquêtés présentent une prévalence double) se situe en dehors de l'intervalle de confiance à 95 % de la prévalence observée. Ces deux pays présentent les taux de non réponse les plus élevés (respectivement 5,4 et 5,0 %) ainsi que les prévalences observées les plus élevées (d'où une erreur aléatoire moindre et des intervalles de confiance proportionnellement plus serrés). Sous l'hypothèse d'une prévalence relative de 2, la prévalence ajustée se situe légèrement au-dessus de l'intervalle de confiance à 95 %. Nous avons déjà évoqué précédemment (section 3.3.1) que l'hypothèse d'une prévalence relative de 2 était une hypothèse de maximisation du biais mais qu'il était plus probable que la prévalence relative se situe entre 0,75 et 1,5, notamment

si nous extrapolons aux ménages non enquêtés les résultats que nous allons obtenir concernant les individus non testés (section 3.3.4). Or, sous l'hypothèse d'une prévalence relative de 1,5, la prévalence ajustée pour le Lesotho et le Zimbabwe est alors respectivement de 24,13 % et de 18,55 %. Si nous devons rester prudent, il semble néanmoins que l'imprécision due aux ménages non enquêtés reste inférieure à celle de l'erreur aléatoire.

Tableau 3.11

Ajustement de la prévalence du VIH selon la proportion de ménages non enquêtés pour 17 enquêtes nationale en population générale

Pays	Année	Prévalence observée (%)	IC 95 %	Ménages non enquêtés (%)	Prévalence ajustée (1/2)*	Prévalence ajustée (2)†
Burkina Faso	2003	1,8	1,5 - 2,1	0,7	1,79	1,81
Cameroun	2004	5,5	5,1 - 6,0	3,0	5,42	5,67
Côte d'Ivoire	2005	4,7	4,3 - 5,2	4,5	4,59	4,91
Éthiopie	2005	1,4	1,2 - 1,6	1,4	1,39	1,42
Ghana	2003	2,2	1,9 - 2,5	1,3	2,19	2,23
Guinée	2005	1,5	1,2 - 1,8	1,0	1,49	1,52
Kenya	2003	6,7	6,1 - 7,4	3,7	6,58	6,95
Lesotho	2004	23,5	22,3 – 24,7	5,4	22,87	24,77
Malawi	2004	11,8	10,9 - 12,7	2,3	11,66	12,07
Mali	2001	1,7	1,4 - 2,1	2,4	1,68	1,74
Niger	2006	0,7	0,5 - 0,9	2,0	0,69	0,71
Ouganda	2004	6,4	6,0 - 6,8	3,2	6,30	6,60
Rwanda	2005	3,0	2,7 - 3,4	0,4	2,99	3,01
Sénégal	2005	0,7	0,5 - 0,9	1,8	0,69	0,71
Tanzanie	2003	7,0	6,8 - 7,2	1,6	6,95	7,11
Zambie	2001-02	15,6	14,5 - 16,8	1,7	15,47	15,87
Zimbabwe	2005-06	18,1	17,4 - 18,8	5,0	17,65	19,01

Sources: http://www.measuredhs.com et rapport final de chaque enquête.

EDS : Enquête démographique et de Santé ; AIS : AIDS Impact Survey ; HSBS : HIV/AIDS Sero-Behavioural Survey ; EIS : Enquête sur les Indicateurs du SIDA.

IC 95 % : intervalle de confiance à 95 % de la proportion observée.

La prévalence est ajustée sous l'hypothèse que la proportion d'individus appartenant aux ménages non enquêtés correspond à la proportion de ces mêmes ménages parmi l'ensemble des ménages sélectionnés.

Les résultats sont exprimés en pourcents.

NB : pour ce tableau, nous avons retenu les prévalences observées publiées dans les rapports finaux.

* Hypothèse basse : la prévalence des personnes non enquêtées est égale à la moitié de la prévalence observée.

† Hypothèse haute : la prévalence des personnes non enquêtées est égale au double de la prévalence observée.

3.3.4 Individus non-testés

Une fois l'enquête ménages réalisée, un prélèvement sanguin ou salivaire est proposé, après consentement, aux adultes des ménages retenus pour le test du dépistage du VIH. Cependant, tous les adultes éligibles ne sont pas enquêtés pour le questionnaire individuel et/ou le dépistage du VIH, pour différentes raisons : absence lors des différentes visites, refus de participer, etc. Certains individus peuvent avoir été enquêtés pour le questionnaire individuel et avoir été absents ou avoir refusé le dépistage du VIH. Pour d'autres c'est l'inverse. Nous pouvons distinguer quatre groupes parmi l'ensemble des individus éligibles pour le dépistage du VIH :

- personnes interviewées pour le questionnaire individuel (questionnaire hommes ou questionnaire femmes selon le cas) et testées pour le VIH ;
- personnes interviewées mais non testées ;
- personnes testées mais non interviewées ;
- personnes ni testées ni interviewées.

Tableau 3.12

Proportion d'individus âgés de 15 à 49 ans éligibles pour le dépistage du VIH qui ont été testés et/ou interrogés pour le questionnaire individuel de neuf EDS ou AIS

Pays	Année	Sexe	Ni test ni quest.	Quest. sans test	Test sans quest.	Quest. et test	Individus éligibles
Burkina Faso	2003	H	6,9 %	6,4 %	2,1 %	84,6 %	3 501
		F	2,6 %	5,3 %	0,7 %	91,4 %	4 607
Cameroun	2004	H	6,4 %	4,7 %	1,0 %	87,9 %	5 146
		F	3,8 %	4,6 %	1,7 %	89,8 %	5 759
Éthiopie	2005	H	8,6 %	9,1 %	0,1 %	82,2 %	6 139
		F	3,4 %	9,2 %	0,2 %	87,3 %	6 963
Ghana	2003	H	6,2 %	14,3 %	0,1 %	79,4 %	4 636
		F	4,2 %	6,5 %	0,2 %	89,1 %	5 845
Kenya	2003	H	12,9 %	15,0 %	0,5 %	71,6 %	3 970
		F	5,0 %	16,3 %	0,3 %	78,5 %	4 293
Lesotho	2004	H	16,0 %	18,2 %	0,4 %	65,4 %	2 926
		F	6,0 %	15,0 %	0,3 %	78,7 %	3 672
Malawi	2004	H	13,7 %	25,6 %	0,0 %	60,7 %	3 663
		F	5,6 %	27,3 %	0,0 %	67,1 %	4 057
Sénégal	2005	H	12,6 %	12,0 %	1,0 %	74,5 %	3 997
		F	5,6 %	10,9 %	0,9 %	82,6 %	5 342
Tanzanie	2003	H	9,0 %	15,7 %	0,0 %	75,3 %	6 282
		F	4,2 %	13,6 %	0,0 %	82,2 %	7 231

Sources : EDS et AIS, nos calculs.

Quest. : personne ayant été enquêtée pour le questionnaire individuel, hommes ou femmes.

Test : personne dépistée pour le VIH.

Pourcentages en ligne.

La répartition des adultes éligibles au test de dépistage du VIH selon ces quatre catégories est indiquée dans le Tableau 3.12. Ce tableau a été calculé pour neuf EDS ou AIS pour lesquelles les données des questionnaires et du dépistage du VIH étaient disponibles en téléchargement sur www.measuredhs.com et pour lesquelles il était possible de lier les résultats du test VIH aux données d'enquêtes.

Pour chaque pays, nous disposions de quatre bases de données : membres du ménage, questionnaires hommes, questionnaire femmes et résultat VIH. Les 35 fichiers dont nous disposions[24] ont été fusionnés en une base de données unique. Ce fichier a ensuite été restreint en ne gardant que les individus éligibles pour le dépistage du VIH et âgés de 15 à 49 ans. Nous disposions de trois variables de pondération pour chaque pays : celle du fichier ménages, celle des questionnaires individuels et celle du fichier des résultats VIH. Pour toutes les analyses réalisées dans cette section, nous avons eu recours à la variable de pondération du fichier ménages, seule variable renseignée pour l'ensemble des individus.

Le Tableau 3.12 montre que les taux de personnes non testées sont relativement importants : de 8 à 40 % selon les pays et le sexe. Ces taux élevés de non réponse ont été largement invoqués comme principale source de biais des EDS pouvant expliquer en partie les écarts observés entre EDS et surveillance sentinelle des femmes enceintes (UNAIDS/WHO WORKING GROUP ON GLOBAL HIV/AIDS AND STI SURVEILLANCE 2000, BOERMA 2003, THE UNAIDS REFERENCE GROUP ON ESTIMATES MODELLING AND PROJECTIONS 2004a, UNAIDS 2004b, UNAIDS/WHO WORKING GROUP ON GLOBAL HIV/AIDS AND STI SURVEILLANCE 2005). Plusieurs ont montré en effet, dès la fin des années 1980, que les personnes refusant le test présentaient une prévalence supérieure aux autres (HULL 1988, JENUM 1988)[25].

En 2006, deux études ont essayé d'estimer l'ampleur de ce biais. Le premier article, supposant que les personnes non testées présentaient une prévalence double de celle des personnes observées, calculait des prévalences ajustées 1,03 à 1,34 fois supérieures aux prévalences observées (GARCIA-CALLEJA 2006). Dans le second article, la prévalence du VIH des hommes et des femmes non testés de cinq enquêtes a été estimée à partir de modèles statistiques multivariés, des régressions logistiques en l'occurrence, calculés sur les individus enquêtés. Les auteurs concluaient que, bien que la prévalence du VIH tendait à être supérieure parmi les individus non testés que les individus testés, l'effet global sur les estimations nationales de la prévalence du VIH étaient insignifiants (MISHRA 2006).

[24] Pour l'AIS réalisée en Tanzanie en 2003, le questionnaire était commun aux hommes et aux femmes.

[25] Ces différents points ont déjà été évoqués tout au long du chapitre 1. il s'agit d'une question ancienne et récurrente.

Encadré 3.2

Principe de la régression logistique binaire

La **régression logistique binaire** est une technique statistique qui a pour objectif, à partir d'un fichier d'observations, de produire un modèle permettant de prédire les valeurs prises par une variable catégorielle binaire, de type oui/non, à partir d'une série de variables explicatives continues et/ou catégorielles. Nous ne parlerons ici que du modèle plus courant, à savoir le modèle *logit*.

Nous noterons *X* la variable étudié. Elle prend deux valeurs : 0 si l'évènement ne s'est pas réalisé, 1 s'il s'est réalisé. On modélisera la probabilité que *X* soit égal à 1 connaissant les valeurs d'autres variables, dites variables explicatives et notées Y_1 Y_2 Y_3 … Y_n.[26] Ces différentes variables sont supposées être indépendantes les unes des autres.

On cherchera les coefficients α, β_1, β_2, β_3 … β_n tels que

$$\text{logit}(P[X=1]) = \alpha + \beta_1 y_1 + \beta_2 y_2 + \ldots + \beta_n y_n \text{ où } \text{logit}(p) = \log\left(\frac{p}{1-p}\right).$$

L'intérêt de cette approche est de permettre d'inclure des variables explicatives catégorielles. Une variable ayant quatre catégories sera transformée en trois variables binaires. Une des modalités sera considérée comme modalité de référence. Pour chacune des trois autres modalités, une variable binaire sera crée, valant 1 pour les individus appartenant à cette modalité et 0 sinon.

Pour estimer les différents paramètres du modèle, la probabilité que X soit égale à 1 pour une configuration donnée des variables entrées dans le modèle à la fréquence observée de X parmi les individus répondant à cette configuration.

La force du modèle *logit* est de fournir l'effet de chaque variable, *toutes choses étant égales par ailleurs*, c'est-à-dire indépendamment des autres variables entrées dans le modèle. Autrement dit, l'interprétation d'un modèle est toujours fonction de l'ensemble des variables entrées dans celui-ci. Une variable peut avoir un effet significatif dans un modèle donné et ne plus avoir d'effet lorsque d'autres variables sont ajoutées au modèle.

L'exponentiel des coefficients β correspond à l'*odds ratio* de la variable X par rapport à la variable associée au coefficient considéré. Si l'odds ratio est supérieur à 1, cela signifie que cette modalité augmente la probabilité d'observer X par rapport à la modalité de référence, et inversement (TOULEMON 1995). Bien que ce soit souvent le cas, il importe de ne pas confondre odds ration et risque relatif, notamment lorsque les prévalences ne sont pas petites (TAFFÉ 2004).

Lorsque la régression logistique est utilisée pour prédire le statut d'individus selon X, le modèle est appliqué aux individus pour calculer leur probabilité d'être positifs. Si elle est supérieure ou égale à 0,5, alors ils seront considérés comme positifs, négatifs sinon.

[26] La notation est ici simplifiée par rapport aux notations habituelles. Pour une notation plus rigoureuse, voir les cours de Patrick TAFFÉ : TAFFÉ P., *Cours de Régression Logistique Appliquée*, Lausanne (CH), Institut Universitaire de Médecine Sociale et Préventive et Centre d'Épidémiologie Clinique, 2004.

Nous allons reprendre ici le principe de l'analyse effectuée par Vinod MISHRA et ses collaborateurs. Pour les individus ni testés ni interviewés, nous disposons néanmoins des informations contenues dans le questionnaire ménages. Pour les personnes interviewées mais non testées, nous disposons par ailleurs des informations contenues dans le questionnaire individuel. Nous allons alors calculer des modèles statistiques à partir des individus testés pour prédire ensuite la prévalence des individus non testés. Tous nos modèles seront calculés séparément pour chaque pays et chaque sexe. Pour chacun d'eux, deux régressions logistiques seront calculées. La première ne portera que sur des variables renseignées dans le questionnaire ménages et sera calculée à partir de l'ensemble des individus testés. Les variables inclues dans le modèle sont :

- taille du ménage ;
- milieu de résidence ;
- niveau d'instruction ;
- posséder un poste de radio ;
- posséder un poste de télévision ;
- quintile de bien-être[27] ;
- groupes d'âges quinquennaux ;
- région de résidence.

Une seconde régression logistique sera calculée uniquement à partir des individus testés ayant répondu au questionnaire individuel. En plus des variables précédentes, les variables suivantes ont été ajoutées :

- statut matrimonial ;
- avoir un emploi ;
- IST au cours des douze derniers mois ;
- ulcère génital au cours des douze derniers mois ;
- écoulement génital au cours des douze derniers mois ;
- utilisation du préservatif au dernier rapport sexuel (sauf pour les hommes du Lesotho) ;
- désir de se protéger vis-à-vis du VIH/SIDA ;
- nombre de rapports sexuels au cours du dernier mois ;
- âge au premier rapport sexuel ;
- nombre de partenaires sexuels au cours des douze derniers mois ;

[27] Il s'agit d'un indicateur composite calculé par Measure DHS, pour chaque pays, à partir des caractéristiques des ménages (possession de biens, type d'habitat, de sanitaires, etc.).

Tableau 3.13

Prévalence observée, prévalence prédite pour les non testés et prévalence ajustée, par sexe, pour neuf pays (15-49 ans).

Enquête	Prévalence observée parmi testés (IC 95 %)	Prévalence prédites parmi non testés (IC 95 %)	Ratio observée sur prédite	Prévalence ajustée testés et non testés (IC 95 %)	Ratio ajustée sur observée	Taux de non testés
Burkina Faso 2003						
Hommes	1,8 (1,3-2,2)	2,1 (1,9-2,4)	1,196	1,8 (1,4-2,2)	1,026	13,4
Femmes	1,8 (1,4-2,2)	3,1 (2,5-3,7)	1,760	1,9 (1,5-2,2)	1,060	7,9
Ensemble	1,8 (1,5-2,1)	2,6 (2,3-2,8)	1,443	1,8 (1,6-2,1)	1,045	10,3
Cameroun 2004						
Hommes	4,1 (3,5-4,6)	5,7 (5,2-6,2)	1,406*	4,2 (3,7-4,7)	1,045	11,1
Femmes	6,6 (6,0-7,3)	8,4 (7,7-9,2)	1,272	6,8 (6,2-7,4)	1,023	8,5
Ensemble	5,4 (5,0-5,9)	7,0 (6,5-7,4)	1,281*	5,6 (5,2-6,0)	1,027	9,7
Éthiopie 2005						
Hommes	0,9 (0,6-1,2)	1,2 (1,0-1,4)	1,336	1,0 (0,7-1,2)	1,059	17,7
Femmes	1,7 (1,4-2,0)	3,2 (2,7-3,7)	1,864*	1,9 (1,6-2,2)	1,109	12,6
Ensemble	1,3 (1,1-1,5)	2,1 (1,8-2,3)	1,556*	1,4 (1,3-1,6)	1,083	15,0
Ghana 2003						
Hommes	1,4 (1,0-1,8)	1,9 (1,6-2,1)	1,320	1,5 (1,2-1,8)	1,066	20,5
Femmes	2,7 (2,3-3,1)	2,6 (2,3-2,8)	0,949	2,7 (2,3-3,1)	0,995	10,7
Ensemble	2,2 (1,9-2,5)	2,2 (2,0-2,3)	0,990	2,2 (1,9-2,4)	0,999	15,0
Kenya 2003						
Hommes	4,7 (3,9-5,5)	5,0 (4,6-5,5)	1,074	4,8 (4,2-5,4)	1,021	27,9
Femmes	8,7 (7,8-9,7)	7,5 (7,0-8,0)	0,857	8,5 (7,7-9,2)	0,970	21,3
Ensemble	6,9 (6,3-7,5)	6,2 (5,8-6,5)	0,894	6,7 (6,2-7,2)	0,974	24,4
Lesotho 2004						
Hommes	19,0 (17,3-20,8)	19,2 (18,2-20,2)	1,009	19,1 (17,9-20,3)	1,003	34,2
Femmes	26,0 (24,4-27,5)	25,3 (24,2-26,5)	0,976	25,8 (24,5-27,1)	0,995	21,0
Ensemble	23,2 (22,0-24,4)	21,9 (21,1-22,6)	0,943	22,8 (21,9-23,7)	0,985	26,9
Malawi 2004						
Hommes	10,1 (8,8-11,3)	10,5 (9,8-11,2)	1,044	10,2 (9,4-11,0)	1,017	39,3
Femmes	13,9 (12,6-15,2)	12,9 (12,3-13,5)	0,929	13,6 (12,7-14,5)	0,977	32,9
Ensemble	12,2 (11,3-13,1)	11,7 (11,2-12,1)	0,958	12,0 (11,4-12,6)	0,985	35,9
Sénégal 2005						
Hommes	0,5 (0,2-0,7)	0,5 (0,4-0,7)	1,133	0,5 (0,3-0,7)	1,033	24,6
Femmes	0,9 (0,6-1,2)	0,8 (0,7-0,9)	0,868	0,9 (0,7-1,1)	0,978	16,4
Ensemble	0,7 (0,5-0,9)	0,6 (0,6-0,7)	0,889	0,7 (0,6-0,9)	0,978	19,9
Tanzanie 2003						
Hommes	6,0 (5,3-6,7)	7,1 (6,7-7,5)	1,181	6,3 (5,8-6,8)	1,045	24,7
Femmes	7,5 (6,9-8,2)	8,4 (7,9-8,9)	1,119	7,7 (7,1-8,2)	1,021	17,8
Ensemble	6,9 (6,4-7,3)	7,7 (7,4-8,0)	1,123	7,0 (6,6-7,4)	1,026	21,0

IC 95 % : intervalle de confiance à 95 %.

Tous les résultats, à l'exception des deux ratios, sont exprimés en pourcents.

* La prévalence du VIH prédite parmi les non testés diffère significativement, à 5 %, de la prévalence observée parmi les personnes testées (test t).

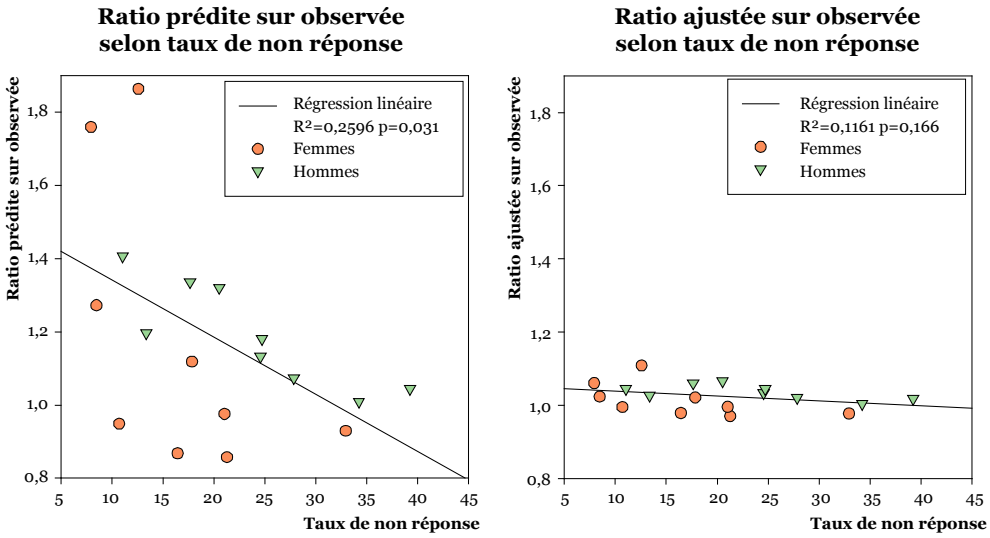
Sources : nos calculs.

- fumer (sauf pour les hommes du Burkina Faso, du Cameroun, du Sénégal, de Tanzanie et les femmes de Tanzanie) ;
- circoncision pour les hommes ;
- excision pour les femmes (sauf celles du Lesotho et du Malawi).

Les variables retenues ont été sélectionnées en fonction de leur disponibilité dans les différentes bases de données et afin de couvrir différents domaines associés plus ou moins directement à l'infection à VIH : comportements sexuels, IST, pauvreté, caractéristiques sociodémographiques… Elles ne sont pas toutes des déterminants directs de la prévalence du VIH. Cependant, elles permettent de couvrir indirectement les déterminants non mesurés dans les EDS. Ainsi, les variables régions et milieu de résidence, fortement discriminantes, reflètent en partie l'historique migratoire des différentes régions ou les variations locales de la prévalence de certaines IST comme l'infection à HSV-2, très fortement corrélée à l'infection à VIH. Pour chaque variable, les non réponses ont été considérées comme une modalité afin de ne pas exclure de l'analyse les individus ayant répondu partiellement aux questionnaires.

Figure 3.14

Corrélations entre les taux de non réponse et les ratios prévalence prédite sur prévalence observée et prévalence ajustée sur prévalence observée



Une fois les deux modèles calculés, la probabilité qu'un individu non testé soit séropositifs a été calculée en utilisant le premier modèle pour les personnes ni testées ni interviewées et le second modèle pour les individus non testés mais interviewés. La prévalence des non testés est ensuite calculée comme étant la moyenne des probabilités individuelles d'être séropositifs. Ce type d'approche permet de procéder à un ajustement simultané sur l'ensemble des variables inclues dans les modèles. Le Tableau 3.13 présente les prévalences observées parmi les

personnes testées, les prévalences prédites parmi les non testés et les prévalences ajustées.

Le ratio prévalence prédite parmi les non testés sur prévalence observée varie de 0,857 à 1,864. Pour 5 enquêtes sur les 9, ce ratio est inférieur à 1 pour les femmes. Pour les hommes par contre, la prévalence estimée des non testés est systématiquement supérieure à celle observée parmi les personnes enquêtées. La prévalence prédite des non testés n'est cependant significativement différente à 5 % de la prévalence observée que pour les hommes du Cameroun et les femmes d'Éthiopie. La Figure 3.14 montre une corrélation négative, statistiquement significative, entre le ratio prévalence prédite sur prévalence estimée et les taux de non réponse. Cette corrélation est plus marquée pour les hommes que pour les femmes. Plus les taux de non réponses sont faibles, plus les non testés présentent une prévalence plus élevée que celles des personnes enquêtées. À l'inverse, plus les taux de non réponse sont élevés, moins nous observons un effet de sélection.

La prévalence ajustée dépend de deux paramètres : la proportion de non testés et la prévalence relative de ce même groupe. Comme cette dernière est corrélée négativement avec le taux de non réponse, ces deux effets se compensent mutuellement et le ratio prévalence ajustée sur prévalence observée demeure ainsi très proche de 1 (entre 0,970 et 1,109). Il n'y pas de corrélation significative entre ce ratio et le taux de non réponse (Figure 3.14) et les prévalences ajustées ne diffèrent pas significativement des prévalences observées. Les différences entre ces deux valeurs n'excèdent 0,32 points et la prévalence ajustée est systématiquement largement incluse au sein de l'intervalle de confiance à 95 % de la prévalence observée.

3.3.5 Ajustement final

Nous venons d'aborder et d'estimer quatre sources de biais des enquêtes nationales en population générale : les populations hors ménages, l'ancienneté de la base de sondage, les ménages non enquêtés et les individus non testés. À cela, s'ajoute le biais lié aux tests de dépistage, à savoir la proportion d'individus non observables car situés dans la fenêtre sérologique (voir section 3.1).

Si l'impact de chacun de ces biais est relativement mineur sur l'estimation nationale de la prévalence du VIH, la question reste de savoir si, en se cumulant, l'impact devient beaucoup plus important quand ces différents biais se cumulent.

Le Tableau 3.14 fournit les prévalences par sexe et région du Burkina Faso, du Cameroun et du Kenya après ajustement sur les personnes non testées (section 3.3.4). Il permet de procéder alors à l'ajustement selon la structure par sexe et région (section 3.3.2). Nous pouvons ensuite tenir compte du taux de non réponse des ménages (section 3.3.3) puis de la proportion d'individus vivant hors ménage (section 3.3.1). Enfin, nous pouvons corriger la prévalence obtenue en fonction de la fenêtre sérologique en considérant une sensibilité globale de 99 % et une spécificité de 100 % (section 3.1.5). En modifiant l'Équation 3.1, nous pouvons calculer la prévalence réelle à partir de la prévalence observée.

Équation 3.7
Prévalence réelle selon la prévalence observée, la spécificité et la sensibilité du test

$$p_r = \frac{p_o + sp - 1}{sb + sp - 1}$$

p_o : *prévalence observée*
p_r : *prévalence réelle*
sb : *sensibilité*
sp : *spécificité*

L'ensemble de ces ajustements successifs sont retranscrits dans le Tableau 3.15. Les ajustements cumulés, hypothèses haute et basse, induisent des prévalences proches de la prévalence nationale observée initialement. Les prévalences ajustées se situent toutes dans les intervalles de confiance à 95 % de la prévalence observée. Nous avons également indiqué les intervalles de confiance à 75 %. En acceptant d'augmenter les risques de se tromper, nous obtenons des intervalles plus précis. Là encore, les prévalences ajustées, à l'exception de l'hypothèse haute au Cameroun, se situent à l'intérieur des intervalles de confiance.

Tableau 3.14

Prévalence du VIH, 15-49 ans, par région, avec ajustement sur les individus non testés, pour le Burkina Faso et le Kenya en 2003 et le Cameroun en 2004

Région	Effectif hommes	Effectif femmes	Prévalence hommes	Prévalence femmes
Burkina Faso 2003				
Boucle du Mouhoun	252	328	2,43 %	2,17 %
Centre	552	638	3,15 %	4,39 %
Centre-Sud	203	253	0,41 %	0,73 %
Plateau Central	164	238	0,92 %	1,36 %
Centre-Est	255	393	0,91 %	1,52 %
Centre-Nord	242	377	1,13 %	0,63 %
Centre-Ouest	227	338	3,01 %	1,81 %
Est	256	320	3,14 %	0,55 %
Nord	285	431	1,24 %	0,89 %
Cascades	111	144	1,79 %	2,72 %
Hauts-Bassins	538	623	1,02 %	2,15 %
Sahel	259	310	0,00 %	0,25 %
Sud-Ouest	158	213	4,52 %	3,64 %
Ensemble	*3 501*	*4 605*	*1,83 %*	*1,87 %*
TOTAL	*8 106*		*1,85 %*	
Cameroun 2004				
Adamoua	457		6,92 %	
Centre	2 059		6,85 %	
Est	548		8,47 %	
Extrême-Nord	1 584		2,10 %	
Littoral	1 758		4,96 %	
Nord	961		1,72 %	
Nord-Ouest	1 142		9,00 %	
Ouest	1 051		4,76 %	
Sud	487		6,43 %	
Sud-Ouest	858		8,03 %	
TOTAL	*10 905*		*5,59 %*	
Kenya 2003				
Nairobi	453	444	7,44 %	11,14 %
Central	609	642	2,09 %	7,33 %
Coast	285	340	4,68 %	6,88 %
Eastern	636	691	1,94 %	5,70 %
Nyanza	536	609	12,44 %	17,82 %
Rift Valley	949	999	3,54 %	6,81 %
Western	433	490	4,12 %	5,71 %
North Eastern	68	78	0,00 %	0,00 %
Ensemble	*3 968*	*4 293*	*4,79 %*	*8,47 %*
TOTAL	*8 262*		*6,70 %*	

* Ouagadougou a été réintégrée à la région Centre.

† Yaoundé a réintégrée à la région Centre.

‡ Douala a été réintégrée à la région Littoral.

Sources : nos calculs à partir des bases de données des EDS disponibles sur www.measuredhs.com.

Tableau 3.15

Ajustement de la prévalence du VIH, 15-49 ans, des EDS 2003 et 2004 du Burkina Faso, du Cameroun et du Kenya

	Burkina Faso 2003	Cameroun 2004	Kenya 2003
Prévalence nationale observée dans l'EDS (%)	1,77	5,44	6,88
Intervalle de confiance à 75 %	1,59 – 1,96	5,18 – 5,71	6,51 – 7,27
Intervalle de confiance à 95 %	1,49 – 2,11	5,00 – 5,91	6,27 – 7,54
Prévalence ajustée sur les non-testés (1)	1,85	5,59	6,70
Prévalence ajustée sur la structure par région et sexe* (2)	1,79	5,37	6,70
Ajustement sur les non testés et la structure par région (1 & 2)	1,82	5,51	6,71
Taux de non réponse des ménages (%)	0,7	3,0	3,7
Ajustement sur non réponse des ménages hypothèse haute† (3)	1,81	5,67	6,95
Ajustement sur non réponse des ménages hypothèse basse† (3')	1,79	5,42	6,58
Ajustement 1, 2 & 3 hypothèse haute	1,83	5,68	6,96
Ajustement 1, 2 & 3' hypothèse basse	1,81	5,43	6,59
Proportion d'adultes hors ménage, camps exclus (%)	0,43	1,81	2,34
Proportion d'adultes en camps de réfugiés (%)	0	0	0,71
Prévalence observée en camps de réfugiés (%)	-	-	2,35
Ajustement sur population hors ménage hypothèse haute† (4)	1,78	5,54	7,01
Ajustement sur population hors ménage hypothèse basse† (4')	1,77	5,39	6,77
Ajustement 1, 2, 3 et 4 hypothèse haute	1,84	5,78	7,09
Ajustement 1, 2, 3' et 4' hypothèse basse	1,81	5,38	6,48
Ajustement sur fenêtre sérologique‡ (5)	1,79	5,49	6,95
Ajustement 1, 2, 3, 4 et 5 hypothèse haute	1,86	5,84	7,16
Ajustement 1, 2, 3', 4' et 5 hypothèse basse	1,82	5,43	6,55
Ajustement 1, 2, 3, 4 et 5 hypothèse haute alternative §	1,85	5,70	6,95

* Pour le Cameroun, nous ne disposions que d'une structure par région. Seul un ajustement national a donc été calculé.

† Hypothèse haute : prévalence relative de 2. – Hypothèse basse : prévalence relative de ½.

‡ Sous l'hypothèse d'une sensibilité de 99 % et d'une spécificité de 100 % (voir section 3.1.5).

§ Hypothèse haute alternative : prévalence relative de 1,5.

Nous avons évoqué plusieurs fois précédemment que nos différentes hypothèses maximisaient les biais : l'ensemble des individus hors ménages sont considérés comme ayant 15 à 49 ans, utilisation de prévalences relatives de ½ et 2, alors qu'il est plus vraisemblable, en extrapolant les résultats sur les non testés, qu'elles soient de l'ordre de 0,75 à 1,5. Nous avons retranscrit dans le Tableau 3.15 les résultats obtenus en posant une hypothèse alternative correspondant à une prévalence relative de 1,5 pour l'ajustement du taux de non réponse des ménages et de la proportion d'adultes hors ménage. Sous cette hypothèse, la prévalence ajustée du Cameroun serait alors de 5,70 %, comprise dans l'intervalle de confiance à 75 % de la prévalence observée.

*
**

Ces résultats mettent en évidence que, pour les EDS des trois pays analysés, bien que les sources de biais soient nombreuses, leur impact sur l'estimation nationale de la prévalence du VIH est inférieur à l'imprécision inhérente à l'échantillonnage de ces enquêtes. Les EDS permettent donc de fournir, au niveau national, une estimation précise du niveau de la prévalence du VIH, tant que cette estimation nationale est interprétée en tenant compte de l'intervalle de confiance à 95 % dans laquelle elle s'inscrit. Nous pouvons donc affirmer avec un risque réduit que la prévalence du VIH des 15-49 ans au Burkina Faso est située entre 1,5 et 2,1 %.

3.4 Représentativité de la surveillance sentinelle des femmes enceintes

Les enquêtes sentinelles effectuées auprès des femmes enceintes consultant en cliniques prénatales sont depuis longtemps utilisées par l'OMS et l'ONUSIDA pour produire des estimations de la prévalence nationale au sein de l'ensemble de la population adulte (15-49 ans). L'hypothèse sous-jacente repose sur le fait que ces femmes seraient représentatives de l'ensemble de la population adulte, hommes et femmes (UNAIDS/WHO 2003). D'autres auteurs formulent cette hypothèse légèrement différemment en considérant que les femmes enceintes sont représentatives de l'ensemble des femmes (BOISSON 1996). Ces conclusions sont basées le plus souvent sur une comparaison, à un niveau local, entre les prévalences mesurées au sein d'une clinique prénatale et celles mesurées en population générale sur les quelques sites où des enquêtes de ce type ont été menées pendant les années 1990-2000, à un niveau local.

Cependant, la surveillance sentinelle des femmes enceintes n'a pas été effectuée sur l'ensemble des cliniques prénatales du pays mais sur une sélection de ces dernières, dont le nombre peut varier grandement d'un pays à l'autre, de l'ordre d'une demi-douzaine jusqu'à une cinquantaine. La sélection et la localisation de ces cliniques peut donc constituer une source importante de biais (JACKSON 1999, SCHWARTLANDER 1999).

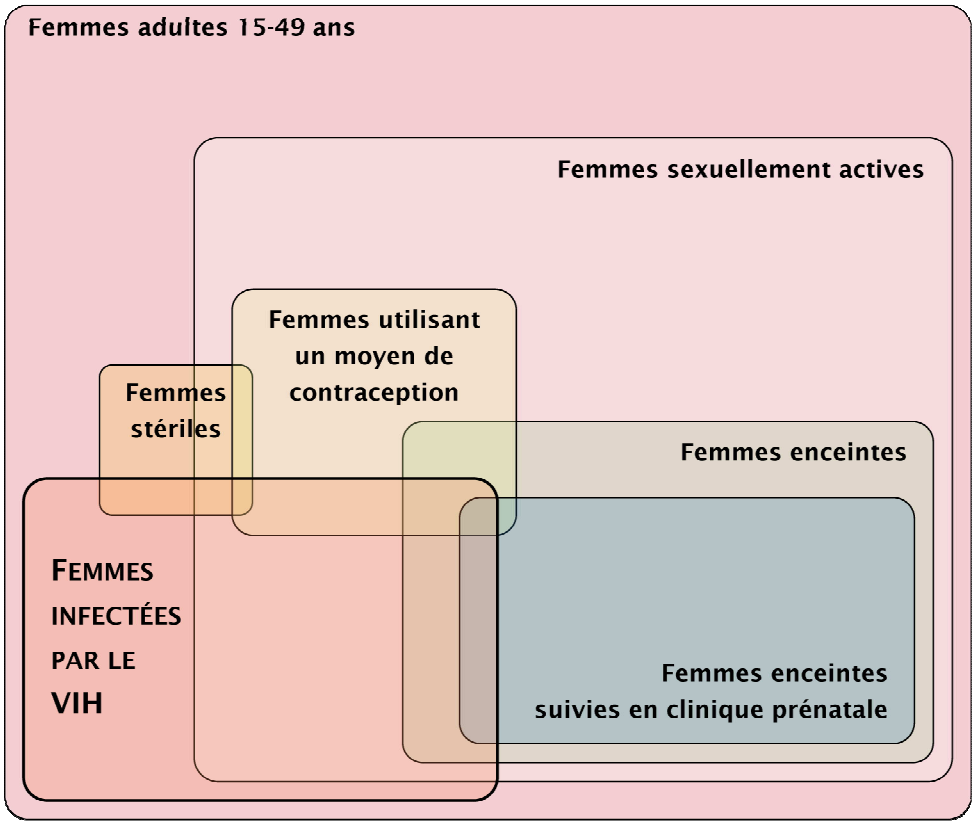
La question de la représentativité de la surveillance sentinelle des femmes enceintes se situent donc à différents niveaux. Dans un premier temps, nous regarderons dans quelle mesure, à un niveau local, les femmes enceintes en cliniques prénatales peuvent être considérées comme représentatives de l'ensemble des femmes adultes (section 3.4.1). Puis, nous nous demanderons si elles sont représentatives de la population générale adulte, hommes et femmes confondus (section 3.4.2). Enfin, nous nous intéresserons à la question de la sélection et de la localisation des cliniques prénatales retenues pour la surveillance sentinelle. Les femmes enceintes enquêtées sont-elles représentatives de l'ensemble des femmes enceintes du pays (section 3.4.3) ?

3.4.1 Femmes enceintes et ensemble des femmes

Les femmes enceintes consultant en clinique prénatale constitue un sous-ensemble des femmes enceintes, elles-mêmes un sous-ensemble des femmes sexuellement actives, elles-mêmes un sous ensemble des femmes adultes (Figure 3.15). Les

femmes qui consultent en clinique prénatale présentent un profil différent de celui de l'ensemble des femmes. Dans une étude réalisée sur des données provenant du Cameroun, du Kenya et de Zambie, il a été montré que celles consultant en clinique prénatale étaient plus jeunes, plus éduquées, plus souvent mariées et avaient plus souvent recours à la contraception (GLYNN 2001a). Or, dans la même étude, il apparaissait sur les trois sites analysés que la prévalence du VIH variait selon l'âge, le statut matrimonial, l'occupation professionnelle, l'utilisation d'une méthode contraceptive et le fait d'avoir déjà eu un enfant. D'autres études ont mis en évidence des variations selon le niveau d'éducation (FYLKESNES 1998, BLANC 2000, FYLKESNES 2001), le statut matrimonial (GREGSON 1995, GREGSON 2002b), la religion (GREGSON 1995) ou encore l'histoire migratoire (GREGSON 2002b).

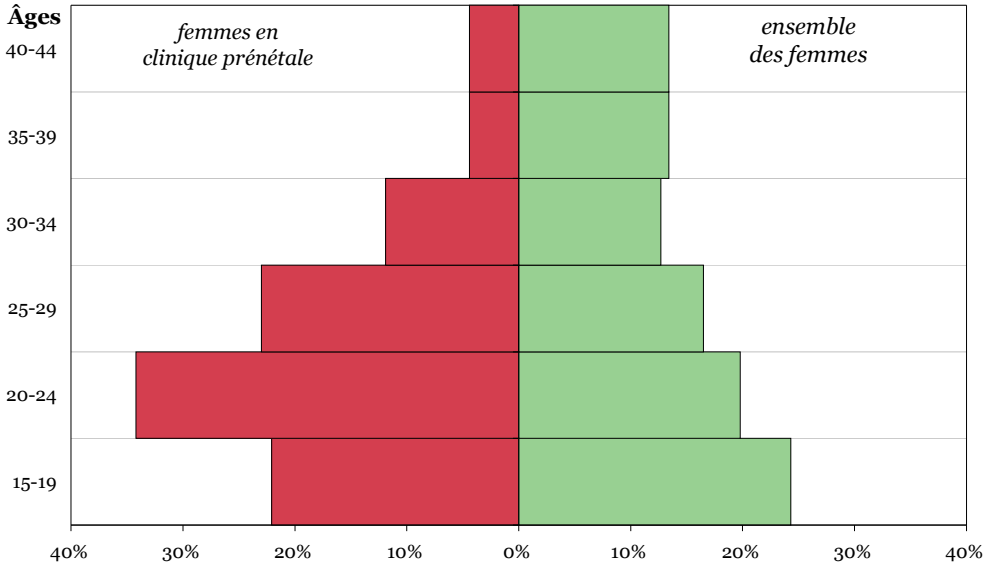
Figure 3.15
Les femmes enceintes suivies en clinique prénatale : un sous-ensemble de l'ensemble des femmes adultes



L'âge présente des différentiels de prévalence très marqués. Or, la structure par âges des femmes enceintes suivies en cliniques prénatales diffèrent singulièrement de celle de l'ensemble des femmes (Figure 3.16), du fait, entre autres, de la variation des taux de fécondité selon l'âge.

Figure 3.16

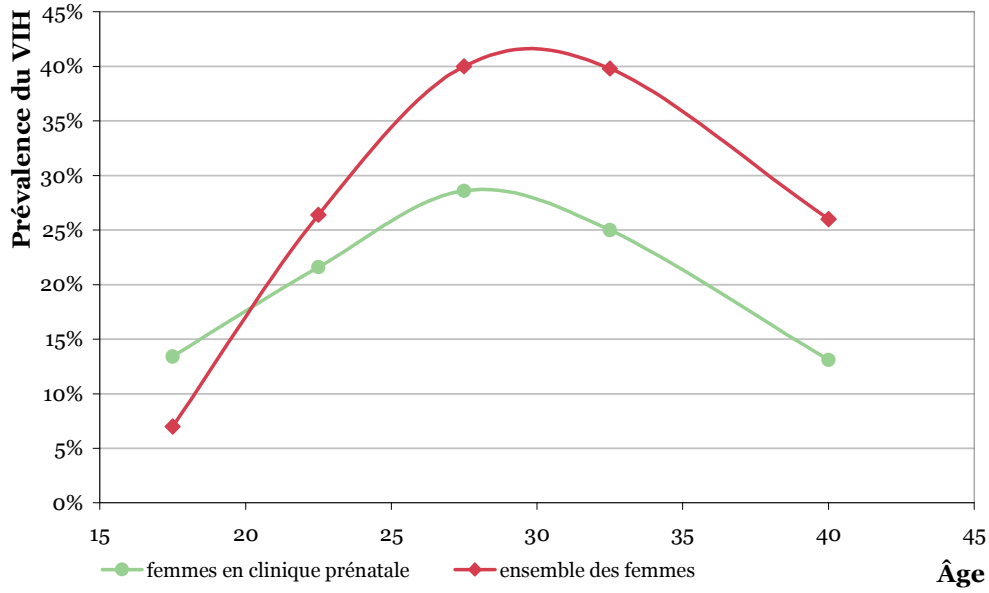
Structure par âges des femmes consultant en clinique prénatale et de l'ensemble des femmes à Manicaland, Zimbabwe (1998-2000)



Source : (GREGSON 2002b).

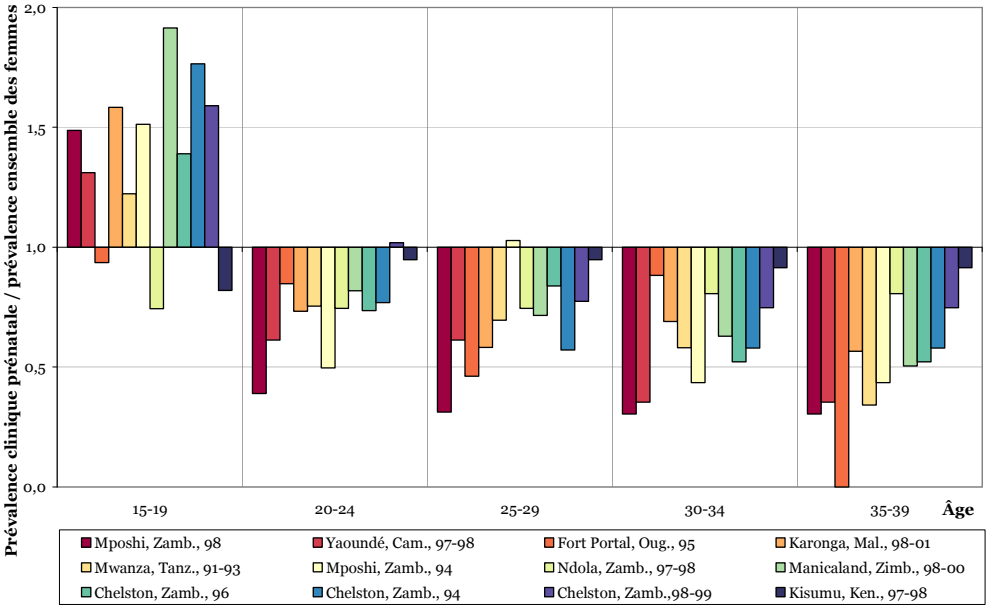
Figure 3.17

Prévalence du VIH selon l'âge, pour les femmes suivies en clinique prénatale et l'ensemble des femmes, à Manicaland, Zimbabwe (1998-2000)



Source : (GREGSON 2002b).

Figure 3.18
Ratio prévalence du VIH mesurée en clinique prénatale sur prévalence de l'ensemble des femmes, selon l'âge, pour 12 séries de données d'Afrique subsaharienne entre 1995 et 2001



Sources : (KILIAN 1999, FYLKESNES 2001, GLYNN 2001a, CHANGALUCHA 2002, GREGSON 2002b, CRAMPIN 2003).

Les prévalences observées en clinique prénatale sous-estiment la prévalence de l'ensemble des femmes à tous les âges, excepté aux jeunes âges (Figure 3.17). Ce phénomène est aujourd'hui bien avéré et a été observé dans plusieurs pays (Figure 3.18) dont la Tanzanie (KWESIGABO 2000, CHANGALUCHA 2002), le Malawi (CRAMPIN 2003), la Zambie (FYLKESNES 1998), le Zimbabwe (GREGSON 1995, GREGSON 2002b) et l'Ouganda (CARPENTER 1997, GRAY 1998).

Cette sous-estimation s'explique en partie par un différentiel de fécondité chez les femmes séropositives (ZABA 1998). Ces dernières sont moins fertiles (WIDY-WIRSKI 1988) et ont significativement plus d'avortements spontanés (BROCKLEHURST 1998). Elles ont, par ailleurs, plus fréquemment une stérilité préexistante des suites d'autres IST (ROSS 1999). Des facteurs comportementaux existent également mais leurs effets sont *a priori* plus faibles. Ainsi, il a été montré que, dans la région de Rakai en Ouganda, les femmes séropositives ont des rapports sexuels moins fréquents (GRAY 1998). À Kinshasa, Zaïre, elles utiliseraient plus fréquemment un moyen contraceptif (RYDER 1991). Une femme séropositive a plus souvent un partenaire séropositif, d'où moins de rapports sexuels du fait de la maladie du partenaire, et un risque de veuvage plus élevé (ZABA 1998). D'autre part, le VIH entraîne une baisse de la production de spermatozoïdes chez les hommes (KRIEGER 1991, MARTIN 1992). Bien que le statut sérologique soit rarement connu, une suspicion de séropositivité pourrait parfois entraîner une rupture du couple

(NDINYA-ACHOLA 1990) et défavoriser le remariage des veuves et des divorcées (NTOZI 1997). La sous-fécondité des femmes séropositives augmente avec l'âge, en particulier parce que la baisse de fécondité des femmes VIH+ s'accentue avec la durée d'infection.

Seules les plus jeunes (15-19 ans) ont une fécondité plus élevée que les femmes séronégatives du même âge (CARPENTER 1997, GRAY 1998, CARPENTER 2002). Ceci s'explique par un effet de sélection : les femmes de moins de vingt ans qui sont enceintes ont commencé leur vie sexuelle plus tôt et ont donc été plus soumises au risque d'être contaminées par le VIH.

Sur les douze séries de données de la Figure 3.18, la prévalence brute, c'est-à-dire sans ajustement, observée en clinique prénatale s'avère être inférieure de 25 % en moyenne à la prévalence de l'ensemble des femmes, avec des variations d'un site à l'autre de -7 à -50 % (Tableau 3.17).

Plusieurs méthodes d'ajustement des données issues de la surveillance sentinelle des femmes enceintes ont été proposées à la fin des années 1990 pour estimer la prévalence de l'ensemble des femmes. La méthode la plus simple consiste à ajuster les prévalences par groupes d'âges en fonction de la structure par âge de la population féminine générale. Cette méthode a pour effet de sous-estimer encore plus la prévalence de l'ensemble des femmes[28], par rapport aux résultats bruts non ajustés, dans la mesure la structure par âge des femmes enceintes vient compenser légèrement la sous-estimation globale de la surveillance sentinelle (LARMARANGE 2004).

Les autres méthodes proposées tiennent compte, quant à elles, de la fécondité différentielle des femmes VIH+ par rapport aux femmes VIH-. La première méthode consiste à corriger la prévalence de chaque groupe d'âges en fonction du ratio fécondité des femmes VIH+ sur fécondité des femmes VIH- (BOISSON 1996, KARON 1997, LEE 1998, NICOLL 1998). La prévalence en population féminine générale s'obtient donc selon l'Équation 3.8[29]. Reste à déterminer les bonnes valeurs de fécondité relative à appliquer aux données sentinelles. Une première étude menée en 2003 a testé cette méthode sur 8 séries de données provenant d'Ouganda, de Tanzanie et de Zambie (FABIANI 2003) avec des coefficients d'ajustement calculés à partir de données provenant de Masaka, en milieu rural, en Ouganda (CARPENTER 1997). Les résultats ont été reproduits dans le Tableau 3.16. Il

[28] Dans une majorité des cas : 10 séries de données sur 12, voir Tableau annexe 1.2 et Figure annexe 1.3

[29] Voir l'annexe 1 pour plus de détails.

apparaît que pour 7 séries de données sur les 8 étudiées, cette méthode a permis de réduire l'écart avec la prévalence observée en population générale féminine.

Équation 3.8

Ajustement de la prévalence en clinique prénatale selon la fécondité relative

$$P_{PFG} = \sum_{\substack{groupes \\ d'âges}} F_{PFG}\left(\frac{P_{CPN}}{FR - FR \times P_{CPN} + P_{CPN}}\right)$$

P_{PFG} : *prévalence en population générale féminine*

P_{CPN} : *prévalence en clinique prénatale*

FR : *fécondité relative*

F_{PFG} : *proportion du groupe d'âges dans la population générale féminine*

Tableau 3.16

Comparaison sur 8 séries de données entre les prévalences du VIH ajustées selon deux méthodes et les prévalences observées en population féminine générale et en cliniques prénatales

	Prévalences observées (%)		Prévalences ajustées (%)	
Série de données Classe d'âges	**Femmes en population générale**	**Cliniques prénatales* (ER)**	**Première méthode† (ER)**	**Seconde méthode‡ (ER)**
Fort Portal, Uganda, 1994-95	26,0	17,9	25,9	NC
15-49 ans		(-31,2)	(-0,4)	
Mwanza, Tanzanie, 1990-1991	15,1	11,5	16,7	NC
15-49 ans		(-23,8)	(+10,6)	
Mwanza, Tanzanie, 1991-1993	4,7	3,2	5,3	4,6
15-44 ans		(-31,9)	(+12,8)	(-2,1)
Mposhi, Zambie, 1994	17,4	12,4	17,3	NC
15-39 ans		(-28,7)	(-0,6)	
Chelston, Lusaka, Zambie, 1994	29,9	23,7	30,9	32,0
15-39 ans		(-20,7)	(+3,3)	(+7,0)
Chelston, Lusaka, Zambie, 1996	29,9	23,5	31,0	NC
15-39 ans		(-21,4)	(+3,7)	
Chelston, Lusaka, Zambie, 1998	27,5	25,3	33,8	34,4
15-39 ans		(-8,0)	(+22,9)	(+25,1)
Ndola, Zambie, 1998	36,7	28,0	39,1	35,3
15-40 ans		(-23,7)	(+6,5)	(-3,8)

NC : non calculé, faute des données disponibles

ER : erreur relative, en %, par rapport à la prévalence observée auprès des femmes en population générale.

* ajustée par âge à partir de la structure par âge observée en population générale féminine.

† ajustement selon les ratios par âge de fécondité relative entre femmes VIH+ et VIH-

‡ ajustement selon la parité et les catégories de fécondité

Source : (FABIANI 2003).

Nous avons mené en 2004 une analyse comparable à partir de douze séries de données, portant sur neuf sites répartis dans sept pays, et de cinq séries de coefficients d'ajustement (LARMARANGE 2004). Les résultats détaillés ont été reproduits en annexe 1 et les principaux résultats sont présentés dans le Tableau

3.17. Il apparaît que la qualité de cette méthode d'ajustement est fortement dépendante de l'écart initial entre la prévalence observée en clinique prénatale et la prévalence de l'ensemble des femmes. La meilleure série de coefficients d'ajustement est donc dépendante de chaque contexte. Pour deux séries de données sur les douze étudiées, cette méthode d'ajustement induit un écart plus important que l'écart initial entre prévalence en clinique prénatale et prévalence de l'ensemble des femmes.

Tableau 3.17

Prévalences observées et estimées, erreurs relatives et moyenne des valeurs absolues des erreurs relatives pour chaque série de coefficients

Site	Prévalence du VIH (%)								Erreur Relative (ER en %)							Effectifs	
	PFG	CPN	CPN aj.	A	B	C	D	(B+C)/2	CPN	CPN aj.	A	B	C	D	(B+C)/2	PFG	CPN
Rural, Mposhi, Zambie, 1998 (Fylkesnes 2001)	16	8,3	7,8	11	9,1	11	13	9,6	-48,7	-52	-35,1	-43,7	-32	-23	-40,6	425	300
Urbain, Yaoundé, Cameroun, 1997-98 (Glynn 2001)	8,6	5,6	5,2	7,7	6,6	7,8	9,2	7	-34,6	-39,6	-10,2	-22,4	-9,3	7,3	-18,8	829	1 525
Urbain, Fort Portal, Ouganda, 1995 (Kilian 1999)	27	19	17	25	23	24	27	23	-28,7	-35,2	-7,1	-16,2	-11	0,1	-15,8	470	458
Rural, Karonga, Malawi, 1998-01 (Crampin 2003)	18	10	11	20	17	16	24	16	-40,3	-34,9	12,4	-2,1	-6,7	39,7	-6,4	287	2 908
Rural, Mwanza, Tanzanie, 1991-93 (Changalucha 2002)	4,8	3,7	3,4	5,5	4,7	5,3	6,4	4,8	-22,9	-29,2	13,8	-2	10,1	32,4	0,2	5. 089	2 153
Rural, Mposhi, Zambie, 1994 (Fylkesnes 2001, Fylkesnes 1998)	17	13	12	17	15	18	20	16	-27,7	-28,4	-2	-13,8	2	13,7	-9,2	426	422
Urbain, Ndola, Zambie, 1997-98 (Glynn 2001)	37	27	28	39	35	37	43	36	-25,2	-23,8	5	-4,6	0,8	18,1	-3,3	730	1 002
Rural, Manicaland, Zimbabwe, 1998-2000 (Gregson 2002)	25	21	20	28	25	27	31	25	-15,8	-22,1	8,6	-2,6	6,3	23,6	-0,6	4 452	1 162
Urbain, Chelston, Zambie, 1996 (Fylkesnes 2001)	30	26	24	31	28	32	34	29	-12,6	-21,4	2,3	-7,7	5,8	14,7	-3,4	1 211	532
Urbain, Chelston, Zambie, 1994 (Fylkesnes 2001, Fylkesnes 1998)	30	25	24	31	27	32	34	29	-17,6	-20,8	2,2	-8,2	7,5	14,8	-3,1	1 211	443
Urbain, Chelston, Zambie, 1998-99 (Fylkesnes 2001)	29	26	26	35	31	34	39	32	-9	-10	21,7	10,3	20,6	36,7	13,2	1 206	776
Urbain, Kisumu, Kenya, 1997-98 (Glynn 2001)	33	31	30	40	36	40	44	37	-7,2	-8,5	20,5	9,8	19,1	34,1	12,5	739	1 447
Moyenne des valeurs absolues des ER*									24,20	27,20	11,70	12,00	10,90	21,50	10,60		

PFG : Population Féminine Générale – CPN : femmes consultant en Clinique Prénatale – CPN aj. : prévalence observée en CPN ajustée par la structure par âge en PFG. A, B, C et D : prévalence en population générale estimée à partir des coefficients A, B, C ou D (voir annexe 2). ER : Erreur relative = (prévalence estimée – prévalence PFG) / prévalence PFG. * Lecture : L'écart moyen des prévalences observées en CPN par rapport à celles en PFG est de 24,2 %. L'écart des prévalences estimées avec la série de coefficients A par rapport à la prévalence observée en PFG est en moyenne de 11,7 %. **Source :** (LARMARANGE 2004).

Une fois ajustées, les prévalences estimées[30] à partir de la surveillance sentinelle des femmes présentent, en moyenne, un écart relatif avec la prévalence de l'ensemble des femmes de l'ordre de 10 à 12 %, avec des variations allant de -44 à +22 %. Alors que la prévalence observée en clinique prénatale induit systématiquement une sous-estimation, les prévalences ajustées sont tantôt supérieures tantôt inférieures à la prévalence de l'ensemble des femmes.

Le Tableau 3.16 présente également les résultats, pour quatre séries de données, d'une autre méthode d'ajustement proposée initialement par ZABA et ses collaborateurs (ZABA 2000). Les femmes sont divisées en plusieurs catégories de fécondité selon le fait qu'elles aient déjà eu un enfant (parité). Les différentes catégories sont les suivantes :

- Femmes n'ayant jamais eu d'enfant
    - N'a jamais eu de rapport sexuel
    - A déjà eu des rapports sexuels, actuellement sexuellement inactive
    - Sexuellement active, utilisant une méthode contraceptive
    - Sexuellement active, inféconde
    - Sexuellement active, enceinte du premier enfant (catégorie de référence pour les femmes sans enfants)
- Femmes ayant déjà eu au moins un enfant
    - Actuellement sexuellement inactive
    - Sexuellement active, utilisant une méthode contraceptive
    - Sexuellement active, inféconde
    - Sexuellement active, enceinte de second enfant ou enfant de rang supérieur (catégorie de référence pour les femmes ayant déjà eu un enfant)

Pour chaque catégorie de femmes, ZABA propose des coefficients d'ajustement ou prévalences relatives (PR) par rapport à la catégorie de référence (Tableau 3.18). Deux séries de coefficients sont proposées, l'une pour les populations où l'usage de méthodes contraceptive est faible, calculée à partir de données collectées à Kisesa, Tanzanie, entre 1994 et 1998 (BOERMA 1999) et à Maska, Ouganda, entre 1995 et 1996 (CARPENTER 1997) ; la seconde lorsque le recours à une contraception est élevée, estimée à partir de données collectées à Manicaland, Zimbabwe, en 1994 (GREGSON 1998). Seules les deux catégories de référence sont observées en clinique prénatale. La prévalence des autres catégories sera estimée à partir des coefficients de prévalence relative de ces catégories par rapport aux catégories de référence. La prévalence de l'ensemble des femmes sera ensuite calculée à partir du poids de

[30] Coefficients D exclus.

chaque catégorie dans l'ensemble de la population féminine. Ces poids nécessitent d'être déterminés à partir d'enquêtes en population générale. Les Enquêtes Démographiques et de Santé, notamment, permettent de calculer la répartition des femmes selon ces différentes catégories[31]. La procédure est résumée par l'Équation 3.9.

Équation 3.9

Ajustement de la prévalence en clinique prénatale selon les catégories de fécondité

$$P_{PFG} = \sum_{parité} P_{ANC} \times \left(\sum_{\substack{catégories \\ de\ fécondité}} F_{PFG} \times PR_{CPN} \right)$$

P_{PFG} *: prévalence en population générale féminine*

P_{CPN} *: prévalence en clinique prénatale*

PR_{CPN} *: prévalence relative par rapport à celle en clinique prénatale*

F_{PFG} *: proportion de la catégorie dans la population féminine générale*

Tableau 3.18

Prévalence du VIH et prévalence relative, par catégorie de fécondité, estimées en population générale dans des contextes de fort et de faible usage de méthodes contraceptives

Catégories de fécondité	Contraception faible		Contraception importante	
	Prévalence VIH % (n)	Prévalence relative	Prévalence VIH % (n)	Prévalence relative
Femmes sans enfant	*(1 197)*		*(1 354)*	
N'a jamais eu de rapports sexuels	0,6 (510)	0,10	2,1 (933)	0,09
Sexuellement inactive	1,6 (129)	0,27	20,9 (67)	0,92
SA, utilise une contraception	14,3 (35)	2,48	22,0 (41)	0,96
SA, inféconde	19,5 (159)	3,38	52,0 (50)	2,28
SA, féconde (référence)	5,8 (364)	1,00	22,8 (263)	1,00
A déjà eu un enfant	*(4 206)*		*(3 775)*	
Sexuellement inactive	16,4 (256)	2,09	41,5 (615)	1,49
SA, utilise une contraception	15,6 (243)	1,99	26,4 (615)	0,95
SA, inféconde	12,5 (634)	1,59	39,8 (595)	1,43
SA, féconde (référence)	4,8 (3 073)	1,00	27,8 (935)	1,00
Total	*(5 373)*		*(5 129)*	

Source : (ZABA 2000). - SA : sexuellement active

[31] Y compris avant que des modules de dépistage du VIH n'y soient insérés.

Sur les quatre sites où elle a été appliquée (Tableau 3.16), cette méthode s'est avérée efficace pour trois sites, et sur deux sites elle a plus réduit l'écart avec la prévalence de l'ensemble des femmes que la première méthode. Cependant, elle s'avère souvent plus difficile à mettre en œuvre dans la mesure où il est plus aisé d'obtenir la structure par âge de l'ensemble des femmes que leur structure par catégories de fécondité. De plus, avec la mise en place de la surveillance sentinelle de seconde génération, s'il est devenu usuel de collecter l'âge des femmes testées, il n'en est pas de même du rang de leur naissance.

La principale difficulté de ces deux méthodes consiste à déterminer les coefficients d'ajustement adéquats à appliquer à une clinique sentinelle donnée. Massimo FABIANI et ses collaborateurs, ainsi que nous-mêmes, dans les deux études présentées, avons appliqué aux différentes séries de données des coefficients d'ajustement calculés sur d'autres populations. Or, si la fécondité relative des femmes VIH+ présente un pattern semblable en différents points du continent africain, la valeur précise des coefficients varie dans le temps et dans l'espace. Le recours à ces techniques permet néanmoins, en moyenne, de réduire les écarts entre la prévalence des femmes suivies en clinique prénatale et la prévalence de l'ensemble des femmes que l'on cherche à estimer. Cependant, la marge d'incertitude reste relativement importante et il n'est pas possible, *a priori*, de savoir si après ajustement la prévalence estimée sur- ou sous-estime la prévalence que l'on cherche à atteindre. Par contre, nous pouvons considérer que la prévalence observée en clinique prénatale, brute, sans aucun ajustement, constitue systématiquement une sous-estimation de la prévalence de l'ensemble des femmes et donc que cette dernière est au moins égale à la première, localement.

Reprenant une idée proposée par NICOLL et al. (1998), Annabel DESGRÉES DU LOÛ et ses collaborateurs (1999) ont ajusté la prévalence observée parmi les femmes enceintes d'Abidjan, Côte d'Ivoire, à l'aide d'un ratio d'inclusion relative calculé directement à partir de l'histoire génésique des femmes enceintes observées en clinique prénatale. Massimo FABIANI *et al.* (2006) ont repris cette idée dans une étude récente où ils ont estimé les différentiels de fécondité par âge et statut sérologique à l'aide de modèles de Cox sur l'histoire génésique des femmes enquêtées en clinique prénatale. Cela permettrait ainsi de pouvoir calculer les coefficients d'ajustement directement à partir des données sentinelles sans avoir besoin de réaliser une enquête complémentaire en population générale. Cependant, ils ont montré que les fécondités relatives calculées à partir des femmes enceintes suivies en clinique prénatale, si elles présentent des patterns du même ordre que ceux observés en population générale, sous-estimaient ces dernières et s'avéraient donc inappropriées pour extrapoler la prévalence de l'ensemble des femmes à partir des femmes enceintes.

Par ailleurs, il a été montré que ces différentes méthodes d'ajustement étaient moins efficaces dans les zones où les pratiques contraceptives étaient courantes (ZABA 2000, GREGSON 2002b, FABIANI 2003). Les interférences de la contraception sur la surveillance sentinelle en clinique prénatale sont mal connues. Si par endroit il a été observé que les femmes VIH+ utilisaient plus souvent une méthode contraceptive (RYDER 1991), le type de méthode contraceptive utilisé peut également intervenir.

*
**

Au final, les prévalences du VIH mesurées auprès des femmes enceintes en clinique prénatale s'avèrent être, en l'absence d'ajustement, une estimation *a minima* de la prévalence de l'ensemble des femmes de la même zone géographique. Plusieurs méthodes d'ajustement permettent de réduire, dans une majorité des cas, l'écart entre la prévalence observée en clinique prénatale et celle de l'ensemble des femmes que l'on cherche à estimer. Cependant, même ainsi, les écarts entre la prévalence estimée et la valeur de cette dernière parmi l'ensemble des femmes peuvent être importants (entre -50 et +25 %) et ne sont plus systématiquement du même signe. Par contre, les ordres de grandeur sont respectés. Si la prévalence des femmes enceintes est de l'ordre de 2 %, alors nous pourrons en déduire que la prévalence de l'ensemble des femmes est de l'ordre de 2 à 4 %, mais il sera extrêmement improbable[32] qu'elle soit de 10 %.

Par contre, les données de surveillance sentinelle des femmes enceintes ne peuvent être utilisées pour étudier les différentiels de prévalence entre plusieurs catégories. En effet, les différents biais de sélections et les variations des taux de fécondité et des taux de consultation en clinique prénatale induise que les observations en cliniques prénatales peuvent présenter des patterns différents. Ainsi, des comparaisons locales ont pu mettre en évidence des prévalences moindres du VIH parmi les femmes les plus instruites tandis que les données auprès des femmes enceintes montraient quant à elles, pour les mêmes aires géographiques, l'inverse (FYLKESNES 2001, GREGSON 2001, GREGSON 2002b).

3.4.2 Femmes enceintes et population générale

Nous nous interrogeons dorénavant sur la représentativité des femmes enceintes testées en cliniques prénatales vis-à-vis de l'ensemble de la population adulte, hommes et femmes confondus, toujours à un niveau local. Il nous faut donc prendre en compte les différentiels de prévalence selon le sexe. Si les liens entre prévalences féminine et masculine sont peu connus, il a été montré depuis

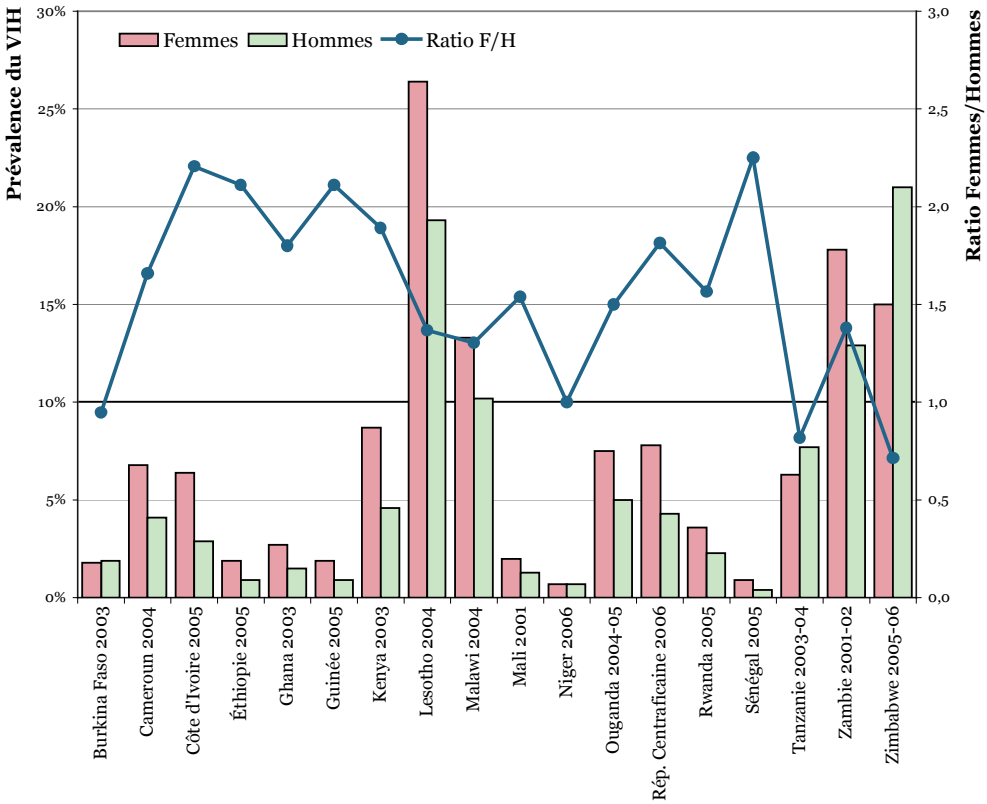
[32] Rappelons qu'il existe toujours un risque d'erreur radical.

longtemps que la prévalence du VIH était en général plus faible chez les hommes que chez les femmes (BERKLEY 1990, GREGSON 2000, GLYNN 2001b). Cependant, les ratios prévalence des femmes sur prévalence des hommes peuvent varier grandement d'un pays à l'autre (Figure 3.19).

Quelques auteurs ont proposé d'estimer la prévalence des hommes à partir des femmes enceintes consultant en cliniques prénatales en leur demandant l'âge du père de l'enfant et en appliquant à ce dernier le statut sérologique dépisté chez la mère (GLYNN 2001a). Cependant, cette méthode d'ajustement n'a pas produit de résultats concluant pour tous les sites où elle a été testée. D'une part, les jeunes hommes sont peu représentés parmi les compagnons des femmes testées, d'où une marge d'erreur importante quant à l'estimation de leur prévalence, alors qu'ils représentent une part importante de la structure par âge de la population générale. Par ailleurs, cette technique présuppose que les hommes présentent le même statut sérologique que leur compagne, ce qui n'est pas systématique (HIRA 1990, SERWADDA 1995, GRAY 1998, CARPENTER 1999).

Figure 3.19

Prévalence nationale du VIH à 15-49 ans selon le sexe et ratio prévalence des femmes sur prévalence des hommes, pour 18 pays d'Afrique subsaharienne



Sources : rapport final de chaque enquête disponible sur http://www.measuredhs.com. Il s'agit d'EDS, à l'exception de la Côte d'Ivoire (EIS), de l'Ouganda (HSBS), de la République Centrafricaine (ESV) et de la Tanzanie (AIS). Ce sont toutes des enquêtes nationales réalisées en population générale.

Tableau 3.19
Prévalence du VIH observée localement en cliniques prénatales et en population générale (hommes et femmes)

Site	Âges	Prévalence en CPN (n)	IC 95 %	Prévalence en PG (n)	IC 95 %	Écart relatif
Fort Portal, Ouganda, 1995	15-49	18,4 (477)	15,1-22,2	22,9 (875)	20,2-25,9	-19,7 %
Lusaka, Zambie, 1995-96	15-39	26,1 (532)	22,5-30,1	25,7 (1 909)	23,8-27,7	1,6 %
Mposhi, Zambie, 1994-961	15-39	12,6 (422)	9,7-16,2	16,7 (760)	14,2-19,6	-24,6 %
Chelston, Zambie, 1998-99	15-39	25,9 (776)	22,9-29,2	23,0 (1 768)	21,1-25,0	12,6 %
Mposhi, Zambie, 1998-99	15-39	8,3 (300)	5,5-12,2	16,7 (724)	14,1-19,7	-50,3 %
Yaoundé, Cameroun, 1997-98	15-49	5,5 (1 532)	4,4-5,8	6,1 (1 913)	5,1-7,3	-9,8 %
Kisumu, Kenya, 1997-98	15-49	30,5 (1 480)	28,2-32,9	25,9 (1 515)	23,7-28,2	17,8 %
Ndola, Zambie, 1997-98	15-49	27,3 (1 021)	24,6-30,2	28,4 (1 534)	26,2-30,7	-3,9 %
Manicaland, Zimbabwe, 98-2000	15-44	21,5 (1 215)	19,2-23,9	22,5 (9 119)	21,6-23,4	-4,4 %
Mwanza, Tanzanie, 1991-93	15-44	3,6 (2 265)	2,9-4,5	4,7 (5 675)	4,2-5,3	-23,4 %
Karonga, Malawi, 98-2001	15-49	10,4 (3 013)	9,3-11,6	17,0 (342)	13,3-21,5	-38,8 %
Kagera, Tanzanie 1987-90	15-49	22,8 (1 292)	20,6-25,2	29,2 (325)	24,4-34,5	-21,9 %
Kagera, Tanzanie 1993	15-49	17,3 (2 816)	15,9-18,8	18,7 (395)	15,0-23,0	-7,5 %
Kagera, Tanzanie 1996	15-49	13,0 (2 893)	11,8-14,3	14,4 (787)	12,1-17,1	-9,7 %
Bukoba, Tanzanie, 1987-90	15-54	22,4 (1 292)	20,2-24,8	24,4 (553)	20,9-28,2	-8,2 %
*Écart relatif moyen**						*16,9 %*

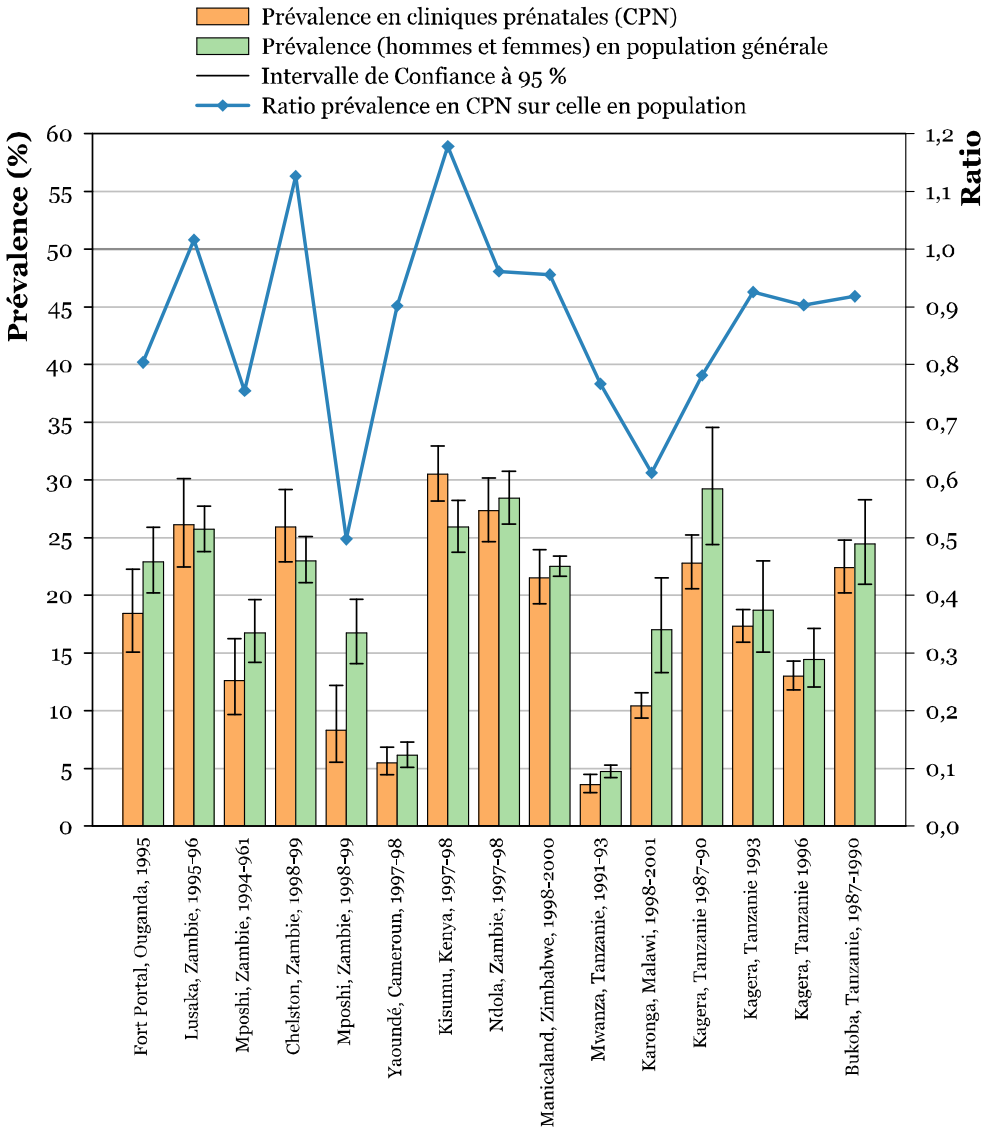
Les prévalences sont exprimées en pourcents. n : effectifs.
IC 95 % : intervalle de confiance à 95 %.
CPN : cliniques prénatales – PG : population générale (hommes et femmes).
* Il s'agit de la moyenne des valeurs absolues des écarts relatifs.
Sources : (KWESIGABO 1996, FYLKESNES 1998, KILIAN 1999, KWESIGABO 2000, FYLKESNES 2001, GLYNN 2001a, CHANGALUCHA 2002, GREGSON 2002b, CRAMPIN 2003), nos calculs.

Nous avons montré à la section précédente que la prévalence brute observée chez les femmes enceintes sous-estimait systématiquement la prévalence de l'ensemble des femmes. Or, comme la prévalence du VIH est en général plus faible chez les hommes, ces deux biais se compensent en partie. Ainsi, la prévalence brute observée chez les femmes enceintes est, le plus souvent, plus proche de la prévalence de l'ensemble des adultes que de celle de l'ensemble des femmes. Sur quinze sites où une comparaison locale est possible, nous avons calculé une erreur

relative moyenne de 17 % (Tableau 3.19), avec des variations de -50 % à +18 %. Cependant, alors que, par rapport à l'ensemble des femmes, la prévalence en clinique prénatale était systématiquement inférieure, par rapport à l'ensemble des adultes elle est tantôt inférieure, tantôt supérieure (Figure 3.20).

Figure 3.20

Comparaisons locales entre prévalence du VIH observée en clinique prénatale et prévalence du VIH en population générale (hommes et femmes)



Sources : (KWESIGABO 1996, FYLKESNES 1998, KILIAN 1999, KWESIGABO 2000, FYLKESNES 2001, GLYNN 2001a, CHANGALUCHA 2002, GREGSON 2002b, CRAMPIN 2003), nos calculs.

Au final, la surveillance sentinelle des femmes enceintes pourra fournir une estimation, au niveau local, de l'ordre de la prévalence parmi la population générale adulte. Cependant, cette estimation restera relativement imprécise.

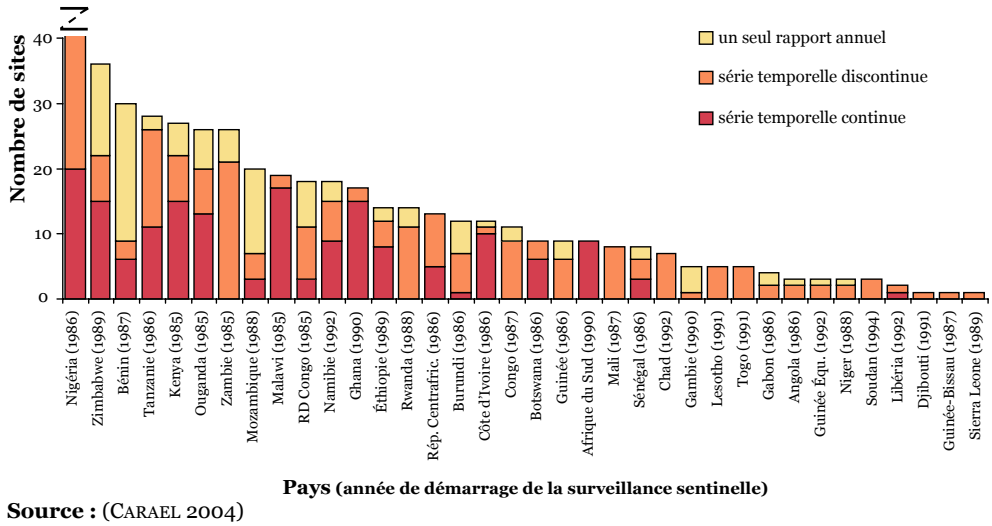
3.4.3 Sélection et localisation des cliniques sentinelles

Si, localement, la prévalence des femmes enceintes testées en clinique prénatale est du même ordre de grandeur que la prévalence de la population générale adulte, au niveau national cette question est beaucoup plus complexe. En effet, elle dépend largement du choix et de la localisation des cliniques prénatales retenues pour la surveillance sentinelle (JACKSON 1999, SCHWARTLANDER 1999). La Figure 3.22, la Figure 3.23 et la Figure 3.24 présentent la localisation des sites sentinelles enquêtés récemment (sur la période 2002-2005) au Burkina Faso, au Cameroun et au Kenya, afin d'avoir trois exemples à l'esprit.

La situation est très inégale d'un pays à l'autre, certains ayant un très grand nombre de sites sentinelles, comme le Nigéria ou le Zimbabwe, tandis que d'autres comptabilisent moins d'une demi-douzaine de sites (Figure 3.21). Par ailleurs, le nombre et la localisation des sites sentinelles varient au cours du temps au sein d'un même pays[33].

Figure 3.21

Continuité de la surveillance sentinelle des femmes enceintes consultant en clinique prénatale en Afrique subsaharienne



Source : (CARAEL 2004)

Pendant les années 1990, les femmes enceintes ont constitué la principale source de données pour l'estimation des prévalences nationales du VIH[34]. Très tôt, les experts d'ONUSIDA avaient conscience de la faible représentativité des sites sentinelles par rapport au milieu rural, les Epidemiological Fact Sheets distinguant

[33] Voir également Figure 1.4 page 32 et Figure 1.7 page 40 sur la qualité de l'implémentation des systèmes de surveillance sentinelle.

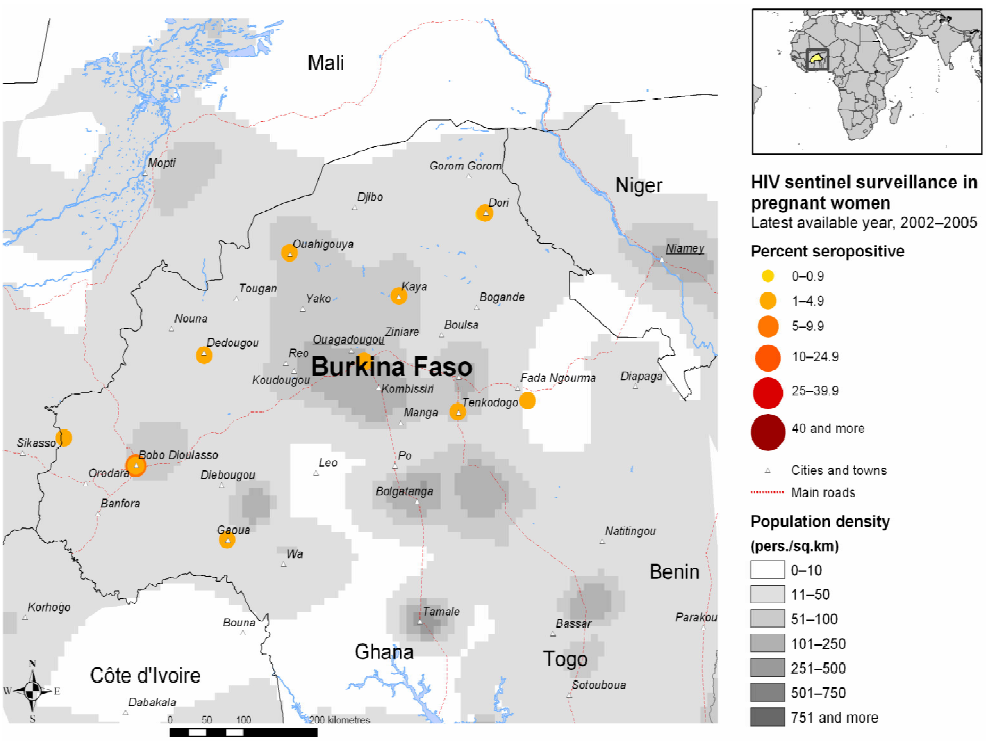
[34] *Cf.* la section 1.6 page 37 et suivantes.

d'ailleurs principales zones urbaines et milieu périurbain. Pour les estimations nationales, la prévalence observée parmi les femmes enceintes en milieu périurbain était ainsi ajustée selon un facteur de réduction (SCHWARTLANDER 1999), le plus souvent 80 % (GHYS 2004).

Avec l'arrivée des enquêtes nationale en population générale, il s'est avéré que les estimations nationales réalisées à partir des femmes enceintes étaient le plus souvent surélevées, avec des écarts plus ou moins importants selon le pays (*cf.* Figure 1.8 page 52).

Plusieurs études récentes ont été publiées récemment comparant, nationalement ou régionalement, des estimations réalisées à partir de femmes enceintes et des données d'enquêtes en population générale. Les conclusions des auteurs divergent d'un article à l'autre.

Figure 3.22
Localisation des sites sentinelles au Burkina Faso



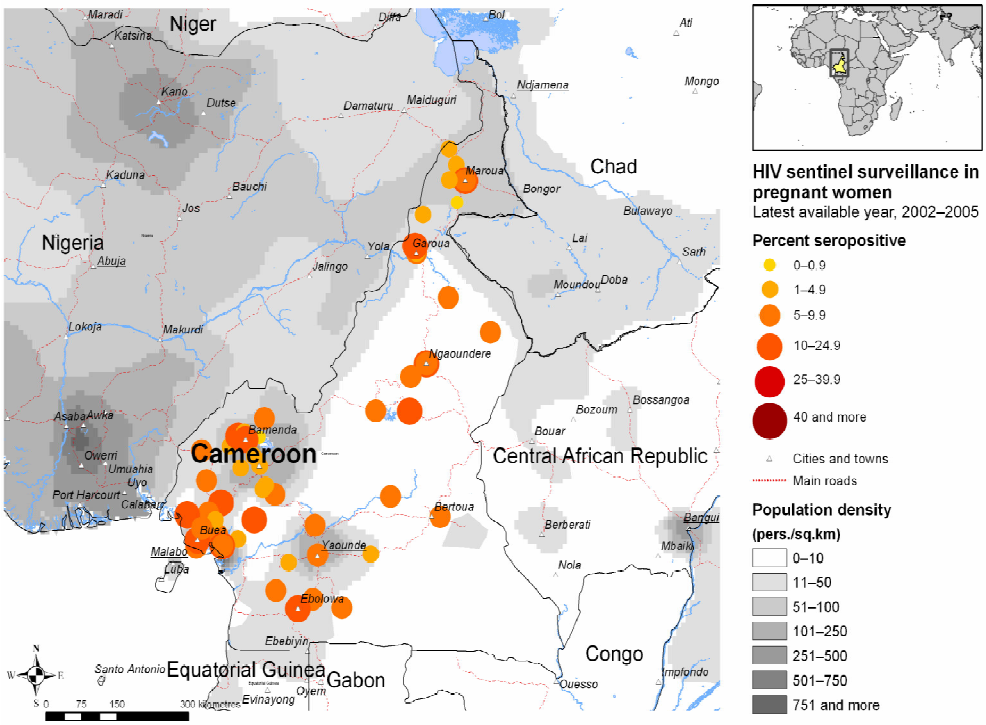
Source : (WHO 2006a).

À partir d'une enquête réalisée en 2003-2004 dans deux régions rurales de Tanzanie (Manyara et Singida), Khadija I. YAHYA-MALIMA et ses collaborateurs concluaient que la prévalence du VIH en clinique prénatale était comparable dans la région à la prévalence observée en population générale et suggèrent donc que les femmes enceintes sont bien représentatives de la population adulte pour l'estimation des prévalences (YAHYA-MALIMA 2007). Cependant, leur article ne

fournit pas la prévalence brute observée en clinique prénatale mais simplement les prévalences par âges. Si celles-ci sont proches de celles observées pour les femmes de plus de 20 ans en population générale, elles divergent relativement pour les hommes et les jeunes femmes.

Kumbutso DZEKEDZEKE et Knut FYLKESNES ont comparé les résultats de l'EDS 2001-2002 de la Zambie avec la surveillance sentinelle de la même année au niveau national (DZEKEDZEKE 2006). La prévalence nationale 15-49 ans de l'EDS de 15,6 % s'est avérée proche de la prévalence de 16,9 % observée en clinique prénatale à la même époque et ajustée selon la structure par âge et par milieu de résidence du pays (la prévalence nationale non ajustée étant de 20,1 %). Les auteurs concluent que la prévalence nationale du VIH peut être obtenue à partir de la surveillance sentinelle en clinique prénatale dès lors que la couverture du pays est importante et que les taux de consultation en clinique prénatale et la fécondité sont élevés.

Figure 3.23
Localisation des sites sentinelle au Cameroun



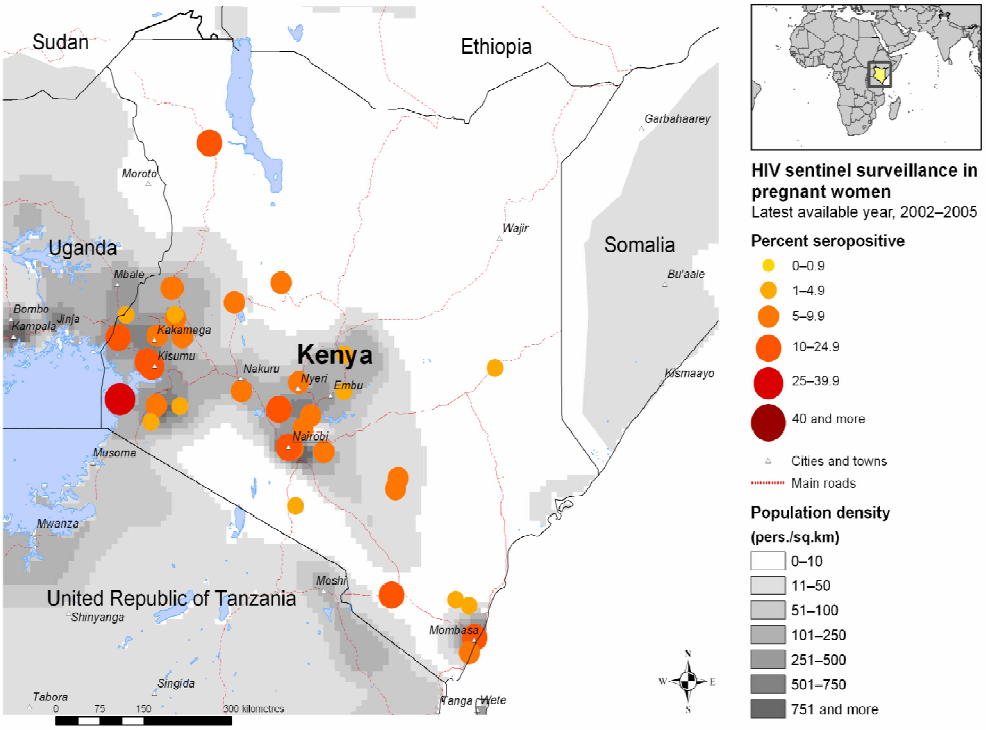
Source : (WHO 2006c).

En Éthiopie, par contre, il a été montré que la surveillance en clinique prénatale dans la région d'Afar surreprésentait les résidents urbains et surestimait ainsi la prévalence du VIH de l'ensemble de la région (ASSEFA 2003).

Une étude, menée en 2005 en Afrique du Sud, en milieu rural dans la région de KwaZulu-Natal, a mis en évidence que, dans une région de fécondité faible pour l'Afrique, la surveillance sentinelle surestimait la prévalence du VIH (RICE 2007).

Des résultats similaires ont été trouvés en Asie, la surveillance sentinelle des femmes enceintes surestimant la prévalence du VIH dans le sud de l'Inde (DANDONA 2006) et au Cambodge (SAPHONN 2002).

Figure 3.24
Localisation des sites sentinelle au Kenya



Source : (WHO 2006b).

Dans une étude présentée en 2005 et portant sur dix pays d'Afrique sub-saharienne (BIGNAMI-VAN ASSCHE 2005), les prévalences nationales du VIH estimées à partir des systèmes de surveillance sentinelle étaient systématique supérieures à celles obtenues à partir d'enquêtes nationales en population générale (voir Tableau 3.20 ci-dessous). Or, comme nous avons montré à la section 3.3 que les biais liés aux enquêtes nationales en population générale sont relativement faibles, les écarts entre les estimations de ces deux sources devraient être imputables majoritairement à des biais de représentativité de la surveillance sentinelle des femmes enceintes.

Tableau 3.20
Comparaison des estimations des prévalences du VIH (en %) des adultes selon des enquêtes nationales en population générale et des données de surveillance sentinelle des femmes enceintes, selon le milieu de résidence

Pays	Population générale (urbain)	Surveillance sentinelle (urbain)	Population générale (rural)	Surveillance sentinelle (rural)
Burkina Faso	3,1	5,3	1,0	4,2
Burundi	13,0	16,0	2,5	4,5
Rép. Dominicaine	0,9	1,2	1,2	2,2
Ghana	2,3	5,1	2,0	5,1
Kenya	10,0	14,4	5,6	11,6
Mali	2,2	5,8	1,5	3,2
Niger	2,1	2,0	0,6	2,5
Afrique du Sud	16,7	27,6	8,3	26,2
Zambie	23,1	26,8	10,8	14,4
Zimbabwe	5,0	30,6	18,0	28,5

Source : (BIGNAMI-VAN ASSCHE 2005).

Il n'y a pas de règle générale concernant la capacité des systèmes de surveillance sentinelle auprès des femmes enceintes à fournir des estimations de la prévalence nationale du VIH proches de la réalité.

Au niveau d'une clinique prénatale, le recrutement des femmes enceintes pour la surveillance sentinelle correspond en règle générale à l'ensemble des femmes se présentant pour une première consultation sur une période de temps fixée ou bien jusqu'à atteindre un certain effectif. Éventuellement, seule une fraction des femmes peut être enquêtée (la moitié ou le tiers par exemple). Il s'agit d'une méthode d'échantillonnage tout à fait classique assurant ainsi à l'échantillon testé une bonne représentativité vis-à-vis de l'ensemble des femmes enceinte consultant la clinique considérée. Puisque les femmes ne sont recrutées que lors de leur première visite, les doublons sont évités.

À la section précédente, nous avons montré que malgré l'ensemble des biais intervenant, la prévalence brute observée parmi les femmes enceintes testées dans une clinique prénatale était plus ou moins représentatives de l'ordre de grandeur de la prévalence du VIH de l'ensemble des adultes, hommes et femmes, de la même zone géographique. S'il est difficile de définir celle-ci avec précision, elle correspond plus ou moins au périmètre au sein duquel vivent les femmes qui viennent consulter dans cette clinique. Sa superficie est donc relativement restreinte.

Qu'en-t-il par contre de la sélection des cliniques dans lesquelles les enquêtes de surveillance sont conduites ?[35] Le choix de ces cliniques repose sur un choix raisonné réalisé par les institutions en charge de la surveillance sentinelle. Depuis le développement de la surveillance de seconde génération, une majorité de pays essaye de sélectionner au moins un site sentinelle dans chaque région et d'étendre la surveillance au milieu rural qui était, et reste parfois, largement sous-représenté. Quelque soit la pertinence des arguments ayant guidé au choix de tel ou tel site, il ne s'agit jamais d'une méthode d'échantillonnage probabiliste sur laquelle nous pourrions nous appuyer pour déterminer le meilleur mode de calcul à partir des données d'observations brutes[36].

De ce fait, quelle méthode retenir pour l'estimation de la prévalence nationale du VIH à partir des résultats de l'ensemble des sites sentinelles retenus ? Plusieurs techniques ont pu être employées mais aucune n'est pleinement satisfaisante. Les Epidemiological Fact Sheets fournissent pour leur part la valeur médiane, en milieu urbain et en milieu périurbain, des prévalences observées sur chaque site. Certains rapports nationaux utilisent la prévalence moyenne (CNLS-IST 2004) ou bien présentent moyenne et médiane (GROUPE ÉPIDÉMIOLOGIE PNLS 2004). D'autres calculent la prévalence sur l'ensemble de l'échantillon collecté en fusionnant les échantillons de chaque site[37] (NATIONAL AIDS CONTROL COMMITTEE 2001, 2003). Parfois, la prévalence nationale est estimée par extrapolation pour chaque province puis ajustement sur le nombre d'habitants de chacune d'elles (GROUPE DE SURVEILLANCE SÉRO-ÉPIDÉMIOLOGIQUE 1999). Dans d'autres cas, la moyenne, la médiane ou un sous-échantillon global sera utilisé pour calculer des prévalences par milieu de résidence et la prévalence nationale sera ensuite calculée en tenant compte de la répartition de la population nationale par milieu de résidence.

Ces différentes approches reposent chacune sur des hypothèses anticipatrices différentes. La médiane est le plus souvent préférée à la moyenne car elle est moins influencée par les valeurs extrêmes. Dans les deux cas, chaque site sentinelle a le même poids dans le calcul de la prévalence nationale. Or, ils représentent des populations de tailles différentes. Plusieurs pays ont sélectionné un site sentinelle par région. L'hypothèse est donc faite que ce site sentinelle est représentatif de la

[35] Nous n'aborderons pas ici les problèmes de comparabilité des données de surveillance sentinelle qui seront, pour leur part, abordés ultérieurement dans la thèse.

[36] Il peut y avoir de rares exceptions comme l'enquête ESSIDAGUI réalisée en Guinée en 2001. Dans cette enquête, un tiers des centres de santé du pays était tirés aléatoirement pour participer au dépistage du VIH des femmes enceintes. L'échantillon des centres de santé était donc relativement représentatif de l'ensemble des centres de santé, ce qui simplifie la problématique même si cela n'est pas suffisant car il reste la question des poids à attribuer à chaque centre de santé. BARRY A., KABA D. et DIOP I., *Rapport final ESSIDAGUI/2001*, Conakry (GN), Stat-View International, 2002.

[37] La prévalence nationale correspond alors tout simplement au nombre total de femmes dépistées positives sur le nombre total de femmes testées, tous sites confondus.

région à laquelle il appartient. Mais toutes les régions ne sont pas aussi peuplées les unes que les autres. Le recours à la moyenne surreprésentera les régions ayant un petit effectif de population tandis que les régions importantes seront sous-représentées.

Calculer la prévalence comme la proportion de résultats positifs sur l'ensemble de l'échantillon induit un ajustement sur le nombre de personnes testées par site. Si le nombre de femmes enquêtées est le même dans chaque site, cette opération sera équivalente à la moyenne de la prévalence de chaque site. Mais, en pratique, le nombre de femmes testées par site peut varier de manière plus ou moins importante. Dans ce cas-là, les sites les plus enquêtés seront plus représentés que les sites où un nombre moindre de femmes a été testé. Si ce sont les sites des régions les plus peuplées qui ont les échantillons les plus importants, alors l'échantillon global se rapprochera de la structure par région du pays. À l'inverse, la structure par région de l'échantillon global peut s'éloigner sensiblement de la structure par région de la population nationale.

Les ajustements par milieu de résidence ou par région sont, de ce point de vue, préférables. Reste qu'ils nécessitent toujours de poser l'hypothèse que les prévalences du VIH observées dans les sites sentinelles retenus sont représentatifs chacun de leur région. Nous connaissons mal les variations infrarégionales de la prévalence du VIH. Nous avons montré que la prévalence variait de manière importante selon le milieu de résidence (voir Figure 3.11 page 152) et selon les régions (voir Figure 3.13 page 154). Elle doit donc également varier de manière plus ou moins importante à l'intérieure d'une même région. En Afrique du Sud, dans la région du Western cap, une étude menée en 2001-2004 a mesurée la prévalence du VIH au sein des 344 cliniques prénatales de la province (SHAIKH 2006). Les variations de la prévalence du VIH d'un district de santé étaient particulièrement importantes, de 0,6 à 22 % en 2001 et de 1 à 33 % en 2004. Dès lors, la représentativité d'un site sentinelle vis-à-vis de la région qu'il est censé représenter dépend de sa localisation et des variations infrarégionales de l'épidémie. Si le site se situe dans une zone présentant une prévalence proche de la prévalence moyenne régionale, il fournira une estimation plus ou moins bonne. Mais s'il se situe au sein d'un pic épidémique, il peut surestimer grandement la prévalence réelle de sa région.

Or, une majorité de sites sont situés en milieu urbain ou périurbain, ce qui peut expliquer en partie que la surveillance sentinelle des femmes enceintes a tendance à surestimer la prévalence réelle. Dans une étude menée en 2003-2004 en milieu rural au nord de la Tanzanie, des cliniques prénatales reculées ont été sélectionnées pour une enquête de séroprévalence (YAHYA-MALIMA 2006). Les auteurs ont ainsi comparés la prévalence du VIH mesurée dans les cliniques prénatales sélectionnées d'ordinaire pour la surveillance nationale et celle observée dans les cliniques prénatales plus reculées non investiguées d'ordinaire. La prévalence du VIH

observée était de 2,6 % dans les premières et de 1,3 % dans les secondes, montrant ainsi que les cliniques prénatales retenues pour la surveillance nationale surestimaient l'épidémie réelle de la région étudiée.

Simona Bignami-Van Assche et ses collaborateurs ont eu recours à des modèles statistiques multivariés pour analyser au Kenya les caractéristiques sociodémographiques des individus enquêtés dans le cadre de la surveillance sentinelle des femmes enceintes et de l'EDS 2003 (Bignami-Van Assche 2005). Les auteurs ont montré que les caractéristiques individuelles des femmes enceintes enquêtées ne pouvaient expliquer les différences observées avec l'EDS et ils ont conclu que ces dernières devaient probablement être dues principalement à une localisation des sites sentinelles dans des zones présentant des prévalences élevées.

⁂

Ainsi, si la surveillance sentinelle des femmes enceintes peut permettre de fournir, localement, une estimation de l'ordre de grandeur de la prévalence du VIH parmi les adultes en population générale (hommes et femmes confondus), au niveau national c'est, pardonnez nous l'expression, « au petit bonheur la chance ». La situation est très inégale d'un pays à l'autre. Cependant, les sites sentinelles retenus n'ayant chacun qu'une représentativité locale, les niveaux de prévalence qu'ils estiment peuvent donc surestimer ou sous-estimer l'aire géographique qu'ils sont censés représenter. En l'absence d'une connaissance précise des variations infrarégionales de la prévalence du VIH et de la zone couverte par chaque site sentinelle, il est difficile voir impossible d'élaborer une méthode d'ajustement satisfaisante pour l'estimation de la prévalence nationale, à moins de poser des hypothèses fortes, non vérifiées empiriquement, sur la capacité des sites sentinelles à traduire correctement les niveaux de prévalence de chaque région. Le Tableau 3.20 montre d'ailleurs que les écarts peuvent être particulièrement importants par endroits, la prévalence du VIH estimée à partir des femmes enceintes pouvant être jusqu'à quatre à six fois supérieure à celle observée en population générale.

3.5 Cliniques prénatales et EDS : des échelles différentes

L'analyse de la représentativité nationale des Enquêtes Démographiques et de Santé et de la surveillance sentinelle s'est révélée déséquilibrée.

Pour les EDS et les enquêtes de même type, nous avons pu aborder avec précision cinq sources différentes de biais : populations hors ménages, ancienneté de la base de sondage, ménages non enquêtés, individus non testés et fenêtre sérologique des stratégies de dépistage utilisées. L'ampleur de ces biais a été estimée et nous avons montré que, bien qu'en maximisant ces différents biais, la prévalence nationale ajustée était proche de la prévalence nationale observée. De plus, l'erreur due à ces biais s'est avérée inférieure à l'imprécision statistique liée à l'échantillonnage, les prévalences ajustées pour le Burkina Faso, le Cameroun et le Kenya se situant à l'intérieur des intervalles de confiance à 95 % des prévalences observées. Enfin, les EDS, de part leur échantillonnage, sont également représentatives au niveau régional et selon le milieu de résidence. Elles permettent également de calculer des prévalences selon les caractéristiques des personnes interrogées, par âge ou par niveau d'instruction par exemple. Par contre, en raison de leurs effectifs limités, les EDS ne sont pas appropriées pour l'estimation des prévalences à un niveau local.

La surveillance sentinelle des femmes enceintes, quant à elle, est soumise à de nombreuses sources de biais difficilement identifiables et quantifiables. Localement, nous pouvons considérer que la prévalence mesurée parmi les femmes enceintes fournit une estimation *a minima* de la prévalence de l'ensemble des femmes et une estimation de l'ordre de grandeur de la prévalence parmi la population générale, hommes et femmes. Par contre, elle ne peut être utilisée pour le calcul de prévalences du VIH ventilées selon les caractéristiques sociodémographiques des individus car les différents biais de sélection peuvent varier grandement d'une catégorie à une autre et les patterns observés parmi les femmes enceintes peuvent être fortement divergents de ceux de la population générale. Au niveau national, les estimations du niveau de la prévalence du VIH seront plus ou moins proches de la prévalence réelle en fonction de la configuration particulière de la localisation des sites sentinelles retenus pour la surveillance et des variations infrarégionales de la prévalence du VIH dans le pays considéré. Suivant les situations, les biais de mesure pourront se compenser et l'estimation nationale réalisée parmi les femmes enceintes être proche du niveau réel de l'épidémie dans le pays. À l'inverse, dans d'autres contextes les écarts pourront être particulièrement importants. Dans certains pays, la prévalence parmi les femmes enceintes s'est avérée jusqu'à quatre à six fois plus élevée que celle observée en population générale.

Ainsi, les enquêtes nationales en population générale sont appropriées pour des estimations de prévalence au niveau national et régional tandis que la surveillance sentinelle des femmes enceintes fournit des estimations essentiellement locales. Nous ne pouvons donc les comparer en toute rigueur directement puisque leurs ensembles populations-espace-temps ne se recoupent pas.

Une majorité d'EDS collecte par ailleurs les coordonnées géographiques (longitude et latitude) des grappes enquêtées. Il semble donc possible, à partir de ces informations, de pouvoir réaliser, à partir des EDS, des estimations locales de la prévalence du VIH, estimations qui pourraient alors être comparées aux données de surveillance sentinelle des femmes enceintes.

Nous aborderons dans le chapitre suivant comment nous pouvons tirer partie de ces informations afin d'estimer les variations spatiales de la prévalence du VIH à des niveaux infrarégionaux.

Chapitre 4
Des populations aux territoires : l'apport cartographique

La collecte des coordonnées géographiques (longitude/latitude) des zones d'enquêtes dans les Enquêtes Démographiques et de Santé ouvre la voie à la cartographie des variations spatiales des prévalences du VIH à l'aide des techniques de géostatistiques[1].

Nous aborderons dans un premier temps le principe de l'interpolation spatiale et la reconstruction qu'elle induit du concept opératoire de prévalence en celui de prévalence localisée (section 4.1). Puis, nous regarderons les données à notre disposition dans les EDS de deux pays : le Burkina Faso et le Cameroun (section 4.2). Au niveau des zones d'enquêtes, les effectifs sont trop faibles pour permettre de réaliser directement une interpolation spatiale. Il s'avère nécessaire d'estimer au préalable la prévalence de chaque grappe. Nous aurons recours à un pays modèle, que nous avons nommé Alicante[2], pour simuler des EDS (section 4.3) et pouvoir tester ainsi une approche méthodologique.

[1] Techniques statistiques appliquées aux objets géoréférencés.

[2] En référence à un poème de Jacques PRÉVERT, le package R réalisé pour nos analyses se nommant prevR.

Afin de réaliser ces analyses et de les automatiser, nous avons programmé un package additionnel, *prevR*, au logiciel de statistiques R (section 4.4).

Ce pays modèle nous permettra de choisir une technique d'interpolation spatiale, le krigeage ordinaire, qui s'avérera adaptée à notre problématique (section 4.5). Nous inspirant des techniques d'analyse en composantes d'échelle (section 4.6), nous testerons une technique reposant sur des cercles de même rayon pour estimer les prévalences de chaque grappe (section 4.7). Cependant, en raison d'une dispersion très inégale des grappes sur le territoire, les cercles de même rayon s'avèrent inadaptés. Nous aurons alors recours à un lissage adaptatif à l'aide de cercles de même effectif (section 4.8). Si cette approche permet de réduire les aléas statistiques dus à l'échantillonnage, la difficulté réside dans la détermination de meilleure valeur du paramètre N, à savoir l'effectif minimum des cercles de lissages. Le pays modèle permettra de déterminer les valeurs optimums de ce paramètre en fonction des caractéristiques de chaque enquête (section 4.9). Nous verrons alors qu'elles peuvent être modélisées (section 4.10).

Nous avons également pris en compte le milieu de résidence afin de tenir compte des différentiels marqués entre milieu urbain et milieu rural (section 4.11). Dans les zones faiblement peuplées, la méthode des cercles de même effectif induit un lissage à partir de cercles de très grand rayon. Nous avons donc réintégré le paramètre R, correspondant à un rayon maximum des cercles de lissage (section 4.12). Enfin, nous avons élaboré un indicateur de qualité afin de préciser la fiabilité des données estimées (section 4.13). Nous verrons alors comment l'épidémie de notre pays modèle a pu être reconstituée par cette approche (section 4.14) et les résultats de l'application de cette technique aux données du Burkina Faso et du Cameroun (4.15).

Livia MONTANA et ses collaborateurs ont réalisé des travaux cartographiques à partir de l'EDS 2003 du Kenya en utilisant une autre approche que la notre (MONTANA 2007). Nous pourrons comparer ainsi leurs résultats avec les nôtres (section 4.16).

Enfin, nous aborderons quelques pistes de discussion et de développements futurs envisageables à partir de cette approche (section 4.17).

4.1 Interpolation spatiale et prévalence localisée

L'interpolation spatiale est un traitement mathématique de données référencées sur un territoire visant à estimer la valeur d'un attribut pour des sites non échantillonnés, à partir des valeurs des sites échantillonnés, à l'intérieur des limites définies par les positions des sites échantillonnés. L'extrapolation spatiale correspond au même processus mais appliqué à des sites non échantillonnés situés à l'extérieur des limites définies par les positions des sites échantillonnés. L'interpolation spatiale repose sur une hypothèse anticipatrice fondamentale : l'autocorrélation spatiale. Selon ce principe, des objets proches dans l'espace ont tendance à posséder des caractéristiques similaires. Les méthodes d'interpolation spatiale continue présupposent que les variations d'un phénomène se produisent de manière graduelle dans l'espace pouvant être estimée par un modèle mathématique.

Plusieurs techniques d'interpolation spatiale continue existent : surfaces de tendance, moyennes mobiles pondérées par l'inverse de la distance élevée à la puissance, krigeage… Nous les aborderons un peu plus tard. Cependant, elles ont toutes en commun de fournir un procédé mathématique permettant de calculer une surface à partir d'informations ponctuelles. Par exemple, à partir des relevés de pression atmosphérique de différentes stations météorologiques, elles vont permettre de calculer la carte des pressions atmosphériques et les lignes isobares[3]. Si nous concevons aisément que la pression atmosphérique puisse être considérée comme une mesure ponctuelle, quel sens cela peut-il avoir concernant la prévalence du VIH ?

Le concept opératoire de prévalence du VIH s'applique à une population d'individus : il s'agit de la proportion au sein d'une population donnée d'individus infectés par le VIH. Il ne s'applique donc pas directement à un territoire. Pour associer une prévalence à un territoire géographique, nous attribuons à ce dernier la prévalence calculée parmi les individus vivant sur ce territoire. Or, les êtres humains sont mobiles. Dans le cadre des EDS, l'association entre individus et territoire géographique s'effectue en fonction du lieu d'habitation. Pour que la notion de prévalence ait un sens, il est nécessaire que la population soit suffisamment importante. En effet, calculer une prévalence sur une ou dix personnes ne nous apporte que peu d'informations et, en tout cas, une information de nature différente qu'une prévalence applicable à plusieurs centaines ou plusieurs milliers de personnes.

[3] Il s'agit des courbes pour lesquelles la pression atmosphérique est constante.

La prévalence s'appliquant à une population suffisamment grande, nous ne pouvons la relier géographique qu'à une superficie elle-aussi suffisamment grande pour contenir cette population. Lorsque nous réduisons la taille de la zone géographique que nous voulons étudier, nous diminuons du même coup le nombre d'individus situés sur celle-ci. Une fois cette zone géographique réduite à la taille de point, le concept de prévalence n'est plus applicable car nous ne pouvons associer un individu à un simple point. En l'absence d'une population, il devient impossible de calculer une prévalence.

En physique, un problème plus ou moins similaire s'est posé dans la détermination d'une vitesse instantanée. En effet, la vitesse est un concept opératoire défini comme le rapport entre une distance et une durée. Elle s'applique donc à une période de temps donnée, sur une certaine trajectoire. En un point de l'espace à un instant donné, la vitesse ne peut plus être calculée car la durée est devenue égale à zéro. Cependant, lorsque l'on réduit progressivement le laps de temps considéré autour de l'instant qui nous préoccupe, les vitesses successives que l'on calcule tendent vers une valeur unique limite. En développant le concept mathématique de dérivée, il a été possible de définir une vitesse instantanée comme la dérivée de la position selon le temps.

Est-il possible de procéder de la même manière ? Pas tout à fait dans la mesure où, si distance et temps sont des concepts continus, individus et personnes infectées sont de nature discrète. À mesure que l'on réduit l'espace géographique, arrive un moment où il n'y aura plus qu'un individu présent et qui sera soit infecté soit non infecté par le VIH. En réduisant encore notre superficie, il n'y aura plus personne. Nous ne pouvons donc définir une prévalence localisée en un point comme étant la dérivée de la prévalence par la superficie.

Cependant, si nous travaillons à une échelle nationale, nous pouvons contourner la définition mathématique d'un point en tant qu'objet sans dimension et la remplacer par la notion de *pixel*. Ce concept désigne en informatique l'unité de base d'une image numérique. Une image est découpée en plusieurs petits carrés auxquels est associée une couleur. Ainsi, bien que visuellement un pixel soit assimilable un point, il possède une superficie. De même, dans le cadre d'une cartographie réalisée à une échelle nationale, de petites zones géographiques pourront être assimilées à un pixel. Bien que modélisées mathématiquement comme un point, elles auront une superficie et donc il sera possible de leur associer une population et une prévalence. En procédant de la sorte, le concept de prévalence localisée devient opératoire.

Dans le cadre des EDS, nous ne disposons pour chaque zone d'enquête que d'une seule coordonnée géographique correspondant au centre de la grappe en question. Comme il s'agit de zones de dénombrement, c'est-à-dire d'un découpage fin du territoire, nous pouvons sans problème les assimiler mathématiquement à un point lorsque nous travaillons à une échelle nationale.

Depuis la mise en place de tests de dépistage du VIH, Measure DHS a mis en place une procédure pour garantir l'anonymat des personnes enquêtées. Ainsi, les coordonnées géographiques des grappes enquêtées sont décalées aléatoirement dans un rayon de deux kilomètres en milieu urbain et de cinq kilomètres en milieu rural (MEASURE DHS 2006). Cette imprécision constitue un biais mineur pour une analyse à l'échelle nationale. Cependant, elle interdit de retrouver avec précision les grappes effectivement enquêtées et donc de pouvoir identifier les personnes interviewées par recoupement avec les données des questionnaires individuels.

4.2 Les données de départ

Pour chaque enquête avec mesure de la prévalence du VIH et des coordonnées des zones d'enquêtes, Measure DHS fournit un fichier au format SPSS contenant pour chaque individu testé le résultat du test VIH, l'âge et le sexe de la personne, un identifiant contenant le numéro de la grappe d'appartenance de l'individu, ainsi qu'une variable de pondération. Le second fichier, au format *dbf* le plus souvent, indique les coordonnées (longitude et latitude) de chaque grappe, leur milieu de résidence et leur région d'appartenance.

Nous restreignons ici notre analyse aux individus âgés de 15 à 49 ans révolus, disposant d'un résultat positif ou négatif (les résultats indéterminés ont été retirés du fichier) et d'un taux de pondération non nul. Ainsi, dans l'EDS 2004 du Cameroun, 9 900 personnes ont été testées, réparties en 466 grappes. Dans l'EDS 2003 du Burkina Faso, il s'agissait de 7 244 individus répartis en 400 grappes. Le Tableau 4.1, repris à partir du Tableau 3.5, rappelle l'échantillonnage de 17 enquêtes récentes[4].

Pour chaque zone d'enquête, nous avons calculé le nombre d'individus testés, la somme des poids individuels de ceux-ci, ainsi que la prévalence observée au niveau de la grappe (en tenant compte des taux de pondération individuels).

[4] Les écarts, pour le Burkina Faso, concernant le nombre de personnes proviennent du fait que le Tableau 4.1 est construit à partir des rapports finaux de chaque enquête tandis que le chiffre de 7 244 correspond au nombre d'individus effectivement présents dans la base de données, sans avoir recours aux taux de pondération.

Tableau 4.1
Échantillonnage de 17 enquêtes nationales récentes (EDS et apparentées)

Pays	Année	Type	Grappes	Ménages éligibles*	Personnes testées 15-49 ans	Nombre moyen de pers. testées par grappe	Prévalence du VIH (%) 15-49 ans	IC 95 %
Burkina Faso	2003	EDS	400	3179	7151	17,9	1,8	1,5 - 2,1
Cameroun†	2004	EDS	466	5319	9900	21,2	5,5	5,1 - 6,0
Côte d'Ivoire‡	2005	EIS	249	4368	8436	33,9	4,7	4,3 - 5,2
Éthiopie†	2005	EDS	540	6689	10540	19,5	1,4	1,2 - 1,6
Ghana	2003	EDS	412	6251	9144	22,2	2,2	1,9 - 2,5
Guinée†	2005	EDS	297	3126	6388	21,5	1,5	1,2 - 1,8
Kenya	2003	EDS	400	4234	6001	15,0	6,7	6,1 - 7,4
Lesotho	2004	EDS	405	4185	5043	12,5	23,5	22,3 – 24,7
Malawi†	2004	EDS	522	4580	5150	9,9	11,8	10,9 - 12,7
Mali§	2001	EDS	403	4087	6475	16,1	1,7	1,4 - 2,1
Niger†	2006	EDS	345	3815	7262	21,0	0,7	0,5 - 0,9
Ouganda#	2004	HSBS	417	9529	16906	40,5	6,4	6,0 - 6,8
Rwanda‡	2005	EDS	462	5136	10016	21,7	3,0	2,7 - 3,4
Sénégal‡	2005	EDS	377	2453	7503	19,9	0,7	0,5 - 0,9
Tanzanie	2003	AIS	345	6499	10747	31,2	7,0	6,8 - 7,2
Zambie§	2001-02	EDS	320	2368	3807	11,9	15,6	14,5 - 16,8
Zimbabwe†	2005-06	EDS	400	9285	12796	32,0	18,1	17,4 - 18,8

Sources: http://www.measuredhs.com et rapport final de chaque enquête..

EDS : Enquête démographique et de Santé ; AIS : AIDS Impact Survey ; HSBS : HIV/AIDS Sero-Behavioural Survey ; EIS : Enquête sur les Indicateurs du SIDA.

IC 95 % : Intervalle de Confiance à 95 %, calculé selon la méthode de Wilson avec correction de continuité.

* Il s'agit du nombre de ménages éligibles pour le dépistage du VIH et effectivement enquêtés.

† Coordonnées des zones d'enquêtes pas encore disponibles. Concernant le Cameroun, ces données étaient disponibles début 2006 avant d'être retirées du téléchargement.

‡ Coordonnées des zones d'enquêtes non collectées.

§ Les résultats du dépistage du VIH ne peuvent être liés aux coordonnées des zones d'enquêtes.

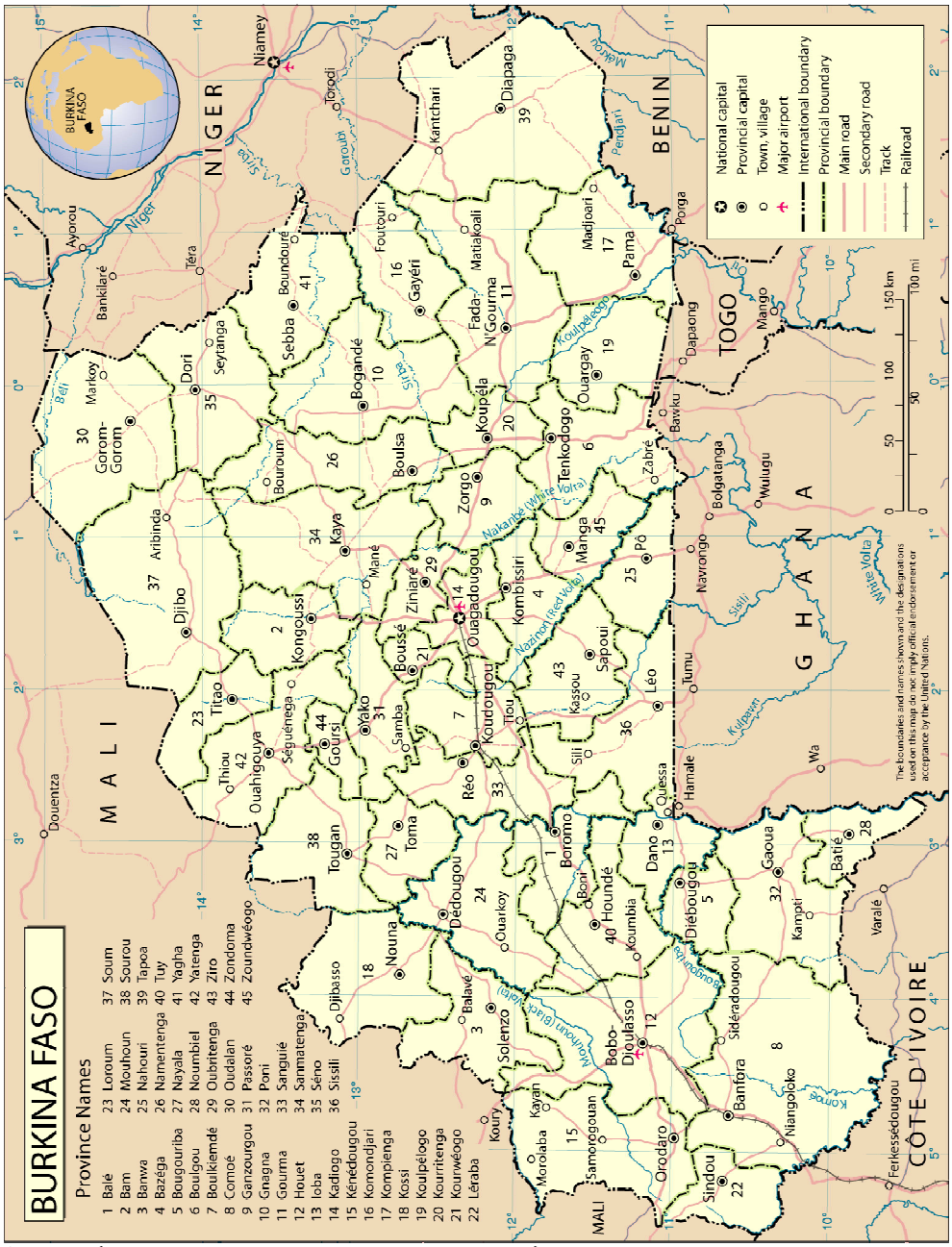
Données en accès restreint.

Pour pouvoir procéder à un travail cartographique, il importe de disposer des frontières du pays, sous une forme géoréférencée, afin de pouvoir positionner correctement les observations. Nous avons eu recours aux données du *Digital Chart of the World* (DCW)[5]. Le DCW a été développée par la société ESRI, à partir d'un travail original de la US Defense Mapping Agency (DMA), en 1991-1992. Les frontières qu'elle fournit reflètent donc la situation à cette date. Cependant, pour les pays que nous étudions, ces frontières nationales n'ont pas évolué depuis.

5 Téléchargeables sur http://www.maproom.psu.edu/dcw/.

Figure 4.1

Carte du Burkina Faso (régions, villes, routes)



Source : (UNITED NATIONS CARTOGRAPHIC SECTION 2004c).

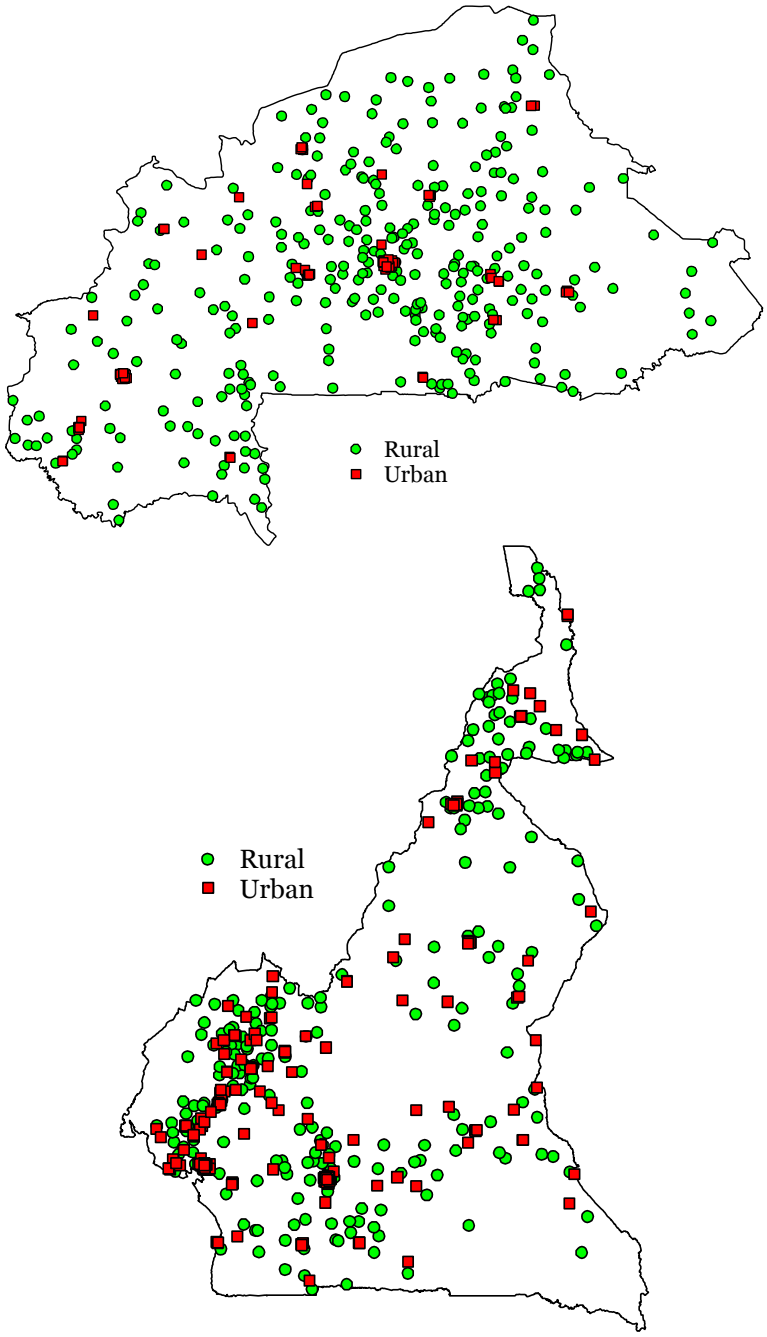
La Figure 4.1 et la Figure 4.2 permettent de situer chacun des deux pays étudiés (Burkina Faso et Cameroun). Y sont représentées les différentes régions, les principales villes et les principales routes.

Figure 4.2
Carte du Cameroun (régions, villes, routes)

Source : (UNITED NATIONS CARTOGRAPHIC SECTION 2004b).

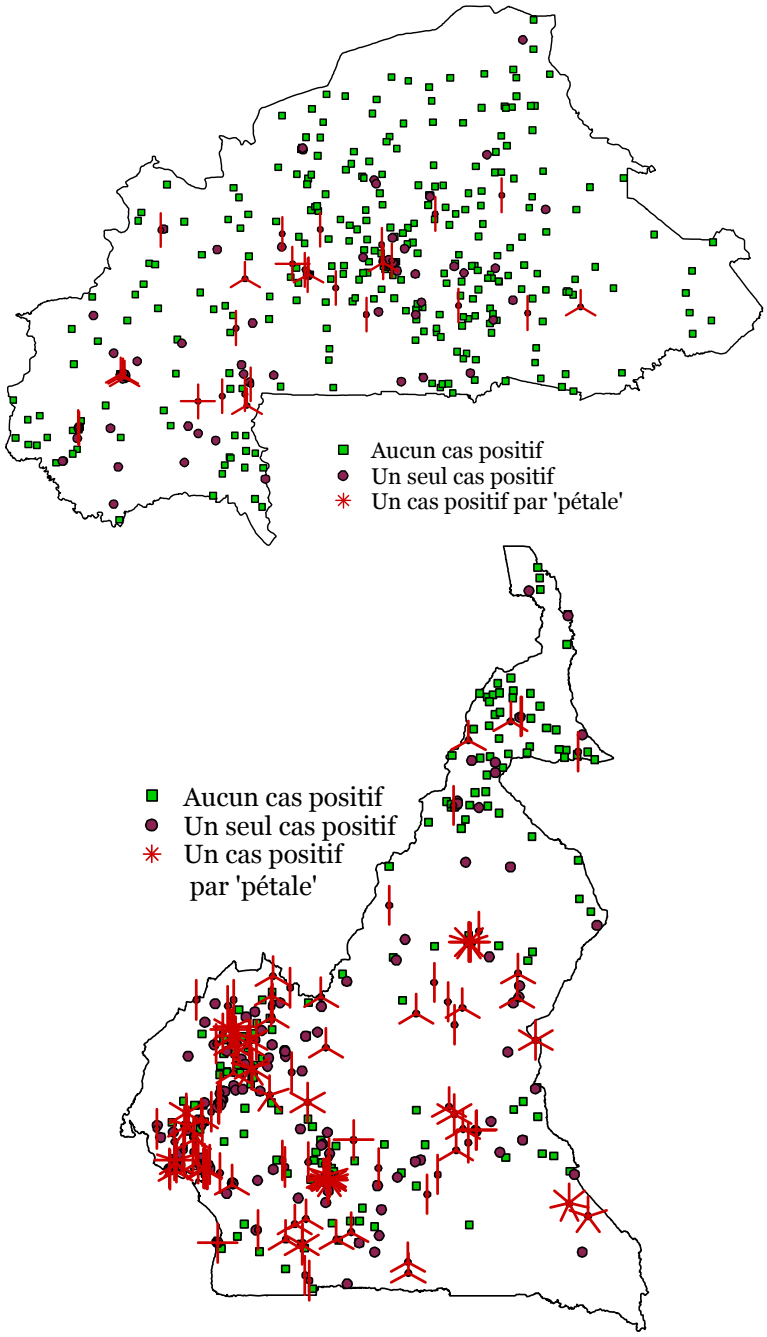
La Figure 4.3 permet de localiser les 400 grappes de l'EDS 2003 du Burkina Faso et les 466 grappes de l'EDS 2004 du Cameroun, selon leur milieu de résidence.

Figure 4.3
Localisation des grappes de l'EDS 2003 du Burkina Faso et de l'EDS 2004 du Cameroun, selon le milieu de résidence



Les différentes grappes ne sont pas localisées uniformément sur le territoire mais sont concentrées dans les zones les plus peuplées. Ainsi, certaines régions ont été enquêtées avec précision tandis que pour d'autres, seules quelques grappes isolées ont été inclues dans l'enquête.

Figure 4.4
Nombre d'infections à VIH observées, par grappe, pour l'EDS 2003 du Burkina Faso et l'EDS 2004 du Cameroun



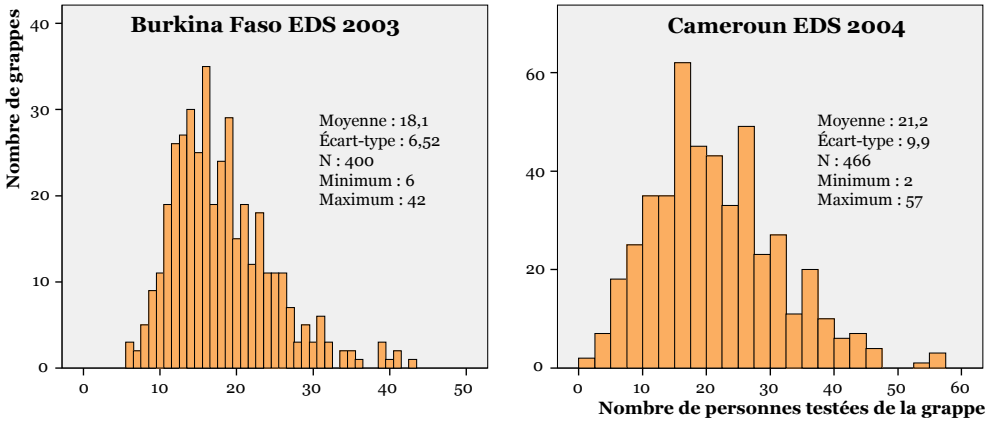
Sur la Figure 4.4, nous avons représenté, pour chaque grappe, le nombre d'individus dépistés positifs. En raison du faible nombre d'individus testés par grappe (18 en moyenne pour le Burkina Faso et 21 pour le Cameroun), dans de nombreuses grappes (la majorité d'ailleurs au Burkina Faso), aucune séropositivité n'a été observée (points verts). Cela ne signifie pas pour autant que l'épidémie de

VIH n'a pas atteint ces zones. Simplement, du fait des faibles effectifs de personnes testées et du niveau local de la prévalence, aucun individu séropositif au VIH n'a été sélectionné. Cela est plus fréquent au Burkina Faso puisque la prévalence nationale du VIH y est plus faible (1,8 % contre 5,5).

La Figure 4.5 met en évidence la variabilité du nombre de personnes testées par grappe[6]. Cette variabilité est due en grande partie à des effets aléatoires lors de l'échantillonnage. Le nombre de ménages éligibles au test de dépistage par grappe est faible (7,9 en moyenne au Burkina Faso et 11,4 au Cameroun). Suivant les grappes, le tirage au sort des ménages favorisera tantôt des ménages nombreux et tantôt des ménages réduits. Cette variation du nombre de personnes testées par grappe ne présente pas de pattern spatial particulier.

Figure 4.5

Distribution des grappes selon le nombre de personnes testées



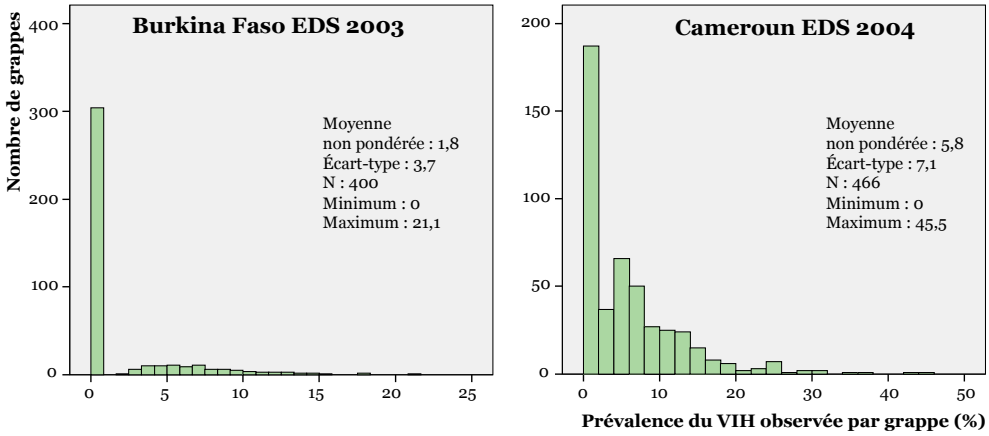
Nous avons calculé la statistique de GEARY (GEARY 1954, AUBRY 2001) pour déterminer si le nombre de personnes testées par grappe était autocorrélé spatialement[7], à partir des données des dernières EDS du Burkina Faso, du Cameroun, du Ghana, du Kenya et de la Tanzanie et du logiciel *CrimeStat® III*[8] (LEVINE 2004). À l'exception du Ghana, le nombre de personnes testées par grappe n'était significativement autocorrélé spatialement.

[6] Nous ne présentons pas ici la cartographie du nombre de personnes testées par grappe. Celle-ci sera abordée à la section 4.12 page 247 (voir également la Figure 4.36 page 258 et la Figure 4.42 page 264).

[7] « *L'autocorrélation spatiale peut être définie comme la ressemblance des valeurs prises par une variable, exprimée en fonction de leur localisation géographique. L'analyse de l'autocorrélation spatiale permet de quantifier la régularité d'un phénomène (une forme de complexité spatiale) et de déterminer la portée de la dépendance spatiale.* » AUBRY P. et PIÉGAY H., « Pratique de l'analyse de l'autocorrélation spatiale en géomorphologie : définitions opératoires et tests », *Géographie physique et Quaternaire*, n°55(2), 2001.

[8] Disponible sur http://www.icpsr.umich.edu/CRIMESTAT/.

Figure 4.6
Distribution des grappes selon la prévalence du VIH observée par grappe



La Figure 4.6 montre la très grande variation des prévalences du VIH calculées par grappe. Nous retrouvons le fait qu'un nombre élevé de grappe n'a pas enregistré de séropositivité. Par ailleurs, les prévalences observées sont très étalées et les maximum élevés. L'écart-type s'avère supérieur à la moyenne, traduisant ainsi que les prévalences calculées par grappe reflètent plus les variations aléatoires dues à l'échantillonnage plutôt que les variations réelles des épidémies. C'est la résultante des faibles effectifs du nombre de personnes testées par grappe.

Chaque grappe est un petit sous-échantillon. Le statut sérologique des individus testés constitue une information fiable. Mais la prévalence observée d'une grappe, en tant que proportion calculée sur un sous-échantillon, ne constitue qu'une estimation de la prévalence réelle de la grappe. La prévalence observée est donc égale à la prévalence réelle plus une erreur aléatoire liée au tirage au second degré dans l'échantillonnage des EDS (Équation 4.1).

Équation 4.1
Lien entre prévalence observée et prévalence réelle

$$P_{obs}(x,y) = P_{grappe}(x,y) + EA$$

P_{obs} : *prévalence observée par grappe*

P_{grappe} : *prévalence réelle du VIH des grappes*

EA : *erreur aléatoire*

Dans le cadre d'une proportion, cette erreur aléatoire suit en première approximation une loi normale d'espérance nulle et de variance $\sqrt{p(1-p)/N}$ en notant p la proportion étudiée et N l'effectif de l'échantillon[9]. Cette erreur aléatoire

[9] Voir l'annexe 3 pour plus de détails, méthode de WALD sans correction de continuité.

est donc d'autant plus importante que la proportion étudiée est proche de zéro et que l'effectif de l'échantillon est faible. Or, au niveau d'une grappe, nous avons vu que seules 10 à 30 personnes étaient testées. Avec un effectif aussi faible, l'erreur aléatoire peut devenir tellement importante au niveau d'une grappe qu'il en devient impossible de conclure sur le sens de la prévalence observée. Si nous observons un cas positif sur 20 individus (prévalence observée de 5 %), la prévalence réelle de la grappe se situe dans un intervalle de confiance à 95 % compris entre 0 et 27 % !

De fait, les prévalences du VIH observées par grappe sont d'une extrême imprécision et ne traduisent pas la valeur réelle des prévalences localisées. Il n'est pas possible de procéder directement à une interpolation spatiale puisque ces techniques nécessitent de disposer, pour chaque point connu, d'une valeur de qualité. Il s'avère donc nécessaire, avant de procéder à une interpolation spatiale, d'estimer la valeur des prévalences localisées de chaque grappe.

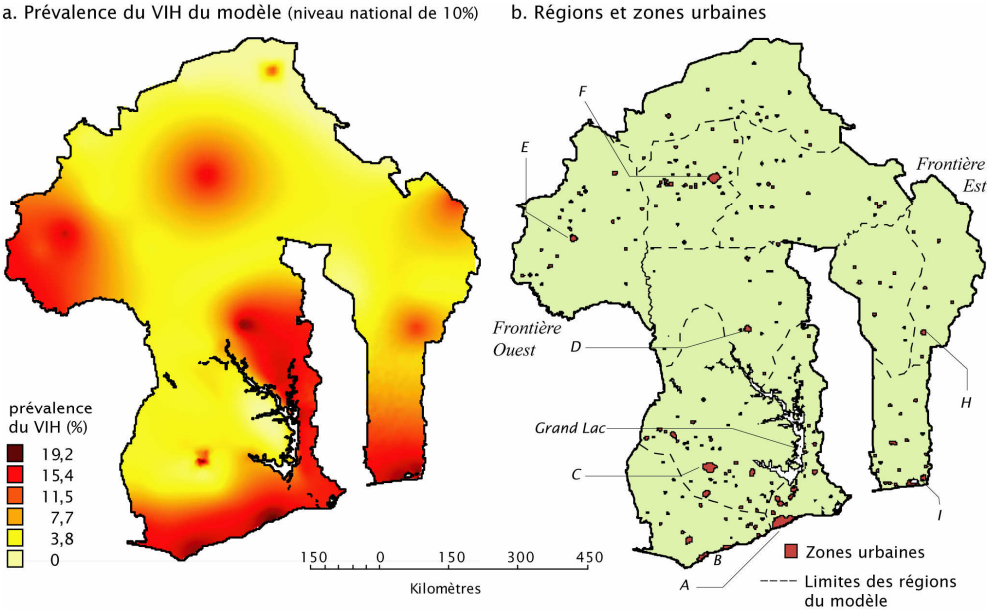
4.3 Création d'un pays modèle et simulation d'EDS

Aucun pays ne dispose d'une enquête à grande échelle avec une mesure fine des prévalences locales. En l'absence de données de référence pour tester et valider une approche méthodologique, nous pouvons avoir recours à un pays modèle à partir duquel nous simulerons la réalisation d'Enquêtes Démographiques et de Santé. Il sera alors possible de comparer les résultats d'enquête à l'épidémie de départ du modèle mais également de voir les différences entre les variations spatiales de l'épidémie de départ et celles que nous aurons reconstituées.

Pour élaborer ce pays modèle, que nous avons baptisé Alicante, nous avons décidé de partir de données réelles afin que la répartition de la population au sein d'Alicante présente un pattern plausible. Pour cela, nous avons agrégé le Bénin, le Burkina Faso et le Ghana. Nous avons volontairement exclu le Togo d'Alicante afin d'obtenir une forme concave qui complexifie l'interpolation spatiale puisque la zone située à l'intérieure de la concavité n'est pas enquêtée. Une telle forme concave se retrouve dans les frontières de certains pays, tels que le Sénégal.

Nous avons alors utilisé des données du *Global Rural-Urban Mapping Project (GRUMP)* que nous avons appliquées à Alicante : densités de population en l'an 2000[10] avec une résolution de 30 secondes d'arc et découpage du territoire en milieu urbain et milieu rural[11]. Nous avons ensuite découpé le territoire d'Alicante en 9 137 grappes (7 818 grappes rurales et 1 319 grappes urbaines) avec une résolution moyenne de 2 minutes d'arc en milieu urbain et 5 minutes d'arc en milieu rural. Nous avons ensuite calculé la superficie de chaque grappe puis la population de chaque grappe en multipliant la superficie de la grappe par sa densité moyenne. Alicante a ensuite été divisée en 11 régions. Les principaux centres urbains ont été renommés par une lettre, allant de A à I (voir Figure 4.7).

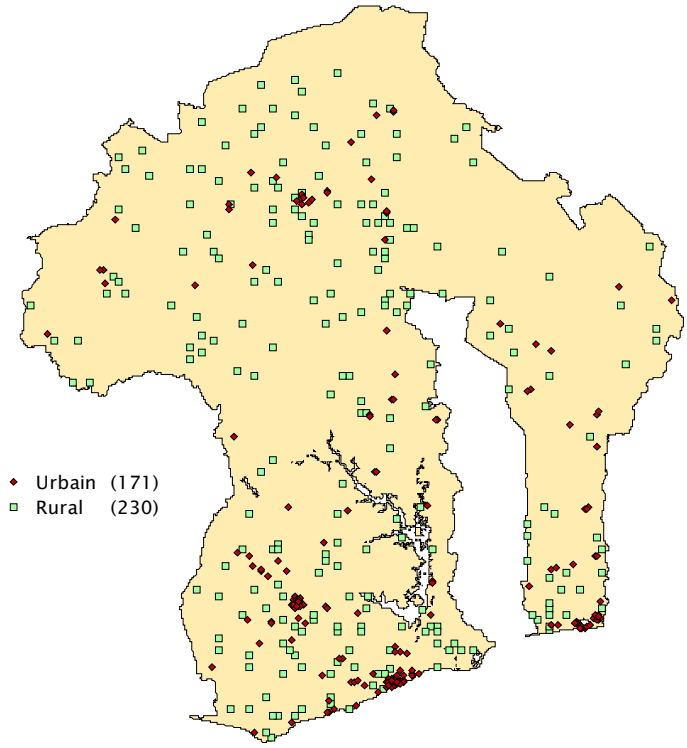
Figure 4.7
Épidémie de VIH, régions et zones urbaines d'Alicante



[10] CENTER FOR INTERNATIONAL EARTH SCIENCE INFORMATION NETWORK (CIESIN) OF COLUMBIA UNIVERSITY, INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE (IFPRI) *et. al.*, *Global Rural-Urban Mapping Project (GRUMP), Alpha Version: Population Density Grids*, Palisades, New York (US), Socioeconomic Data and Applications Center (SEDAC) of Columbia University, 2004c. (http://sedac.ciesin.columbia.edu/gpw)

[11] CENTER FOR INTERNATIONAL EARTH SCIENCE INFORMATION NETWORK (CIESIN) OF COLUMBIA UNIVERSITY, INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE (IFPRI) *et. al.*, *Global Rural-Urban Mapping Project (GRUMP), Alpha Version: Urban Extents*, Palisades, New York (US), Socioeconomic Data and Applications Center (SEDAC) of Columbia University, 2004b. (http://sedac.ciesin.columbia.edu/gpw)

Figure 4.8
Répartition des grappes d'une simulation d'une EDS



Paramètres de la simulation : prévalence nationale de 10 %, 8 000 personnes, 401 grappes.

Dans un second temps, une épidémie fictive a été appliquée au modèle (Figure 4.7.a), totalement indépendante de la situation des trois pays qui ont servi de base à la construction du modèle. Cette épidémie, correspondant à une prévalence nationale de 10%, a été conçue de manière à présenter différents modèles de diffusion : ville importante avec une épidémie concentrée sur celle-ci et faible prévalence autour (C) ; ville importante (F) et moyenne (H) avec diffusion progressive ; pic localisé en zone rurale dans le nord du pays ; rupture de continuité de part et d'autre d'un grand lac ; gradient des côtes vers l'intérieur des terres au sud du pays, avec deux grandes agglomérations (A et I) et deux villes moyennes (B) ; diffusion depuis la frontière ouest et depuis une ville située de l'autre côté d'une frontière (à l'est).

Pour la suite des analyses, il sera possible de faire varier la prévalence nationale en multipliant la prévalence de chaque grappe par un même facteur d'échelle. Nous pourrons ainsi simuler des EDS dans le cadre d'épidémies présentant une prévalence nationale de 5 % ou de 30 %.

La simulation des Enquêtes Démographiques et de Santé est ensuite réalisée à partir de trois paramètres : la prévalence nationale, le nombre total de personnes enquêtées et le nombre de grappes au premier degré. Elle se déroule en trois étapes. Dans un premier temps, chacune des onze régions est divisée en deux

strates, l'une urbaine et l'autre rurale. Le nombre de grappes tirées par strate est proportionnel à la population totale de celle-ci. Les grappes sont tirées aléatoirement, strate par strate, avec une probabilité de sondage proportionnelle à leur effectif de population.

Dans un second temps, l'effectif de personnes testées par grappe est déterminé aléatoirement, selon une loi normale, afin de reproduire la variabilité du nombre de personnes enquêtées par grappe, variabilité que l'on peut observer dans les EDS[12]. Le nombre de personnes enquêtées par grappe est ensuite redressé pour que le total corresponde à l'effectif visé. Enfin, dans un dernier temps, le nombre de personnes présentant la caractéristique étudiée dans chaque grappe est déterminé aléatoirement selon une loi binomiale ayant pour paramètre le nombre de personnes enquêtées et la prévalence de cette grappe dans le modèle.

Un facteur de pondération W_{hi} est calculé pour les individus de la grappe i de la strate h selon la formule suivante, comparable à celle utilisée dans les EDS :

Équation 4.2

Calcul des taux de pondération W_{hi} dans le cadre de la simulation d'EDS

$$W_{hi} = \frac{1}{P_{1hi} P_{2hi}}$$

$$P_{1hi} = \frac{a_h \cdot pop_{hi}}{POP_h}$$

$$P_{2hi} = \frac{n_{hi}}{pop_{hi}}$$

P_{1hi} représente la probabilité que la grappe hi soit incluse dans l'échantillon, P_{2hi} celle qu'un individu de la grappe hi soit enquêté, a_h le nombre de grappes sélectionnées dans la strate h, pop_{hi} l'effectif de population de la grappe hi, POP_h l'effectif total de la strate h et n_{hi} le nombre de personnes enquêtées dans la grappe hi. Ces pondérations sont ensuite corrigées par un facteur d'échelle afin que leur total corresponde à l'effectif total de l'enquête.

[12] Nous rappelons que le nombre de personnes testées par grappe n'est pas autocorrélé spatialement (voir section 4.2).

La Figure 4.8 page 211 présente la répartition des grappes enquêtées pour une simulation d'une EDS avec une prévalence nationale de 10%, 8 000 personnes enquêtées et 400 grappes[13]. Les détails de cette simulation sont fournit dans le package *prevR* (voir section suivante, le CD-Rom annexe et l'annexe 5 page 3).

Pour plus de détails sur la simulation des EDS, voir la fonction *tirage* en annexe (Annexe 7, section 7.3).

Les données générées par simulation sont bien de nature comparable aux données réelles des Enquêtes Démographiques et de Santé. La distribution des prévalences observées générées par la simulation de la Figure 4.8 présente le même profil[14] que celles des EDS du Burkina Faso et du Cameroun (Figure 4.6). Par ailleurs, l'écart type des prévalences par grappe est également très élevé, traduisant les variations aléatoires dues à l'échantillonnage.

4.4 Automatisation des analyses sous R : le package prevR

Les analyses réalisées dans ce chapitre ont été effectuées sous le logiciel de statistiques R (R DEVELOPMENT CORE TEAM 2006). Outre sa puissance de calcul, R présente l'avantage d'être distribué[15] gratuitement sous la licence libre GNU GPL de la Free Software Foundation. Il s'agit par ailleurs d'un projet viable supporté par de nombreuses universités à travers le monde. R fonctionne sous de multiples plateformes et notamment sous Windows, MacOs ou encore Linux.

Nous avons développé plusieurs fonctions, les plus génériques possibles, afin de pouvoir reproduire facilement les différentes analyses menées ici. Notre objectif consistait à pouvoir fournir un outil à la disposition des différents programmes nationaux de lutte contre le SIDA. Ces fonctions sont distribuées gratuitement, sous la licence libre CeCILL-C[16], à partir du site du CEPED[17], sous la forme d'un package additionnel à R que nous avons nommé *prevR*.

[13] Le nombre total de grappes s'avère être de 401 en raison d'arrondis dans le calcul du nombre de grappes par strate.

[14] Graphique non reproduit.

[15] http://www.r-project.org.

[16] Cette licence libre de droit français a été développée par le CEA, le CNRS et l'INRIA (http://www.cecill.info/).

[17] http://www.ceped.cirad.fr/prevR.

prevR est bilingue, français/anglais, bien que la documentation ne soit pour le moment disponible qu'en français. Deux forums de discussion (l'un en français et l'autre en anglais) sont par ailleurs à la disposition des utilisateurs sur notre site personnel[18].

prevR a fait l'objet d'un dépôt auprès de l'Agence de Protection des Programmes (APP) par l'IRD le 22 mars 2007 (Figure 4.9) après que l'ANRS ait renoncé à tout droit sur l'œuvre en question.

Il permet l'import des données (résultats des tests VIH, coordonnées des zones d'enquêtes, frontières nationales du DCW, localisation des villes du GRUMP) directement à partir des fichiers téléchargeables en ligne et leur mise en forme pour exploitation. Plusieurs cartes descriptives peuvent être produites. prevR permet d'utiliser deux techniques d'interpolation spatiale : le krigeage et l'interpolation selon l'inverse de la distance. Pour ces deux techniques, nous avons eu recours aux fonctions du package *gstat* (PEBESMA 2004).

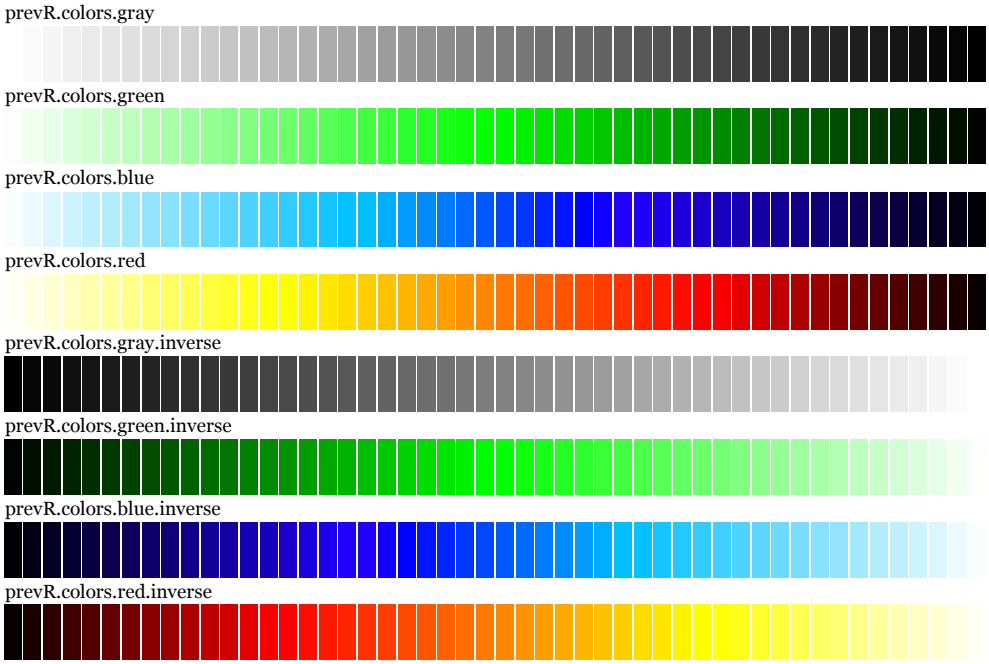
Les résultats peuvent ensuite être exportés dans différents formats standards vers des logiciels de statistiques ou des logiciels de cartographie (SIG).

Figure 4.9
Certificat de dépôt de prevR auprès de l'APP


Inter Deposit Digital Number
certificat délivré par
Agence pour la Protection des Programmes
249, rue de Crimée - 75019 PARIS. Tél 33 (1) 40 35 03 03. Fax 33 (1) 40 38 96 43
IDDN.FR.001.120030.000.S.P.2007.000.31235
(1) (2) (3) (4) (5) (6) (7) (8) (9) (10)
Pour l'oeuvre : prevR
Identité du(des) titulaire(s) des droits :
IRD - Institut de Recherche
pour le Développement
213 rue La Fayette
75480 PARIS Cx 10
Siret : 180006025
INTERDEPOSIT CERTIFICATION
Siret :
Siret :
Adhérent sous le numéro : 89.75.879
Support utilisé : 1 CD-Rom en double exemplaire
Logibox conservé par l'adhérent : 50777
Logibox conservé par l'APP : 50778
Fait à Paris, le 22/03/07
Le titulaire *
Le Président
(1) Inter Deposit Digital Number
(2) Nationalité de l'oeuvre
(3) Numéro de l'organisme d'enregistreme
(4) Numéro d'ordre de l'enregistrement
(5) Numéro de version
(6) Type d'enregistrement
(7) Type de l'oeuvre
(8) Année d'enregistremen
(9) Zone réservée (Clé d'intégrité
(10) Classe de produit
* Le titulaire s'engage à informer l'APP de toute cession ou aliénation, totale ou partielle, de ses droits de propriété intellectuelle
Seules les inscriptions de type S et C permettent un éventuel accès au programme source

[18] http://joseph.larmarange.net/prevR.

Figure 4.10
Palettes de couleur prevR.colors



prevR est joint à la thèse sur le CD-rom annexe. Par ailleurs, le code des différentes fonctions implémentées est publié en annexe 4. L'annexe 5 fournit une description détaillée de chaque fonction. Enfin, un guide d'utilisation ou tutoriel est reproduit en annexe 6.

Afin d'améliorer la lecture des cartes produites avec prevR, quatre palettes de couleurs ont été créées et implémentées dans prevR (Figure 4.10). Outre une évolution progressive et continue d'une teinte vers une autre, les couleurs extrêmes ont été éclaircies ou assombries afin d'augmenter les contrastes. Les fonctions ont été écrites de manière à pouvoir générer des palettes de quelques couleurs à plusieurs centaines.

4.5 Choix d'une méthode d'interpolation spatiale : le krigeage ordinaire

Nous proposons une méthodologie en deux étapes. Tout d'abord une estimation de la prévalence de chaque zone d'enquête, puis une interpolation spatiale de ces prévalences estimées. L'objectif de l'estimation des prévalences localisées de chaque zone d'enquête vise à réduire les aléas statistiques du tirage au second degré des EDS qui, du fait des faibles effectifs de personnes testées par grappe, induit des variations importantes de la prévalence observée dans les zones

d'enquêtes. L'interpolation spatiale, quant à elle, doit permettre de compenser le tirage au premier degré puisque seule une petite partie des zones de dénombrement du pays sont effectivement enquêtées dans les EDS.

Avant de procéder à une estimation de la prévalence localisée de chaque grappe, nous devons vérifier si, à partir de 400 à 500 points seulement, il est possible de reconstruire une épidémie nationale. Pour cela, nous allons avoir recours à Alicante, notre pays modèle. Nous appelons *prévalence réelle* la prévalence attribuée, dans le modèle, à chaque grappe et *prévalence observée* la prévalence calculée, pour chaque grappe, à partir des données d'enquête, c'est-à-dire à partir des résultats des tests VIH des individus sélectionnés et testés dans la grappe. Enfin, nous parlerons dans la suite de chapitre de *prévalence estimée* pour désigner la prévalence localisée de chaque grappe estimée à partir des résultats d'enquêtes de la grappe et de ses voisines.

Si nous prenons les résultats d'enquêtes obtenus à l'aide de la simulation présentée à la Figure 4.8, nous disposons, après le tirage au premier degré (sélection des grappes), de la prévalence réelle pour chaque grappe sélectionnée. Les prévalences sont connues puisqu'il s'agit du modèle (nous connaissons par définition la situation avant simulation d'une EDS). Nous pouvons alors vérifier si les techniques d'interpolation spatiale permettent de reconstruire l'épidémie de départ du modèle. Ce type de vérifications ne peut être effectué à partir de données empiriques puisque nous ne disposons alors que des résultats d'enquête. Les prévalences réelles et l'épidémie nationale sont de ce fait hors de notre portée et ce sont justement elles que nous cherchons à atteindre.

Encadré 4.1
Interpolation spatiale selon l'inverse de la distance

La moyenne selon une pondération inverse de la distance, connue en anglais sous le nom de *Inverse Distance Weighting* (IDW), est une technique permettant d'assigner une valeur en des points inconnus à partir de points connus. Une fonction de pondération simple selon l'inverse de la distance a été proposée par Donald SHEPARD (1968) :

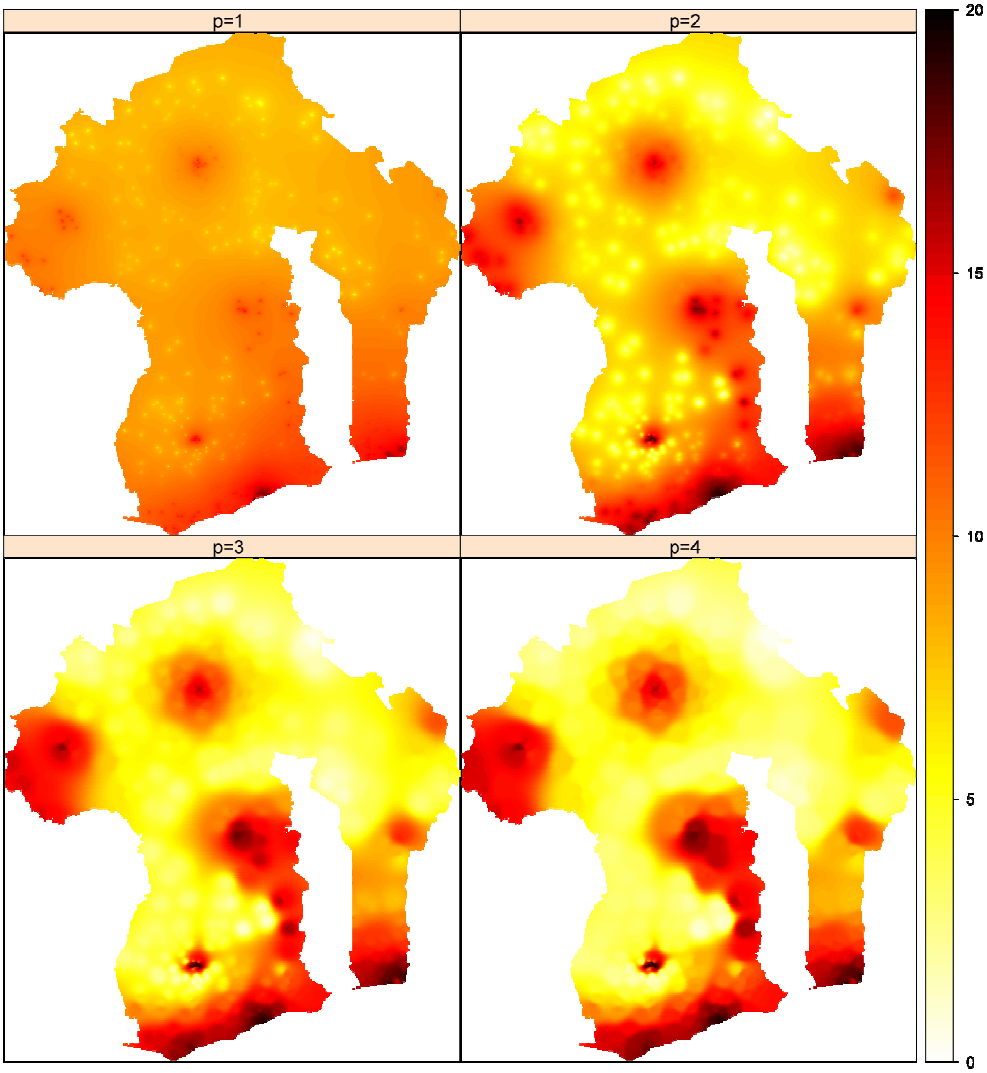
$$Z = \frac{\sum_{i=1}^{N} \frac{Z_i}{d_i^p}}{\sum_{i=1}^{N} \frac{1}{d_i^p}}$$

Z est la valeur du point à interpoler, Z_i les valeurs connues, N le nombre de points connus et d_i la distance entre les points Z et Z_i.

Plus la puissance appliquée à d_i est élevée, plus les observations proches auront de poids au détriment des observations plus éloignées. La valeur la plus commune de p est 2.

Nous allons aborder dans ce chapitre deux des techniques d'interpolation spatiale les plus courantes : l'interpolation par pondération selon l'inverse de la distance et le krigeage.

Figure 4.11
Interpolation selon l'inverse de la distance, pour différentes valeurs de la puissance appliquée à la distance, de la prévalence réelle des grappes sélectionnées lors d'une simulation d'une EDS



La première technique, la *pondération selon l'inverse de la distance* (Encadré 4.1), est relativement aisée à mettre en œuvre. Nous avons interpolé selon cette méthode la prévalence réelle des 401 grappes de notre simulation d'EDS avec différentes

valeurs de puissance appliquées à la distance (Figure 4.11)[19]. Ces cartes peuvent être comparées avec l'épidémie de départ appliquée à Alicante (Figure 4.7).

Pour *p* valant 1, les observations éloignées influent de manière trop importante sur l'interpolation spatiale et nous assistons à un effet d'uniformisation. Lorsque la distance est élevée au carré (puissance de 2), l'épidémie de départ commence à être reconstituée. Cependant, l'épidémie interpolée présente un aspect « granuleux » dépendant fortement de la position des grappes enquêtées.

Les résultats sont meilleurs avec une puissance de 3 ou 4. Les principales tendances de l'épidémie de départ sont effectivement reconstituées. Cependant, le recours à une interpolation selon l'inverse de la distance induit un effet de « vagues », particulièrement visibles autour de l'agglomération F[20], qui dépend essentiellement de la localisation des grappes sélectionnées plutôt du phénomène étudié.

Nous avons également interpolé la prévalence réelle des grappes sélectionnées dans notre simulation d'EDS selon la technique du krigeage ordinaire (voir Encadré 4.2). Les résultats ont été portés sur la Figure 4.12 ci-dessous. La carte obtenue est de meilleure qualité que celles générées par pondération selon l'inverse de la distance et l'épidémie reconstituée est plus proche de celle de départ (Figure 4.7 page 210). Cette technique sera donc privilégiée par la suite.

Les différents patterns épidémiques sont bien traduits à l'exception du pic épidémique local en zone rurale au Nord du pays. En effet, aucune grappe n'a été sélectionnée dans cette zone. De plus, quelque soit la technique utilisée, il est impossible de reconstituer des variations à des échelles inférieures à la densité des grappes sélectionnées, puisque toutes les méthodes d'interpolation présupposent une forme de continuité du phénomène étudié[21]. C'est d'ailleurs pour cette raison, inhérente à toute méthode d'interpolation, que les patterns de l'épidémie de départ de notre modèle, à l'exception du pic local en milieu rural, présentent des variations suffisamment importantes pour être couvertes par plusieurs grappes après le tirage au premier degré.

[19] Les cartes présentées ci-après ont été directement produites avec prevR. Aucune projection n'a été utilisée : nous avons simplement utilisé la longitude comme abscisse et la latitude comme ordonnée. Néanmoins, nous situant entre les tropiques, les déformations sont faibles. Ces cartes sont simplifiées puisqu'elles ne comportent pas d'échelle et que le lac situé à l'intérieur des terres n'est pas représenté. Seuls les résultats finaux feront l'objet de cartes plus précises après exportation dans un logiciel de cartographie et ajout d'informations supplémentaires (dessin des frontières, principales villes et routes, échelle, etc.).

[20] Voir Figure 4.7 pour les noms des agglomérations urbaines d'Alicante.

[21] Voir à ce sujet la première citation de Georges MATHERON de la section 2.8 page 95 et l'exemple de la Figure 2.2 page 96.

Encadré 4.2

Le krigeage, la méthode optimale d'interpolation spatiale

Le krigeage est la première méthode d'interpolation spatiale à avoir tenu compte de la structure de dépendance spatiale des données. Les travaux de l'ingénieur sud-africain KRIGE (1951) sont précurseurs de cette méthode. Cependant, nous devons le terme *krigeage* et le formalisme de cette approche à Georges MATHERON qui en a assuré le développement à l'École des Mines de Paris (MATHERON 1962, 1963b, 1963a).

La valeur de la variable régionalisée étudiée est prédite en un point inconnu par une combinaison linéaire des données ponctuelles connues. « *Les poids λ_i associés à chacune des valeurs régionalisées observées sont choisis de façon à obtenir une prévision non biaisée et de variance minimale. Ces poids dépendent de la localisation des observations et de leur structure de dépendance spatiale. En fait, le krigeage est le nom donné à la meilleure prévision linéaire sans biais, en anglais "best linear unbiased predictor" ou "BLUP", dans un cadre spatial.* » (BAILLARGEON 2005, p. 13)

Il existe trois types classiques de krigeage : le krigeage simple (variable stationnaire de moyenne connue), le krigeage ordinaire (variable stationnaire de moyenne inconnue) et le krigeage universel (variable non-stationnaire, nécessite d'en connaître la tendance). Nous aurons recours ici au krigeage ordinaire, le plus courant, qui ne nécessite ni de connaître la moyenne du phénomène étudié, ni de poser l'hypothèse qu'il soit une tendance particulière.

Le krigeage est une méthode d'interpolation très souple. Il peut être global (toutes les observations sont prises en compte) ou local (seules les observations proches entrent dans l'estimation d'un point inconnu), exact (les valeurs originales sont préservées) ou approximatif (les valeurs originales peuvent être modifiées avec prise en compte d'un résidu ou bruit non spatialement corrélé). Nous utiliserons pour notre part une approche exacte et globale. En effet, dans le cadre d'une approche locale, il est nécessaire de préciser au modèle un rayon de sélection des observations à prendre en compte, ce qui est inadapté dans notre cas (voir section 4.7 page 226). prevR permet néanmoins d'appliquer tout type de krigeage (ordinaire ou universel, local ou global, exact ou approximatif).

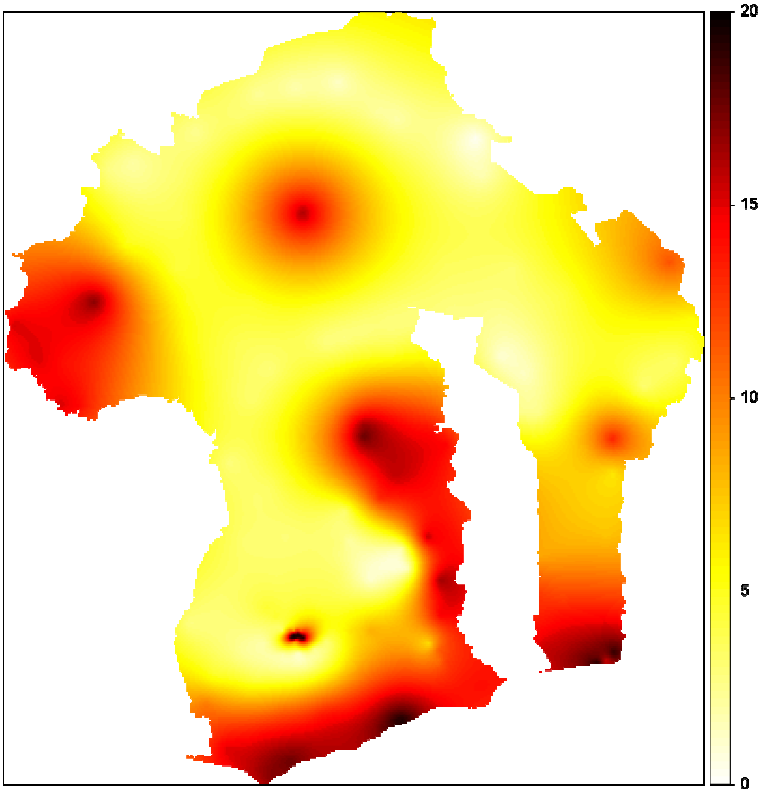
Le krigeage prend en compte la structure de dépendance spatiale des données en modélisant le semi-variogramme. Ce dernier correspond à la moitié du variogramme, c'est-à-dire à la variance totale moins la covariance en fonction de la distance. Pour procéder à une interpolation par krigeage, il faut dans un premier temps mesurer le semi-variogramme expérimental puis le modéliser sous la forme d'une fonction mathématique. Plusieurs familles de fonctions peuvent être utilisées. Pour notre part, nous avons eu recours à des semi-variogrammes de type exponentiel, ces derniers donnant de bons résultats à l'usage.

Nous avons ajusté les modèles de semi-variogramme aux semi-variogrammes expérimentaux selon la méthode des moindres carrés. Lorsqu'un ajustement automatique n'était pas satisfaisant, nous avons procédé à des ajustements visuels. Pour plus de détails, voir l'annexe 6.

L'une des forces du krigeage, c'est sa possibilité de pouvoir utiliser des semi-variogrammes différents selon des directions choisies par l'utilisateur (par exemple Nord-Sud et Est-Ouest). Pour notre part, nous avons utilisé des semi-variogrammes omnidirectionnels.

Le krigeage peut s'appliquer à une grande variété de situations et de distribution des points observés. Il s'avère en général plus efficace que d'autres méthodes telles que la pondération selon l'inverse de la distance ou les splines cubiques (GRATTON 2002). « *Le Krigeage est la méthode optimale, au sens statistique, d'interpolation et d'extrapolation. C'est la méthode d'estimation la plus précise.* » (GRATTON 2002)

Figure 4.12
Krigeage ordinaire à partir de la prévalence réelle des grappes sélectionnées lors d'une simulation d'une EDS



Dans les zones très peuplées (agglomérations urbaines), des variations locales sont correctement mises en évidence. La Figure 4.12 retranscrit correctement la diffusion de l'épidémie autour de F et la concentration de l'épidémie sur l'agglomération C. Les gradients depuis la côte au Sud sont également visibles ainsi que la rupture nette entre les deux rives du grand lac.

Cependant, la carte obtenue après interpolation des prévalences réelles accentue les contrastes et perd certaines variations fines. Si les principales variations et différentiels sont correctement rendus, le niveau estimé de la prévalence en un

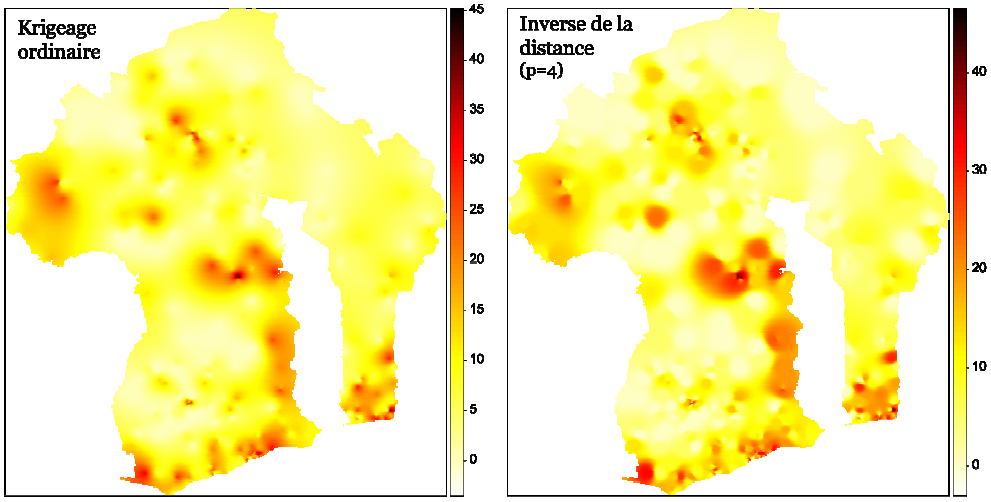
point donné peut différer de sa valeur réelle. Dans l'interprétation ultérieure de nos résultats à partir de données réelles, nous devrons donc garder à l'esprit que les cartes produites rendent bien compte des variations spatiales mais non des niveaux fin de la prévalence en un point donné.

Nous venons de vérifier qu'une interpolation spatiale par krigeage ordinaire permet bien de reconstruire les principales variations de l'épidémie de départ à partir des prévalences réelles.

Nous avons affirmé à la section 4.2 que, du fait des faibles effectifs de personnes testées par grappe, les prévalences observées traduisaient plus les variations aléatoires de l'échantillonnage que les variations spatiales de l'épidémie nationale. Pour vérifier notre propos, nous avons interpolé les prévalences brutes par krigeage ordinaire et pondération inverse de la distance (Figure 4.13 ci-dessous).

Figure 4.13

Interpolations spatiales, par krigeage et inverse de la distance, de la prévalence observée des grappes sélectionnées lors d'une simulation d'une EDS



Note : pour l'interpolation selon l'inverse de la distance, le facteur puissance de la distance était de 4.

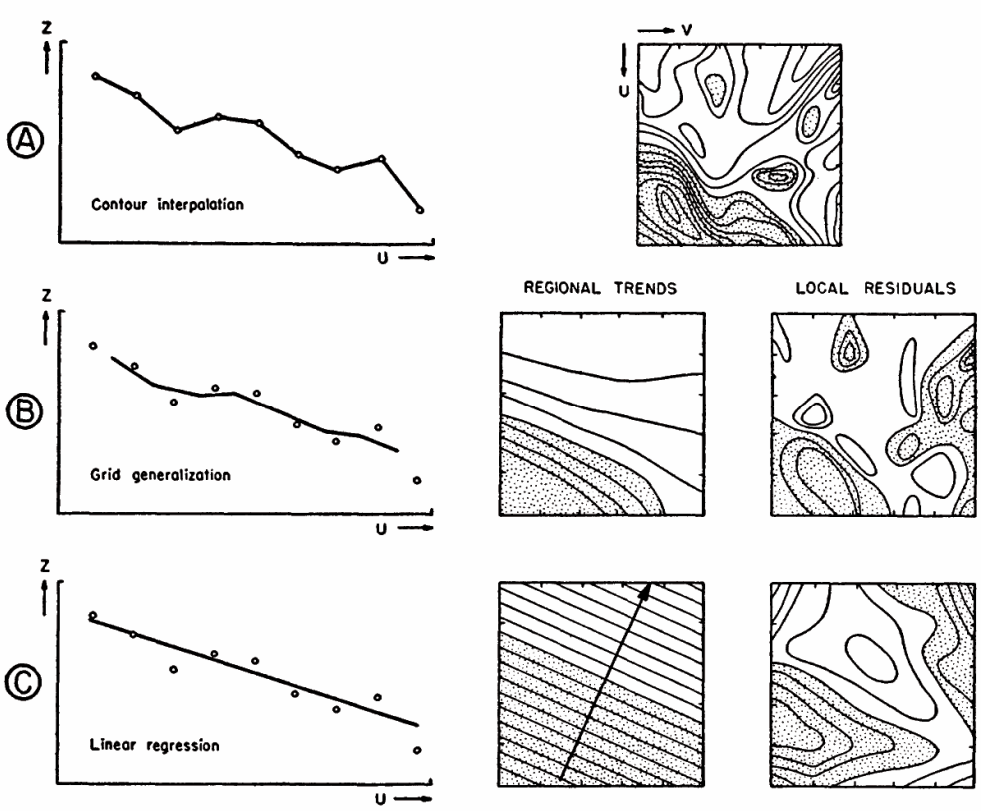
Les deux cartes produites diffèrent très sensiblement de l'épidémie de départ. Les prévalences interpolées atteignent des maximums élevés (environ 45 %) correspondant aux valeurs les plus élevées des prévalences observées. Si les zones les plus touchées sont déjà repérables (les côtes sud, l'agglomération C, les régions situées autour de D et F, le gradient depuis la frontière Nord-Ouest), de nombreuses variations aléatoires locales viennent brouiller les cartes.

Il est donc bien nécessaire d'estimer la prévalence de chaque grappe avant de procéder à une interpolation spatiale afin de réduire les variations aléatoires de l'échantillonnage.

4.6 L'analyse en composante d'échelles

À partir du milieu du XXe siècle, se sont développées les analyses en composante d'échelle (HAGGETT 1968, p. 301-309). Il s'agit de séparer, à l'aide de méthodes cartographiques, les composantes d'échelle supérieure (échelle régionale) des composantes d'échelle inférieure (échelle locale). La Figure 4.14 en fournit un exemple. Les variations spatiales du phénomène étudié (A), sont décomposées en tendances régionales et résidus locaux (B et C).

Figure 4.14
Analyse de la distribution d'une forêt au sein d'une section de 10 000 kilomètres carrés du bassin de Tagus-Sado au Portugal



Source : (CHORLEY 1965).
A : interpolation linéaire des contours ; B : généralisation à partir d'une grille ; C : régression linéaire.

Il s'agit donc d'extraire des variations d'une variable *Z* une tendance régionale *TR* à l'aide d'une technique statistique. Les résidus locaux *RL* s'obtiennent alors par soustraction selon l'Équation 4.3.

Équation 4.3
Décomposition d'une variable spatiale en composantes d'échelle

$$Z(x,y) = TR(x,y) + RL(x,y)$$

Z : *variable étudiée*
TR : *tendance régionale*
RL : *résidus locaux*
x, y : *longitude et latitude*

Nous avons montré précédemment que la prévalence observée par grappe pouvait se décomposer entre la prévalence réelle de la grappe et une erreur aléatoire d'espérance nulle et de variance $\sqrt{p(1-p)/N}$ [22]. La prévalence réelle peut, pour sa part, se décomposer en deux composantes d'échelles. Nous obtenons alors l'Équation 4.4.

Équation 4.4
Décomposition en composantes d'échelle de la prévalence observée

$$P_{obs}(x,y) = TR(x,y) + RL(x,y) + EA$$

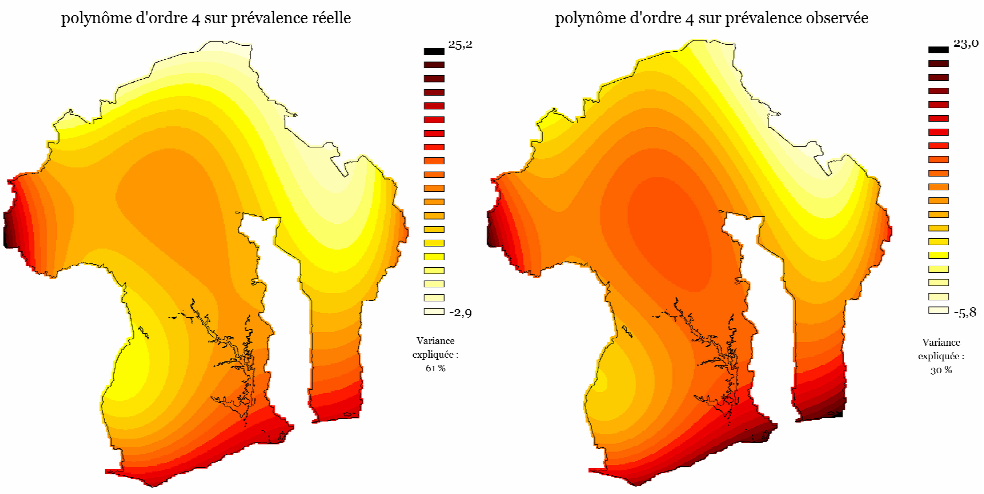
De nombreuses techniques différentes existent pour calculer une surface de tendance qui sera plus ou moins complexe et simplifiera donc plus ou moins le phénomène étudié. Il est possible d'avoir recours à différentes formes de régressions (linéaires, polynomiales…) qui tentent d'ajuster la valeur de la variable étudiée à une fonction mathématique de ses coordonnées (HAGGETT 1968, p. 305-307). La Figure 4.15 ci-dessous présente les surfaces de tendances calculées à l'aide de polynômes d'ordre 4 à partir de la prévalence réelle et de la prévalence observée des grappes de notre simulation[23]. Les coefficients du polynôme sont calculées selon la méthode dite « des moindres carrés ». Il s'agit de minimiser les écarts entre les valeurs prédites par le polynôme et les valeurs de départ. Comme l'erreur aléatoire est nulle en moyenne (les valeurs positives viennent compenser les valeurs négatives), les tendances régionales du polynôme ajusté à partir des prévalences observées s'avèrent proches de celles du polynôme ajusté à partir des prévalences réelles des grappes.

À partir de techniques en composante d'échelle, il est donc possible de dégager les tendances régionales de l'épidémie à partir des prévalences observées, malgré les valeurs élevées de l'erreur aléatoire, car ces dernières se compensent lorsque l'on passe à des échelles de niveaux supérieurs.

[22] En première approximation.

[23] Nous avons utilisé pour générer ces cartes le logiciel gratuit *Philcarto* développé par Philippe WANIEZ (http://philgeo.club.fr/).

Figure 4.15
Surfaces de tendance (polynômes d'ordre 4) calculées à partir de la prévalence réelle et de la prévalence observée des grappes d'une simulation EDS



Nous avons ici comparé le polynôme calculé à partir des prévalences observées à celui calculé à partir des prévalences réelles et non à l'épidémie de départ. Nous avons montré précédemment qu'il était possible de compenser les effets du tirage au premier degré à partir des techniques d'interpolation spatiale et de reconstruire les principales tendances de l'épidémie de départ à partir des prévalences réelles des zones enquêtées. Notre objectif, dans cette partie, vise à compenser les effets du tirage au second degré et donc à compenser les effets des erreurs aléatoires induites. Quelque soit la méthode que nous emploierons, nous ne pourrons arriver à un meilleur résultat que celui obtenu à partir des prévalences réelles. En effet, les variations locales se situant à des échelles inférieures à celle du maillage des grappes ne pourront jamais être estimées. Nous sommes nécessairement limités par la résolution de l'échantillonnage au premier degré. Notre objectif ici est donc d'essayer de reconstruire, à partir des prévalences observées, les variations régionales couvertes par la localisation des zones d'enquêtes, c'est-à-dire celles que nous pouvons estimer à partir des prévalences réelles des grappes enquêtées[24].

L'effet des erreurs aléatoires est visible sur l'estimation des polynômes de puissance 4 au travers de la variance expliquée. Il s'agit du pourcentage de la variance totale, calculée à partir des données de départ, correspondant à la variance des valeurs prédites par le polynôme. La surface de tendance d'ordre 4 explique 61 % de la variance des prévalences réelles, mais seulement 30 % de la variance des

[24] Rappelons que nous ne disposons de ces prévalences réelles que dans le cadre de notre pays modèle où nous les connaissons par définition. Ces prévalences réelles sont bien entendues inconnues dans le cadre des enquêtes réelles.

prévalences observées. En effet pour ces dernières, outre la variance des résidus locaux s'ajoute la variance des erreurs aléatoires.

Enfin, si les surfaces de tendances calculées à partir des prévalences observées sont pertinentes et traduisent correctement les variations régionales du phénomène étudié, nous ne pouvons étudier les résidus locaux. En effet, si pour chaque grappe, nous soustrayons à la prévalence observée la prévalence prédite par le polynôme, nous n'obtiendrons pas seulement les résidus locaux[25] mais la somme des résidus locaux et des erreurs aléatoires. Comme les erreurs aléatoires sont du même ordre de grandeur que les résidus locaux, voire supérieures, il ne nous est pas possible de distinguer ces résidus des erreurs. L'analyse des résidus locaux n'aurait donc ici aucun sens.

S'il nous est possible de calculer des surfaces de tendances polynomiales à partir des prévalences observées, nous ne pouvons ainsi mettre à jour que les grandes tendances de l'épidémie et nous perdons de l'information. Si les principaux gradients sont correctement reconstitués, les variations spatiales sont moins contrastées. Les variations autour des agglomérations urbaines (concentration ou diffusion) ne sont pas correctement rendues et la rupture de continuité de part et d'autre du grand lac a disparu.

[25] Ce qui serait le cas en procédant de manière identique à partir des prévalences réelles.

4.7 Méthode des cercles de même rayon

D'autres techniques de calcul des tendances régionales existent. Des auteurs tels que GRIFFIN, NETTLETON ou KRUBEIN ont ainsi utilisé des moyennes mobiles basées sur des cercles (GRIFFIN 1949, NETTLETON 1954, KRUMBEIN 1956). Plusieurs variantes existent. La plus courante consiste, pour chaque point d'une grille, à tracer un cercle d'un certain rayon autour de celui-ci puis de calculer la moyenne des observations présentes dans le dit cercle. Cette valeur moyenne est ensuite appliquée au point central.

Nous avons alors eu recours à une approche similaire que nous avons nommé *méthode R*. Pour chaque grappe enquêtée, nous avons calculée sa distance aux autres grappes puis nous avons trié ces autres grappes de manière croissante. La Figure 4.16 présente la répartition des grappes de l'EDS 2003 du Burkina Faso en fonction de leur distance à une grappe donnée *x* et des effectifs cumulés du nombre de personnes testées.

Figure 4.16
Sélection des grappes retenues pour l'estimation de la prévalence d'une grappe, méthode R

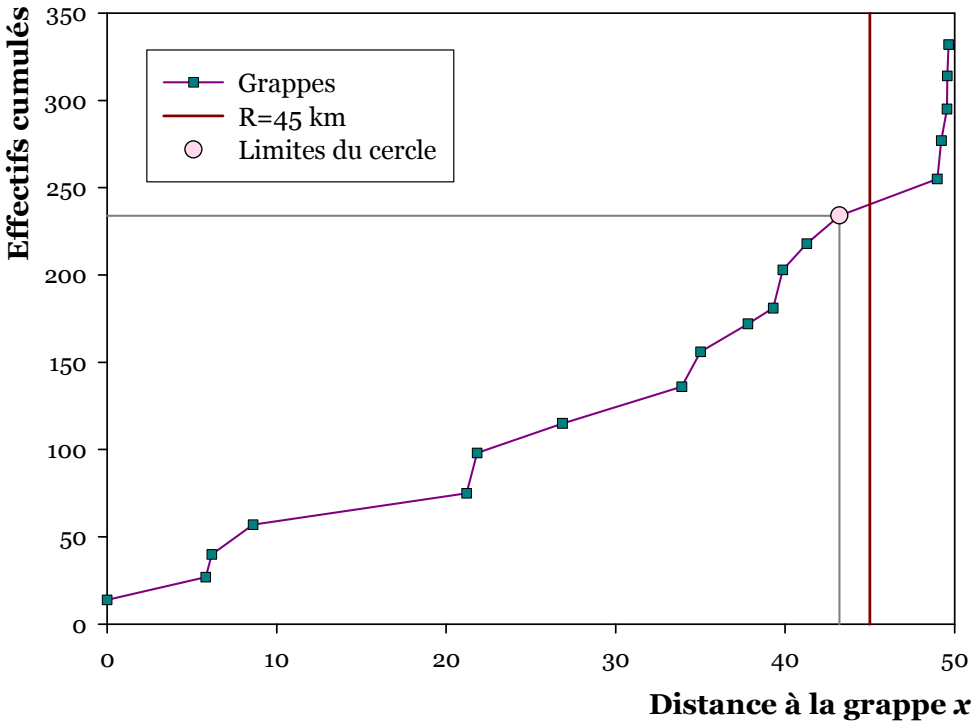
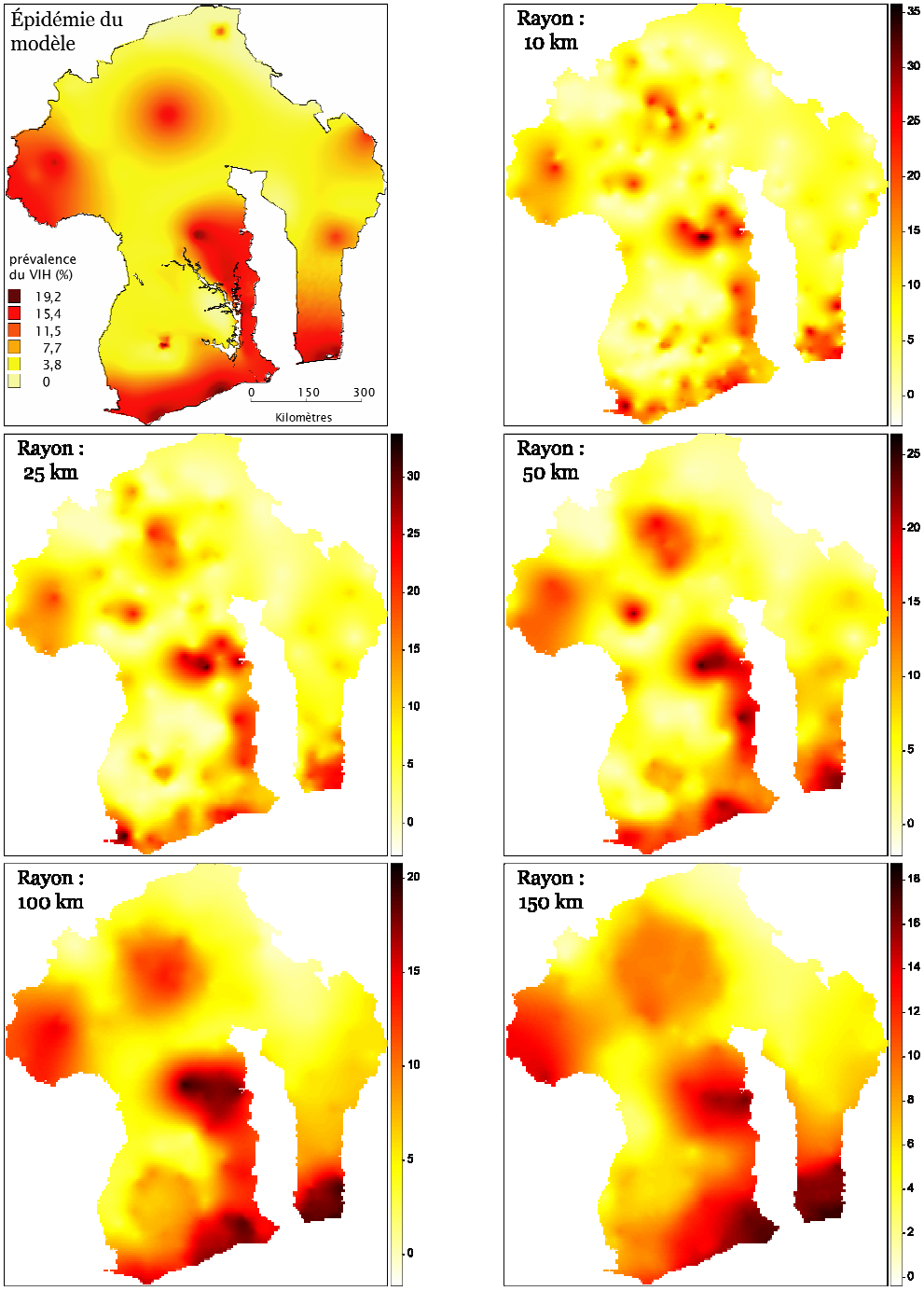
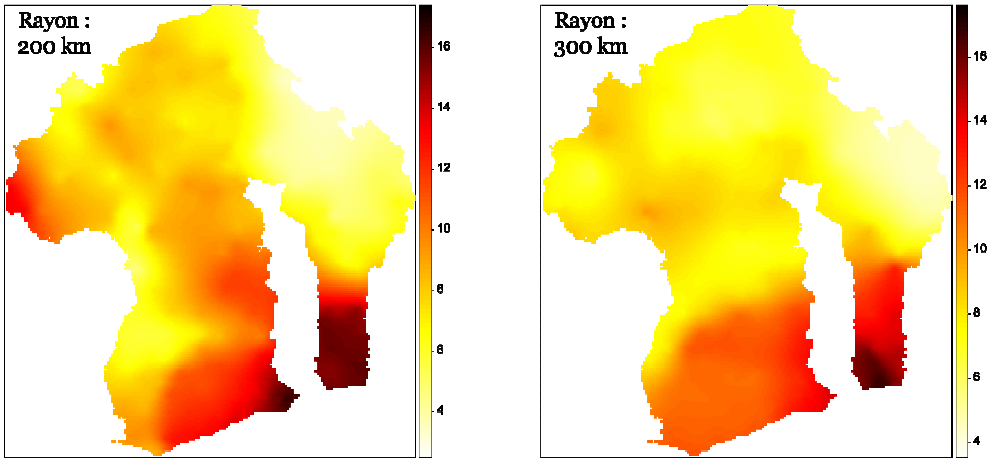


Figure 4.17

Interpolation des prévalences estimées de chaque grappe avec différentes valeurs du rayon et épidémie d'origine du modèle





Si nous choisissons un rayon *R* de 45 kilomètres, alors, nous allons retenir l'ensemble des grappes situées à moins de 45 kilomètres de *x* pour estimer la prévalence de cette grappe. Dans cet exemple, cela correspond à 14 grappes (*x* compris) au sein desquelles 234 individus ont été testés pour le VIH. Nous affecterons alors comme prévalence ajustée à la grappe *x* la prévalence calculée sur ces 234 personnes, en tenant compte de leurs taux de pondération respectifs. Le rayon exact du cercle de lissage s'avère être de 43,2 kilomètres (distance de la grappe sélectionnée la plus éloignée).

Une fois la prévalence de chaque grappe estimée de cette manière, nous utilisons le krigeage ordinaire pour générer nos cartes. La Figure 4.17 présente les résultats obtenus à partir de notre simulation d'EDS avec différentes valeurs de R : 10, 25, 50, 100, 150, 200 et 300 kilomètres. Nous avons également fait figurer, pour comparaison, l'épidémie de départ du modèle. Les différentes cartes sont à la même échelle cartographique. Cependant, l'échelle colorimétrique diffère d'une carte à une autre.

Au fur et à mesure que le rayon de nos cercles augmente, le nombre de grappes sélectionnées pour l'estimation de la prévalence de chacune d'elles s'accroît, ainsi que le nombre de personnes testées. Plus ce dernier est important, meilleure est la compensation des erreurs aléatoires des prévalences observées. Aux petits rayons (10 ou 25 kilomètres), les erreurs aléatoires sont clairement visibles sur les cartes produites. D'ailleurs, ces cartes présentent des prévalences maximales encore élevées (30 à 40 %) alors que l'épidémie de départ ne dépasse pas les 20 %. Lorsque le rayon devient très important (200 à 300 kilomètres), le lissage devient très important et seules les grandes tendances régionales sont mises à jour. L'information devient donc moins précise. Notre objectif vise à obtenir l'information la plus précise possible tant en réduisant au maximum les erreurs aléatoires. Pour la simulation d'EDS utilisée, la valeur de 100 kilomètres comme rayon des cercles semble le meilleur compromis par comparaison avec l'épidémie de départ.

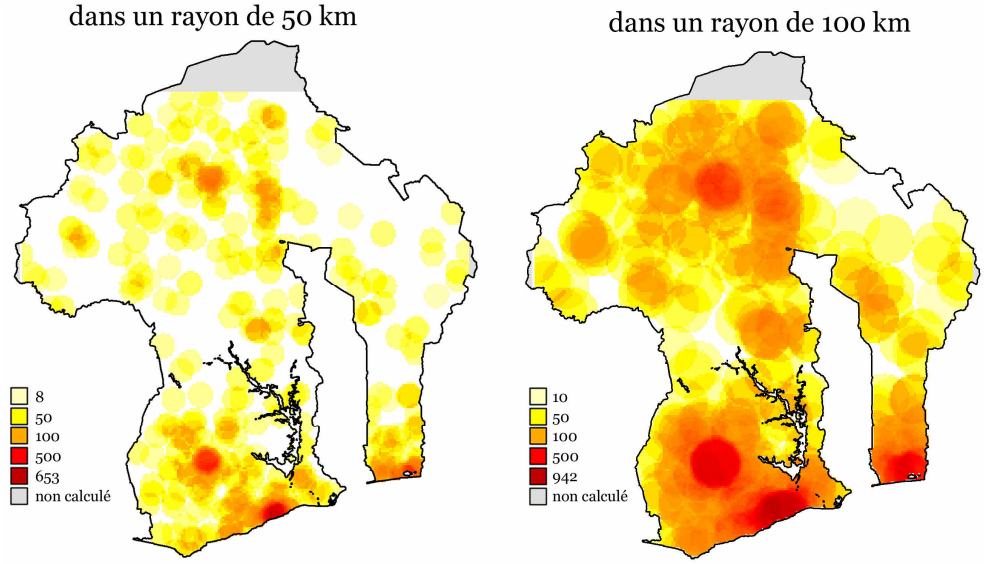
Cependant, avec un rayon aussi important, les estimations réalisées au niveau des agglomérations urbaines sont masquées. Ainsi, la diffusion autour de la petite ville H[26] et depuis la frontière Est ne sont plus visibles. L'épidémie concentrée de l'agglomération C s'est en quelque sorte « diluée » pour former une région d'épidémie moyenne.

Or, dans le sud d'Alicante, fortement peuplée, nous disposons de nombreuses observations. La Figure 4.18 et la Figure 4.19 permettent de s'en rendre compte. Sur la première, nous avons représenté pour chaque point de la grille le nombre de personnes testées se situant dans un rayon de 50 ou 100 kilomètres. Il apparaît que la situation est très inégale d'une zone à l'autre, du fait de la répartition inégale des zones d'enquêtes (voir Figure 4.8 page 211 pour rappel).

Sur la Figure 4.19 nous avons interpolé par krigeage ordinaire le nombre d'individus testés pris en compte pour l'estimation de la prévalence de chaque pour deux valeurs du paramètre R (50 et 100 kilomètres). Avec un rayon de 50 kilomètres, dans de nombreuses zones les prévalences ont été estimées sur moins de 50 personnes (en jaune clair) ou moins de 100 personnes (vert clair). Dans ces régions, les erreurs aléatoires restent donc encore importantes. En augmentant le rayon du cercle à 100 kilomètres, presque la prévalence de la majorité des grappes a été estimée sur au moins cent personnes et le plus souvent sur plus de deux cents individus, d'où une bonne réduction des erreurs aléatoires.

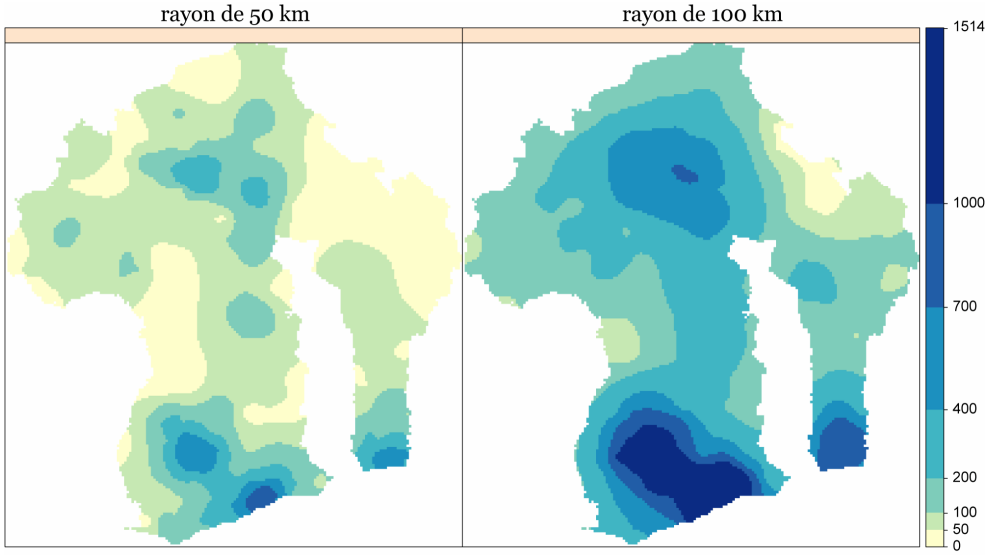
Figure 4.18

Nombre de personnes testées dans un rayon de 50 ou 100 kilomètres



[26] Rappel : les noms des agglomérations urbaines d'Alicante sont mentionnés sur la Figure 4.7 page 210.

Figure 4.19
Interpolation par krigeage ordinaire du nombre d'individus pris en compte pour l'estimation de la prévalence de chaque grappe avec des cercles de 50 ou 100 km

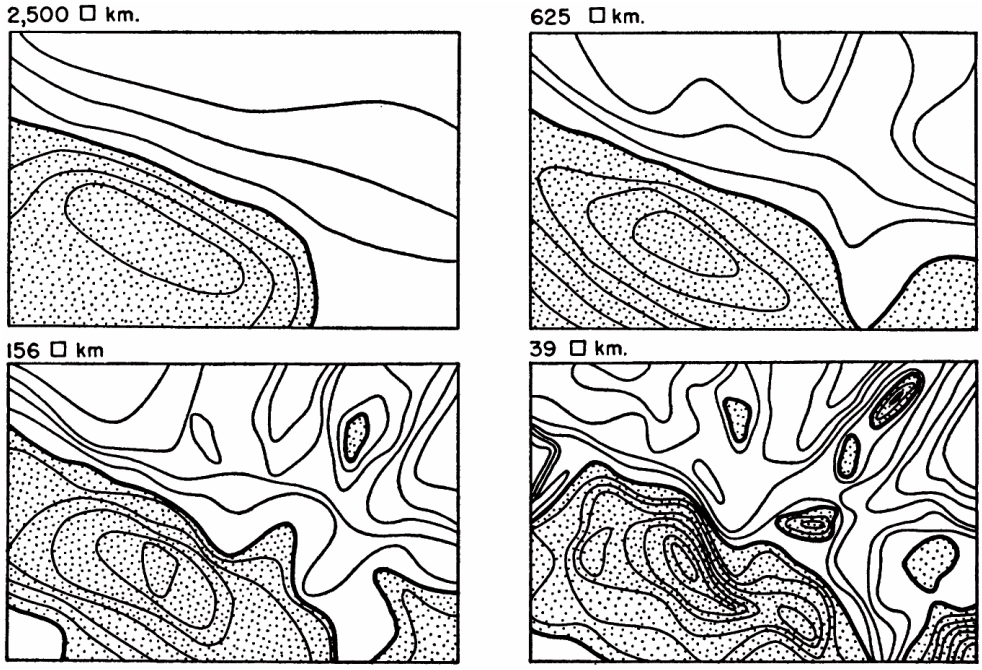


Dans le sud d'Alicante et autour de F, les estimations auront été réalisées à partir de plus de 400 personnes, voire plus de 700 ou encore 1 000 personnes. Pour ces régions là, il serait pertinent d'avoir recours à des cercles plus petits. En effet, nous pourrions toujours obtenir une bonne réduction des erreurs aléatoires en réalisant nos estimations sur moins d'individus tout en augmentant la précision des cartes produites car les estimations seraient réalisées sur un nombre d'individus plus faible.

4.8 Un lissage adaptatif : les cercles de même effectif

La Figure 4.20 présente les surfaces de tendance obtenues sur la distribution d'une forêt au Portugal à l'aide d'une moyenne mobile sauf qu'au lieu d'utiliser des cercles, ce sont des carrés centrés sur chaque point à lisser qui ont été utilisés. Elle met en évidence que plus la superficie de la fenêtre de lissage est réduite, plus la courbe des tendances régionales est complexe et intègre des variations locales.

Figure 4.20
Analyse de la distribution d'une forêt au sein d'une section de 15 000 kilomètres carrés dans le bassin du Tagus-Sado au Portugal par lissage avec des carrés de 2 500, 625, 156 et 39 km²



Source : (CHORLEY 1965).
Le lissage à l'aide de carrés utilisé ici est une variante du lissage avec des cercles.

L'objectif usuel des décompositions en surfaces de tendance consiste à simplifier les variations d'un phénomène afin d'en dégager les principales caractéristiques. La perte d'informations localisées est donc recherchée.

Dans le cas présent, nous avons recours à des techniques de lissage non pas pour simplifier le phénomène que nous étudions, puisque nous cherchons justement à mettre jour des variations à des échelles inférieures à celles des grandes régions telles qu'elles sont codées dans les EDS. La décomposition en composantes

d'échelle vise ici à éliminer les erreurs aléatoires du tirage au second degré tout en essayant de conserver un maximum de précision locale.

Avec des cercles de même rayon, il nous faut prendre des valeurs de R suffisamment élevées pour l'estimation de la prévalence de chaque grappe puisse porter sur un nombre suffisant d'individus. Mais dans le même temps, dans les zones fortement peuplées et enquêtées, il serait possible d'avoir recours à des cercles plus petit car les effectifs sont suffisants. Comme ce qui permet de réduire les erreurs aléatoires est justement le nombre de personnes retenues pour l'estimation des prévalences, il semble plus opportun d'avoir recours à une approche par des cercles non pas de même rayon mais de même effectif : ce que nous appellerons la *méthode N*.

Cette fois-ci, nous fixons un effectif donné *N*, puis, pour chaque grappe, nous estimons la prévalence du VIH à partir des grappes situées dans un cercle tel que le nombre d'individus testés y soit au moins égal à N. Après avoir calculé les effectifs cumulés en fonction de la distance à la grappe considérée, nous sélectionnons la première grappe située au-dessus de N et toutes les grappes situées en dessous (Figure 4.21). Le rayon du cercle de lissage sera donc variable d'une grappe à l'autre, faible là où les grappes sont concentrées, élevé là où il y a peu de zones d'enquêtes.

Figure 4.21
Sélection des grappes retenues pour l'estimation de la prévalence d'une grappe, méthode N

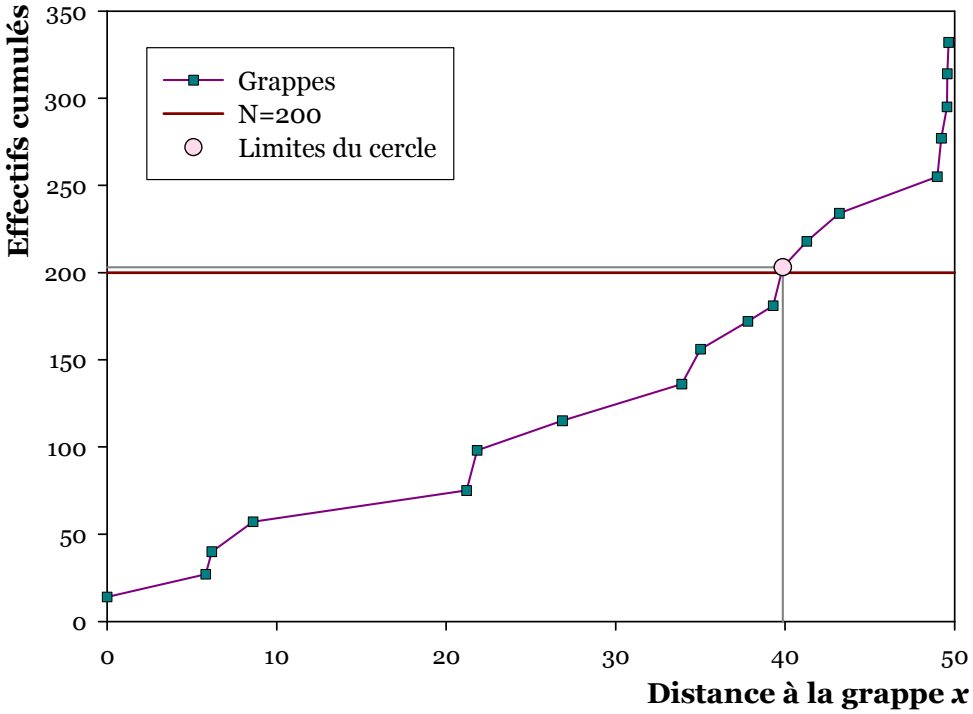
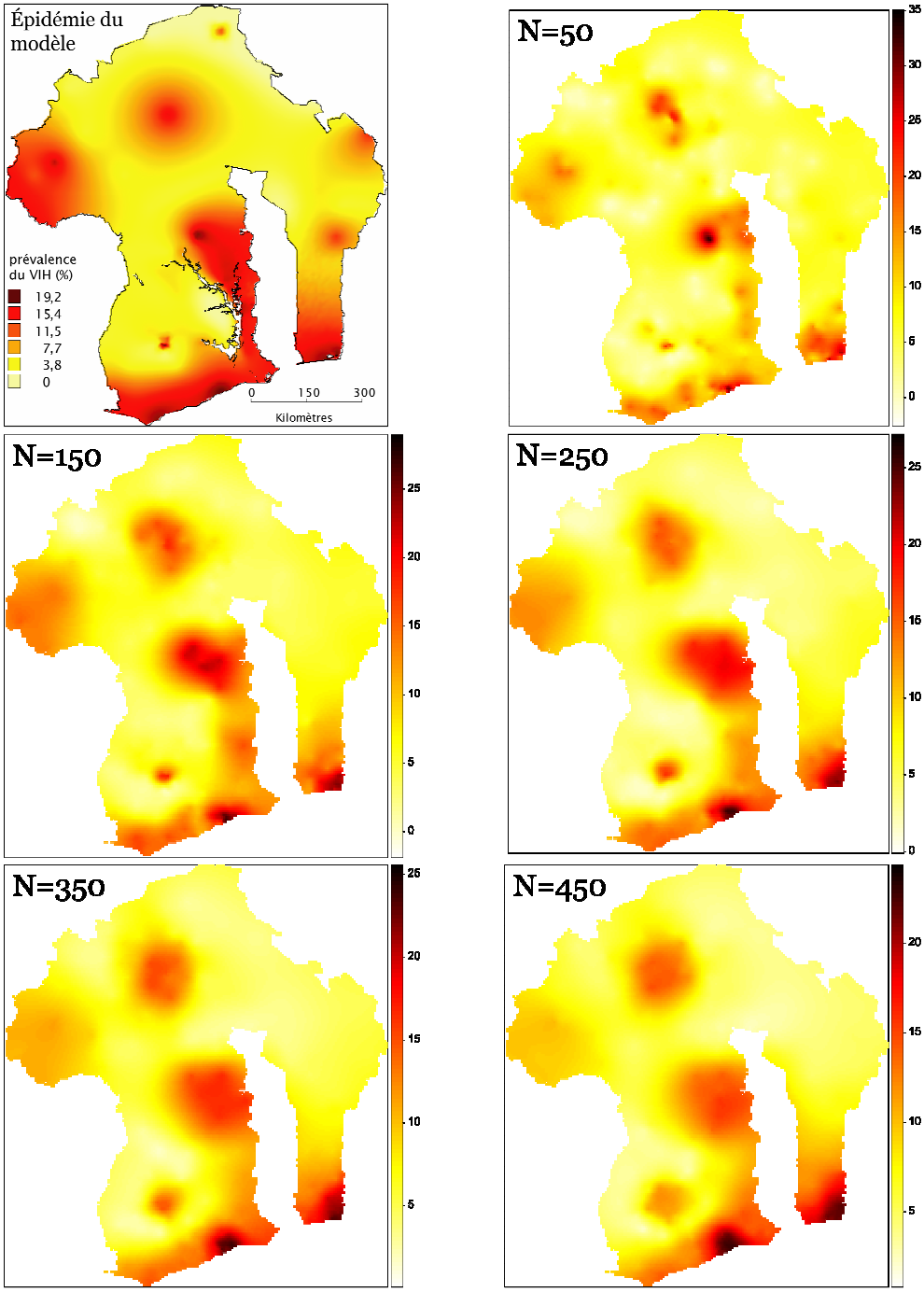


Figure 4.22
Interpolation des prévalences estimées de chaque grappe avec différentes valeurs de N et épidémie d'origine du modèle



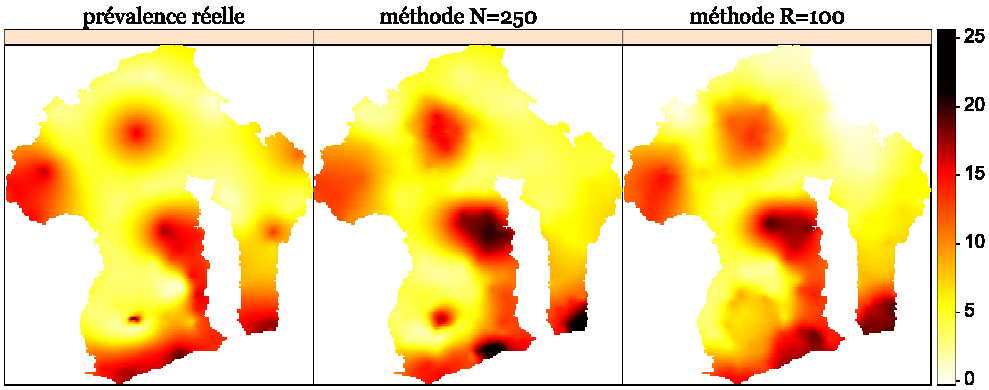
Comme précédemment, les prévalences ajustées sont ensuite interpolées par krigeage ordinaire pour produire des cartes. La Figure 4.22 présente les résultats obtenus pour différentes valeurs de N (de 50 à 450).

Comme précédemment, une valeur insuffisante de N (50) ne permettra pas aux erreurs aléatoires de se compenser suffisamment et une valeur trop élevée induira un lissage excessif. Bien que le rayon des cercles soit désormais variable, il existe un lien entre l'effectif N choisi et ce rayon puisque plus cet effectif augmente et plus le rayon des cercles de lissage augmentera lui aussi.

Par ailleurs, il apparaît que l'évolution du lissage est relativement progressive lorsque le paramètre N augmente de 100 en 100. Cela signifie donc que deux valeurs de N distante entre 25 et 50 produiront des cartes relativement semblables.

Figure 4.23

Comparaison méthodes R et N et interpolation selon prévalence réelle



Si nous comparons les résultats obtenus avec la méthode N et la méthode R (Figure 4.23, la carte obtenue par interpolation des prévalences réelles est donnée à titre de rappel), il apparaît que les contrastes sont plus marqués avec la méthode N qu'avec la méthode R. Les agglomérations urbaines y sont plus visibles. Aucune des deux méthodes n'arrive à rendre la diffusion autour de la petite ville H ou le gradient depuis la frontière Est alors que l'interpolation à partir des prévalences réelles arrive à retranscrire ces deux phénomènes.

Concernant le Nord du pays, faiblement enquêté, les deux méthodes induisent des résultats assez proches puisque les effectifs étant relativement faibles, la méthode N nécessite des cercles relativement grands pour procéder au lissage. Par contre, dans le Sud d'Alicante, là où la méthode R induisait des lissages trop importants, du fait des effectifs élevés, la méthode N procède à un lissage sur des cercles plus petits. L'épidémie autour de l'agglomération C est ainsi relativement concentrée avec la méthode N alors que dans la méthode R le pic épidémique était comme « dilué ».

L'approche par des cercles de même effectif s'avère donc plus efficace que celle par des cercles de même rayon car elle permet de travailler à des échelles différentes selon la quantité d'information (c'est-à-dire de personnes testées) de chaque zone.

4.9 Optimisation du paramètre N

Notre objectif consiste à minimiser l'erreur aléatoire tout en gardant un maximum d'information locale. Quand N augmente, l'erreur aléatoire diminue. Mais dans le même temps, nous lissons plus fortement les données, autrement dit nous simplifions le phénomène ce qui induit une augmentation des résidus locaux. Lorsque N atteint la taille totale de l'échantillon, la tendance régionale se retrouve réduite à une constante (la prévalence nationale[27]) et les résidus locaux sont alors maximums. La meilleure estimation possible à partir des données, c'est-à-dire l'estimation la plus fine géographiquement avec les erreurs aléatoires les plus faibles, sera obtenue lorsque les prévalences estimées seront les plus proches possibles des prévalences réelles.

Dans un usage à partir de données réelles d'enquête, nous ne connaissons pas la prévalence réelle des grappes. Nous avons seulement à notre disposition les prévalences observées. Cependant, pour Alicante, nous avons cette information à disposition. Nous procéderons alors par *analogie*. En simulant des EDS avec Alicante présentant les mêmes paramètres que celle que nous étudions (nombre total de personnes testées, nombre de clusters, prévalence nationale), nous pourrons par cette modélisation déterminer une *valeur optimale de N* dont nous ferons l'hypothèse qu'elle peut s'appliquer à notre EDS particulière.

Si nous reprenons la simulation d'EDS qui nous a servi jusqu'à présent, nous pouvons calculer pour chaque grappe la prévalence ajustée lorsque nous utilisons une certaine valeur du paramètre N puis calculer l'écart entre cette prévalence ajustée et la prévalence réelle, puis la moyenne de ces écarts[28]. La courbe de la Figure 4.24 montre comment évolue cet écart moyen, pour notre simulation exemple, lorsque N augmente. Il commence par diminuer fortement du fait de la compensation des erreurs aléatoires. Puis, après avoir atteint un minima, l'écart moyen augmente progressivement : le lissage des données devient plus important et les résidus locaux augmentent (les variations locales sont « gommées »). La valeur de N qui répond le mieux à notre objectif est donc celle qui minimise l'écart moyen, à savoir 200 à 250.

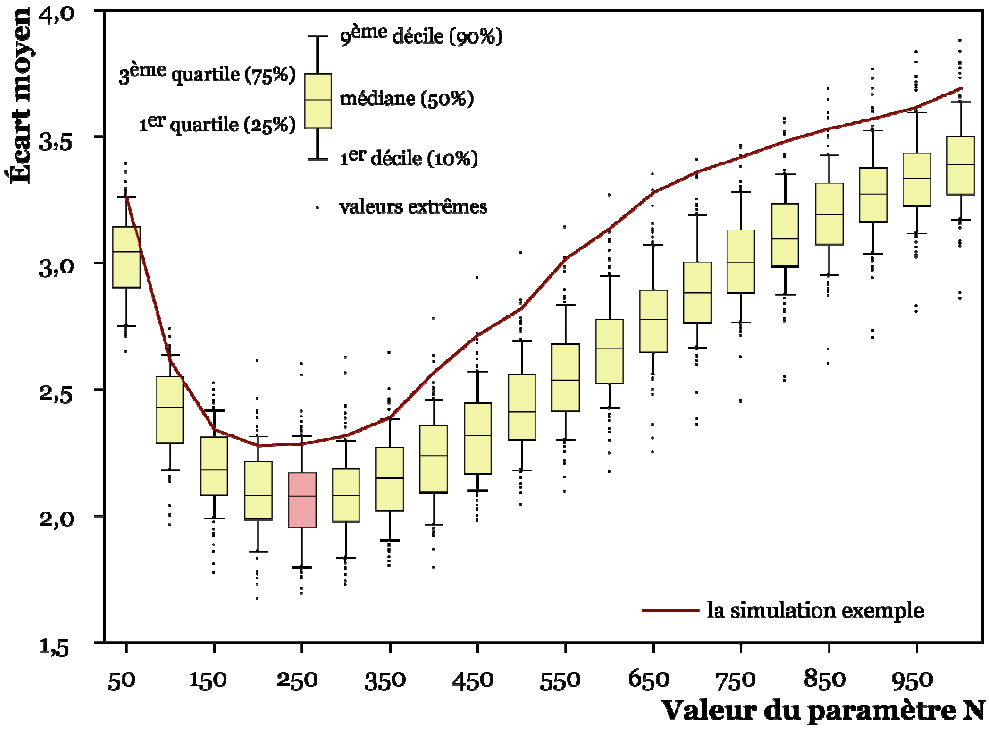
Cependant, rien ne garantit que cette valeur de N pour laquelle l'écart est minium soit la même d'une simulation à une autre. Nous avons donc réalisé 100 simulations avec les mêmes paramètres (8 000 personnes testées, 400 clusters et une prévalence nationale de 10 %). Pour chacune, nous avons calculé l'écart moyen

[27] En effet, si N est égal à la taille totale de l'échantillon, la prévalence est estimée à partir de toutes les observations et est donc égale à la moyenne nationale.

[28] Les écarts pouvant être positifs ou négatifs, il s'agit plus précisément de la moyenne des valeurs absolues de ces écarts.

pour différentes valeurs de N. Les résultats sont représentés par les boîtes à moustache de la Figure 4.24.

Figure 4.24
Écart moyen entre prévalence ajustée et prévalence réelle, selon N, pour 100 simulations d'EDS et la simulation exemple



Paramètres des simulations : 8 000 personnes testées, 400 grappes, prévalence nationale de 10 %.

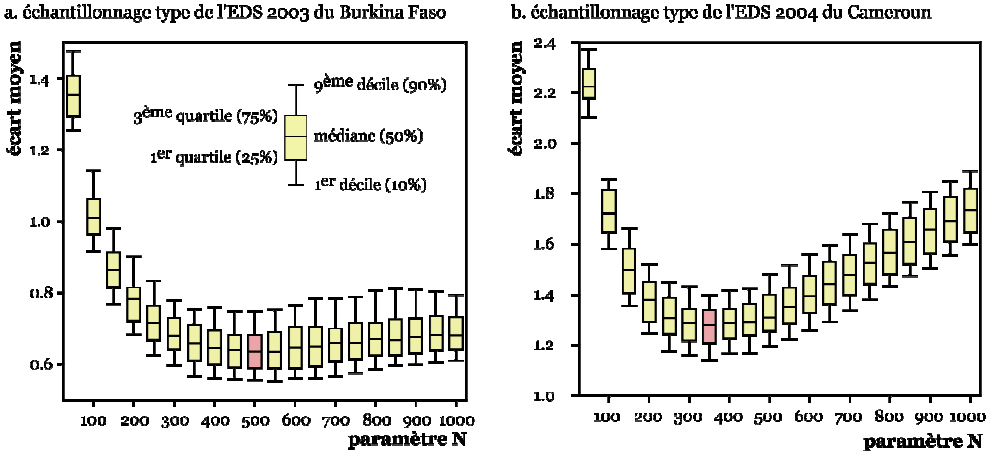
Bien que la valeur de N, pour laquelle l'écart moyen se minimise, puisse varier légèrement d'une simulation à l'autre, l'ensemble des simulations présente un même pattern. Par ailleurs, nous avons vu précédemment que de légères variations de N n'induisaient pas une modification importante de la carte produite. Nous retiendrons alors comme valeur optimale de N celle qui minimise l'écart moyen mesuré à partir d'une centaine de simulations (avec les mêmes paramètres), soit 250 pour notre simulation exemple.

La Figure 4.25 présente les résultats après 100 simulations d'EDS avec les paramètres de l'EDS 2003 du Burkina Faso et 100 simulations avec ceux de l'EDS 2004 du Cameroun[29]. Les valeurs de N qui minimisent l'écart moyen sont alors respectivement de 500 et 350.

[29] Nous rappelons que nous pouvons modifier la prévalence nationale d'Alicante en multipliant par un même facteur d'échelle la prévalence de chaque point.

Figure 4.25

Optimisation du paramètre N à partir de simulations présentant les échantillonnages des EDS du Burkina Faso et du Cameroun



Paramètres des simulations :

- Burkina Faso : 7 244 personnes testées, 400 grappes, prévalence nationale de 1,8 %.
- Cameroun : 9 900 personnes testées, 466 grappes, prévalence nationale de 5,5 %.

Il apparaît que les valeurs optimales de N varient en fonction des paramètres d'échantillonnage, à savoir le nombre total de personnes testées, le nombre de clusters et la prévalence nationale tandis qu'elles sont stables lorsque ces paramètres restent constants. Il importe donc de mieux comprendre comment ces derniers influent sur $N_{optimal}$.

Par ailleurs, la simulation d'une centaine d'EDS et le calcul des écarts moyens nécessite, sur un ordinateur de bureau de puissance moyenne, une dizaine à une quinzaine d'heures. Cette durée pourrait être probablement réduite en optimisant le code informatique du programme que nous avons utilisé. Néanmoins, le calcul devrait toujours nécessiter plusieurs heures. Si nous pouvions modéliser le rapport entre la valeur optimale de N et les paramètres d'échantillonnage, cela permettra un calcul plus rapide d'une valeur adéquate de N.

4.10 Modélisation de $N_{optimal}$

Nous avons simulé des enquêtes de type EDS, sur notre pays modèle, selon plusieurs valeurs des paramètres d'échantillonnage :

- prévalences nationales de 1, 2, 5, 10, 15, 30 et 45 % ;
- nombre total de personnes testées de 5 000, 6 000, 7 200, 8 640, 10 368 12 442 et 14 930 ;
- nombre de grappes enquêtées de 300, 360, 432, 518 et 622[30].

Pour chaque combinaison de ces trois paramètres, 100 simulations d'enquête ont été effectuées soit un total de 24 500 simulations. Pour chaque simulation, nous avons estimé les prévalences des grappes à l'aide de la méthode N en faisant varier N de 50 à 1 000 par paliers de 50 puis calculé l'écart moyen de chacune de ces 490 000 estimations. Pour chaque enquête simulée, nous avons déterminé la valeur de N pour laquelle les écarts moyens se minimisaient. Puis, pour chacune des 245 combinaisons de paramètres[31], la valeur optimale de N a été calculée comme étant la médiane[32] sur 100 simulations de la valeur minimisant l'écart moyen. Nous avons également calculé pour ces 100 simulations le rayon moyen des cercles de lissages pour cette valeur optimale de N. Cette opération aura nécessité 18 mois sur un ordinateur de puissance moyenne.

Les résultats sont présentés sur la Figure 4.26 ci-dessous. Chaque point représente la valeur de $N_{optimal}$ et le rayon moyen des cercles de lissage pour 100 simulations avec les mêmes paramètres d'échantillonnage. Le nombre de personnes testées est représenté par différentes couleurs et la prévalence nationale par différents symboles. Les points de même symbole et de même couleur ne diffèrent que par le nombre de grappes (entre 300 et 622). Ces points étant relativement groupés, il apparaît que le nombre de grappes influe peu sur la détermination du $N_{optimal}$.

Pour faciliter la lecture de ce graphique, nous avons rajouté deux flèches représentant respectivement l'augmentation croissante de la prévalence nationale et du nombre total de personnes testées.

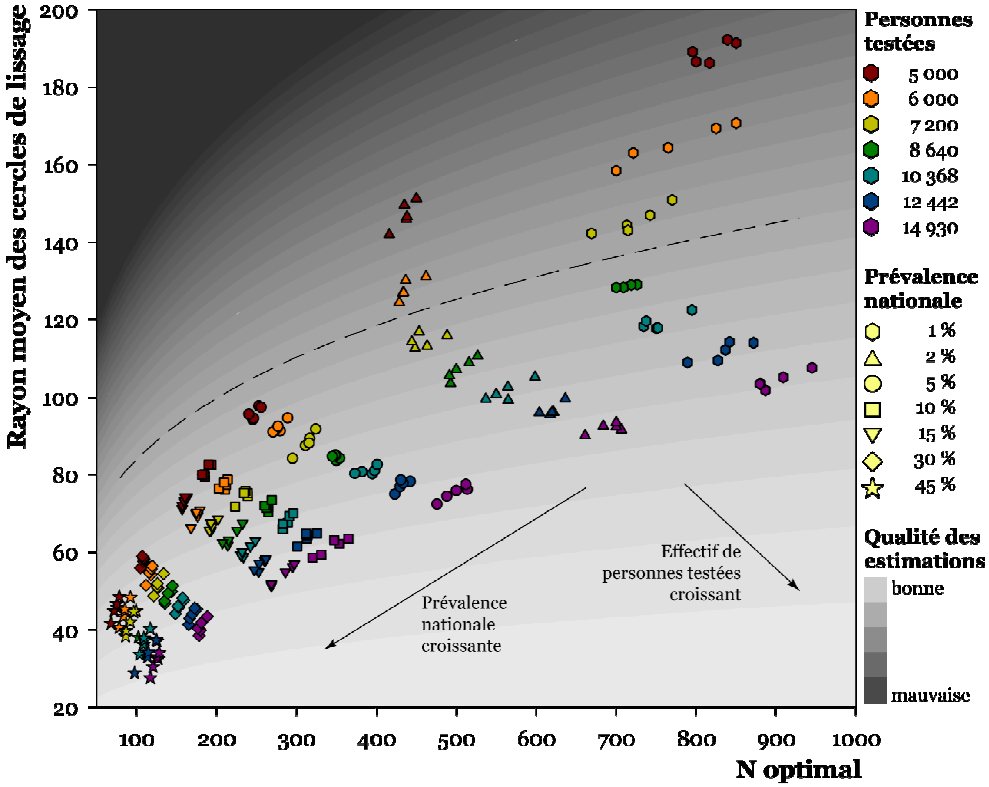
[30] Les valeurs des effectifs totaux et du nombre de grappes ont été choisies de telle manière que leur rapport soit constant. Ainsi, 5 000/300 ≈ 6 000/360 ≈ 7 200/432, etc.

[31] Même prévalence nationale, même nombre de personnes testées et même nombre de grappes.

[32] Plus précisément, nous avons utilisé la médiane de groupes qui considère chaque valeur comme un centre de classes.

Figure 4.26

Valeurs de $N_{optimal}$ et rayon moyen des cercles de lissage selon le nombre de personnes testées, le nombre de grappes et la prévalence nationale



Chaque point est calculé à partir de 100 simulations d'enquête de type EDS avec les mêmes paramètres d'échantillonnage (effectif de personnes testées, prévalence nationale et nombre de grappes). Les points de même forme et de même couleur diffèrent uniquement selon le nombre de grappes (300, 360, 432, 518 ou 622). Un indicateur de qualité globale des estimations peut être obtenu en faisant le rapport du carré du rayon moyen par la racine de N optimal ($R^2/\sqrt{N_{optimal}}$). Plus cette valeur est faible (gris clair), meilleure seront les prévalences estimées.

Équation 4.5

Indicateur de qualité globale de la Figure 4.26

$$IQ = \frac{R^2_{moyen}}{\sqrt{N_{optimal}}}$$

IQ : *indicateur de qualité globale*

R_{moyen} : *rayon moyen des cercles de lissage*

$N_{optimal}$: *valeur optimale pour le paramètre N*

Pour ailleurs, nous avons représenté par dégradé de gris un indicateur de qualité globale des estimations. En effet, les prévalences ajustées seront d'autant plus précises que la superficie des cercles de lissage sera petite, donc que le carré du rayon moyen sera faible. Par ailleurs, plus $N_{optimal}$ sera important, plus les erreurs aléatoires seront réduites, l'amplitude des intervalles de confiance d'une proportion

étant fonction de la racine de l'effectif. Nous avons donc construit un indicateur de qualité selon l'Équation 4.5. Plus la valeur de cet indicateur est faible (gris clair), meilleure sera l'estimation des prévalences de chaque grappe.

Pour la majorité des observations (nombre de personnes testées supérieur ou égal à 5 000 pour une prévalence nationale supérieure à 5 %, supérieur ou égal à 7 200 pour une prévalence nationale de 2 % et supérieur ou égal à 8 640[33]), lorsque l'effectif de personnes testées augmente, la valeur optimale de N augmente également tandis que, simultanément, le rayon moyen des cercles de lissage diminue. Les prévalences estimées sont donc à la fois moins soumises aux erreurs aléatoires, car calculées sur un effectif plus important, et plus localisées, le lissage s'effectuant sur des zones plus petites. Lorsque la prévalence nationale est plus élevée[34], N est optimisé pour des valeurs plus faibles, les variations aléatoires de l'échantillonnage étant moins élevées. Il en résulte que le lissage s'effectue alors sur des cercles de plus petite taille. La réduction du rayon moyen vient compenser la réduction du paramètre N et les prévalences estimées sont ainsi de meilleure qualité quand la prévalence augmente.

Aux petites prévalences, il est nécessaire d'avoir suffisamment de personnes enquêtées pour estimer convenablement une proportion. Lorsque l'effectif total de personnes testées est également faible (points situés au-dessus de la courbe en pointillés), une diminution du nombre de personnes testées induit une augmentation des erreurs aléatoires telles que la valeur de N qui minimise l'écart moyen augmente. Les estimations réalisées par la méthode des cercles de même effectif dans ces conditions spécifiques seront donc fortement lissées, le rayon moyen des cercles de lissage devenant très important.

Nous avons voulu modéliser, pour les points situés sous la courbe en pointillés, la valeur de $N_{optimal}$ en fonction des paramètres d'échantillonnage. Comme le nombre de grappes influe peu, nous avons recherché une famille de fonctions capables de rendre compte de l'impact du nombre de personnes testées et de la prévalence nationale sur la valeur de $N_{optimal}$. Pour cela, nous avons eu recours à la procédure d'ajustement de fonctions du logiciel SPSS, version 15.0 pour Windows, qui permet d'ajuster 11 familles de fonctions aux données. Les relations entre $N_{optimal}$ et l'effectif total, pour une prévalence nationale donnée, et entre $N_{optimal}$ et la prévalence nationale, pour un effectif total donné, ne sont pas linéaires. Les courbes qui s'ajustent le mieux aux données sont les fonctions puissances. Il est donc possible de modéliser la relation entre $N_{optimal}$ et les paramètres d'échantillonnage selon le modèle défini par l'Équation 4.6.

[33] Il s'agit des points situés sous la ligne pointillée sur la Figure 4.26.

[34] C'est-à-dire plus proche de 50 %, puis si la prévalence nationale p était supérieure à 50 %, la situation serait statistiquement équivalente à l'estimation de la prévalence complémentaire 1-p.

Équation 4.6

Modèle utilisé pour exprimer $N_{optimal}$ en fonction des paramètres d'échantillonnage

$$N_{optimal} = b_0 \cdot effectif\ total^{b_1} \cdot prévalence\ nationale^{b_2} \cdot nombre\ de\ grappes^{b_3} + c$$

Pour déterminer la valeur des coefficients b_0, b_1, b_2, b_3 et c, nous avons effectué la procédure de régression non linéaire sous SPSS qui permet de spécifier son propre modèle. Nous obtenons alors l'Équation 4.7.

Équation 4.7

Expression de $N_{optimal}$ en fonction des paramètres d'échantillonnage, calculée à partir de 22 000 simulations d'enquête

$$N_{optimal} = 14{,}172 \cdot effectif\ total^{0{,}419} \cdot prévalence nationale^{-0{,}361} \cdot nombre de grappes^{0{,}037} - 91{,}011$$

Pour vérifier la capacité du modèle à calculer les valeurs de $N_{optimal}$, nous avons comparé les valeurs obtenues, selon les paramétrages de cinq EDS récentes, à l'aide de l'Équation 4.7 et en procédant à cent simulations d'EDS[35] (Tableau 4.2). Les résultats sont proches (écart inférieur à 20). Or, nous avons montré précédemment (section 4.8) que les cartes générées étaient peu sensibles à une petite variation du paramètre N.

Tableau 4.2

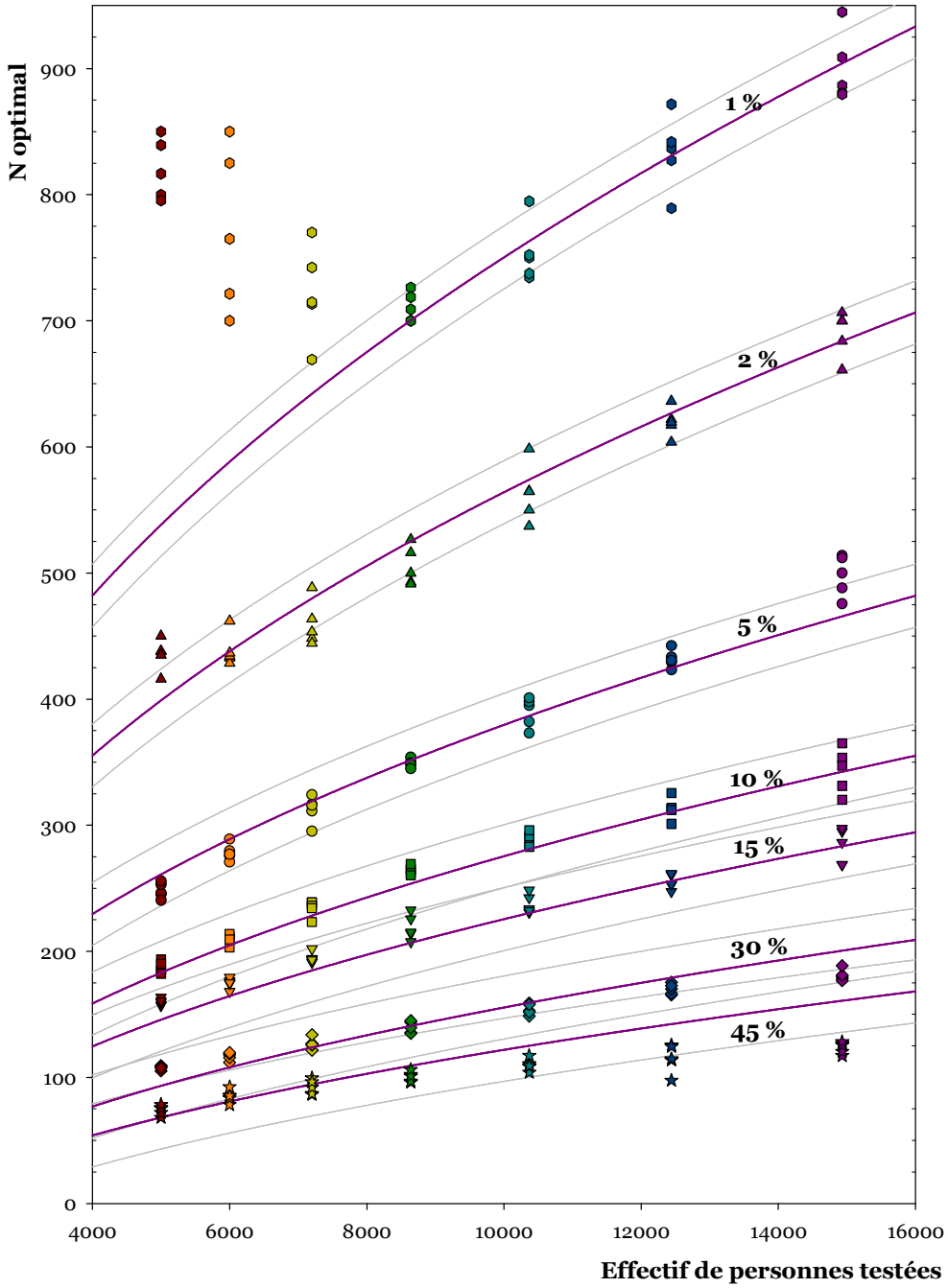
$N_{optimal}$ selon les paramètres d'échantillonnage de cinq EDS récentes

Enquête	Personnes testées (15-49 ans)	Nombre de grappes	Prévalence nationale du VIH	$N_{optimal}$ calculé avec l'Équation 4.7	$N_{optimal}$ calculé sur 100 simulations
Burkina Faso 2003	7 244	400	1,8 %	502	506
Cameroun 2004	9 900	466	5,5 %	363	352
Kenya 2003	6 001	400	6,7 %	250	246
Tanzanie 2003-2004	10 747	350	7,0 %	335	338
Ghana 2003	9 144	412	2,2 %	518	501

La Figure 4.27 montre comment le modèle s'ajuste aux données de la Figure 4.26 pour différents niveaux de prévalence nationale. Dans la majorité des cas, l'écart entre les données de la Figure 4.26 et les valeurs de $N_{optimal}$ prédites par l'Équation 4.7 sont inférieures à 25 (points situés à l'intérieur des bandes délimitées par les courbes grises).

[35] Nous avons utilisé la médiane par groupes pour déterminer les valeurs de $N_{optimal}$.

Figure 4.27
Valeurs de $N_{optimal}$ selon l'effectif de personnes testées, pour différentes prévalences nationales, selon le modèle et les données calculées



Légende : pour la signification des symboles et des couleurs, voir Figure 4.26 page 239. Les courbes violettes représentent les valeurs de $N_{optimal}$ prédites par l'Équation 4.7, selon l'effectif de personnes testées, pour différentes valeurs de prévalence nationale, indiquée en chiffre gras, et un nombre de grappes égales à 432. Les courbes grises encadrent les courbes violettes à +25 et -25. Note : les courbes calculées avec une valeur de 300 pour les grappes sont très proches puisqu'elles sont multipliées par un facteur de 0,987 par rapport à celles avec une valeur de 432. Avec une valeur de 622 grappes, elles sont multipliées par 1,0136.

4.11 Prise en compte du milieu de résidence

Dans la mesure où de nombreux phénomènes présentent un différentiel marqué entre milieu urbain et milieu rural (voir Figure 3.11 page 152), il semble opportun de tenir compte des agglomérations urbaines dans l'estimation des prévalences. La prévalence des grappes n'appartenant pas à une agglomération urbaine sera alors calculée exclusivement à partir de grappes hors agglomérations ; à l'inverse, celle de grappes situées dans une agglomération reposera uniquement sur des grappes de la même agglomération. Nous appellerons cette approche *méthode NU*.

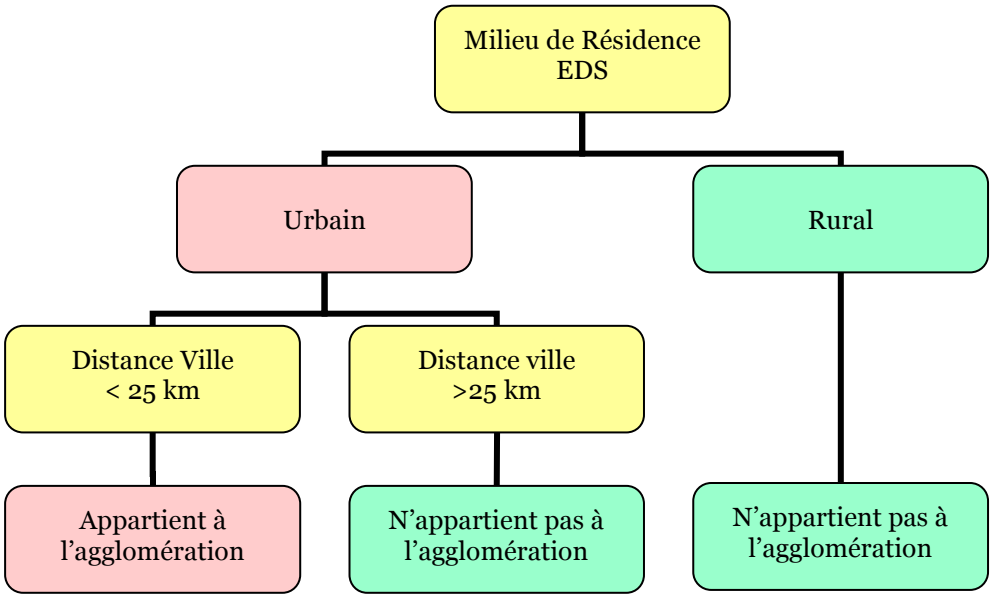
Dans les EDS, l'appartenance d'une grappe à une agglomération urbaine n'est pas renseignée (à l'exception des capitales usuellement considérées comme une région autonome). D'autre part, la définition du milieu de résidence des grappes est telle que des grappes isolées (correspondant à de petites villes et/ou à des chefs-lieux de subdivisions administratives) sont considérées comme urbaines (voir la Figure 4.8 page 211).

Pour déterminer l'appartenance d'une grappe à une agglomération urbaine, nous avons eu recours aux coordonnées longitude/latitude du point central des principales villes fournies par le projet GRUMP[36] (CENTER FOR INTERNATIONAL EARTH SCIENCE INFORMATION NETWORK (CIESIN) OF COLUMBIA UNIVERSITY 2004a). L'appartenance d'une grappe à une agglomération urbaine sera déterminée par sa proximité avec le point central de la dite agglomération. Pour Alicante, nous avons considéré qu'une grappe appartenait à une agglomération urbaine si elle était codée urbaine et située à moins de 25 kilomètres du point central de l'agglomération. Si elle est située à plus de 25 kilomètres ou si elle est codée rurale, elle sera considérée comme n'appartenant pas à l'agglomération considérée (voir Figure 4.29).

La valeur de 25 kilomètres représente une valeur seuil. Elle ne correspond pas à la taille des agglomérations. Il s'agit d'un paramètre permettant de distinguer les grappes urbaines d'une agglomération des grappes urbaines, situées dans le voisinage mais n'appartenant à la dite agglomération. Cette valeur peut être modifiée selon le contexte de chaque enquête. Pour Alicante, la valeur de 25 kilomètres était indiquée en raison de la grande superficie de l'agglomération A. Pour le Burkina Faso et le Cameroun (voir section 4.15 page 255), nous avons utilisé une valeur de dix kilomètres, plus appropriée pour ces deux pays.

[36] Ces données peuvent être téléchargées gratuitement à l'adresse suivante : http://sedac.ciesin.columbia.edu/gpw/.

Figure 4.28
Détermination de l'appartenance à une agglomération urbaine



Il reste à déterminer quelles agglomérations seront retenues pour notre analyse. En effet, lorsqu'une agglomération urbaine est sélectionnée pour la méthode NU, les prévalences des grappes de cette agglomération ne seront estimées qu'à partir des grappes de la même agglomération. Certaines petites agglomérations peuvent avoir été faiblement enquêtées et ne correspondre qu'à une seule ou deux grappes (notamment les petits chefs lieux ruraux isolés). Lorsque nous sélectionnons une agglomération urbaine pour la méthode RNU, nous posons alors une hypothèse anticipatrice en considérant que la prévalence observée à partir de l'EDS dans cette agglomération est un bon indicateur de la prévalence réelle de cette dernière. Le Tableau 4.3 présente les prévalences observées pour les principales agglomérations urbaines d'Alicante.

Tableau 4.3
Prévalence observée des principales agglomérations urbaines d'Alicante

Agglomération	Nombre de grappes	Effectif	Prévalence observée
A	24	489	21,2 %
I	18	352	23,3 %
C	17	346	15,2 %
F	7	151	18,0 %
E	3	48	16,3 %
H	2	32	10,3 %

Le nom des agglomérations fait référence à la Figure 4.7 page 210. Une grappe est considérée comme appartenant à une agglomération urbaine si elle est considérée comme urbaine dans l'enquête et si elle est située à moins de 25 kilomètres du centre de l'agglomération considérée.

Ce tableau va nous permettre de poser nos différentes hypothèses. Les agglomérations A, I, C et F présentent des effectifs de personnes testées importants, répartis en plusieurs grappes. Nous pouvons donc raisonnablement les sélectionner car les effectifs seront suffisants pour réduire les erreurs aléatoires.

Concernant les agglomérations E et H, les effectifs sont trop faibles pour que nous puissions, sur la seule base des données d'enquêtes, décider si les prévalences qui y sont observées traduisent la réalité de l'épidémie au niveau de ces villes ou s'il s'agit d'erreurs aléatoires. Sans information supplémentaire, il est donc préférable de ne pas les sélectionner pour la méthode NU. La prévalence de ces grappes sera alors estimée à partir des grappes voisines comme dans le cas de la méthode N.

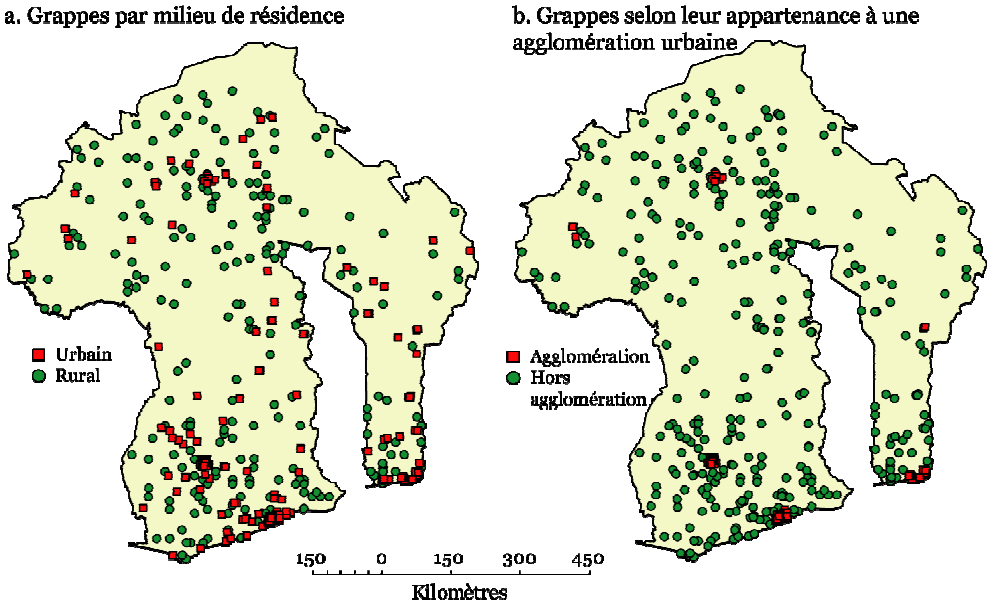
Cependant, nous disposons dans certaines situations d'informations complémentaires qui peuvent nous aider pour poser nos hypothèses (rappelons que le Postulat 2.6 (page 100) nous invite à prendre en compte l'information maximale disponible. Nous avons montré à la section 3.4.2 que la surveillance sentinelle des femmes enceintes pouvait être considérée comme un indicateur de l'ordre de grandeur de la prévalence de la population adulte à un niveau local. Pour les agglomérations urbaines qui disposent de données de surveillance sentinelle, nous pouvons alors comparer leurs résultats avec les prévalences observées, afin de déterminer si la prévalence observée d'une agglomération donnée traduit plutôt la réalité locale de l'épidémie ou des erreurs aléatoires d'échantillonnage. Nous présenterons plus loin (section 4.15 page 255) comment nous avons utilisé les données de surveillance sentinelle pour sélectionner les agglomérations urbaines du Burkina Faso et du Cameroun pour leur prise en compte par la méthode NU.

Nous ne disposons pas de données de surveillance sentinelle pour Alicante. Mais nous pouvons les simuler en utilisant les données de prévalence réelle à notre disposition. Concernant l'agglomération E, les prévalences observées dans la simulation d'enquête, bien que portant sur un petit effectif, s'avèrent proches des prévalences réelles. Il est alors pertinent de retenir cette agglomération pour notre analyse afin d'améliorer notre estimation. L'agglomération H présente une prévalence observée plus faible que la prévalence réelle (qui est de l'ordre de 15 %). Cependant, lorsque nous utilisons la méthode N, les prévalences ajustées pour H sont de l'ordre de 5 %. Si nous sélectionnons H pour l'approche NU, les prévalences ajustées y seront de l'ordre de 10 %, ce qui est certes inférieur aux prévalences réelles mais toujours meilleur que les prévalences de l'ordre de 5 % obtenues avec l'approche N. Il est donc préférable de sélectionner également H pour l'approche NU.

La Figure 4.29 présente le résultat de la recodification du milieu de résidence, selon l'algorithme de la Figure 4.28, des grappes de notre simulation exemple.

Figure 4.29

Recodification du milieu de résidence pour déterminer les grappes appartenant à une agglomération urbaine



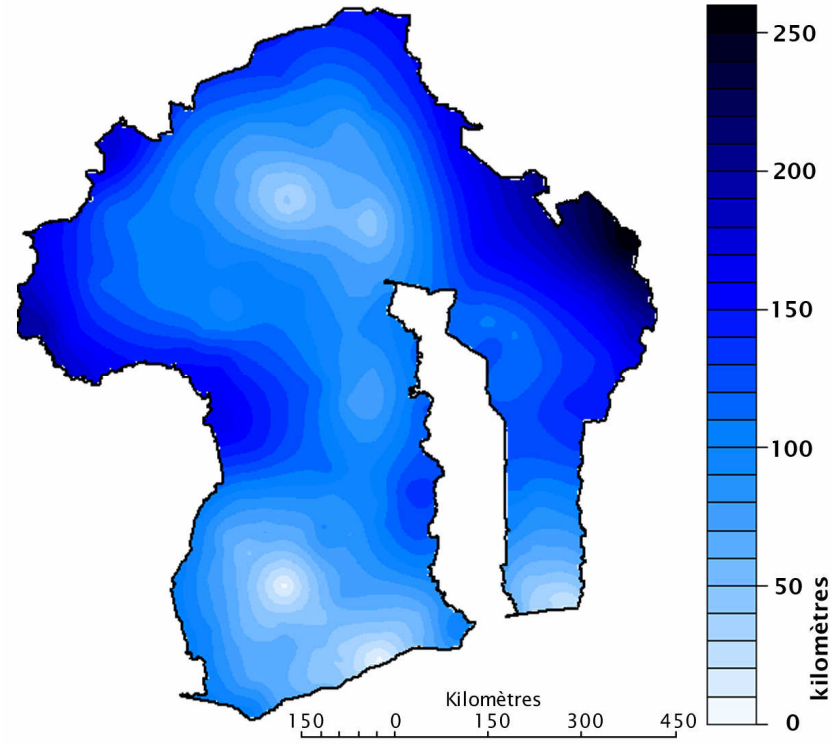
Le milieu de résidence avant recodification correspond à la définition usuelle du milieu de résidence. Les grappes affichées comme urbaines après recodification correspondent aux grappes appartenant à l'une des six agglomérations urbaines sélectionnées (A, I, C, F, E ou H).

Les cartes obtenues avec la méthode NU seront présentées plus loin, à la section 4.14 page 251. Nous avons pris le parti de noter U le nombre d'agglomérations urbaines retenues pour la méthode NU ce qui permet de distinguer des cartes produites en prenant en compte un nombre d'agglomérations urbaines différentes. La méthode N s'avère donc être un cas particulier de la méthode NU, lorsque U est nul.

4.12 Réintégration du paramètre R

Le recours à un effectif minimum comme paramètre pour déterminer la taille des cercles permet à la fois de s'assurer que la prévalence de chaque grappe sera estimée sur un nombre de personnes suffisant et induit un niveau de lissage différent selon les zones enquêtées ; ce dernier étant déterminé par la superficie du cercle de lissage. Ainsi, dans les régions denses et urbanisées, le rayon des cercles est faible (Figure 4.30) et l'information représentée relativement localisée. Par contre, pour les zones peu enquêtées, notamment le long des frontières, le calcul des prévalences fait intervenir des grappes très éloignées les unes des autres.

Figure 4.30
Rayon des cercles de lissages pour N=250 (simulation exemple)

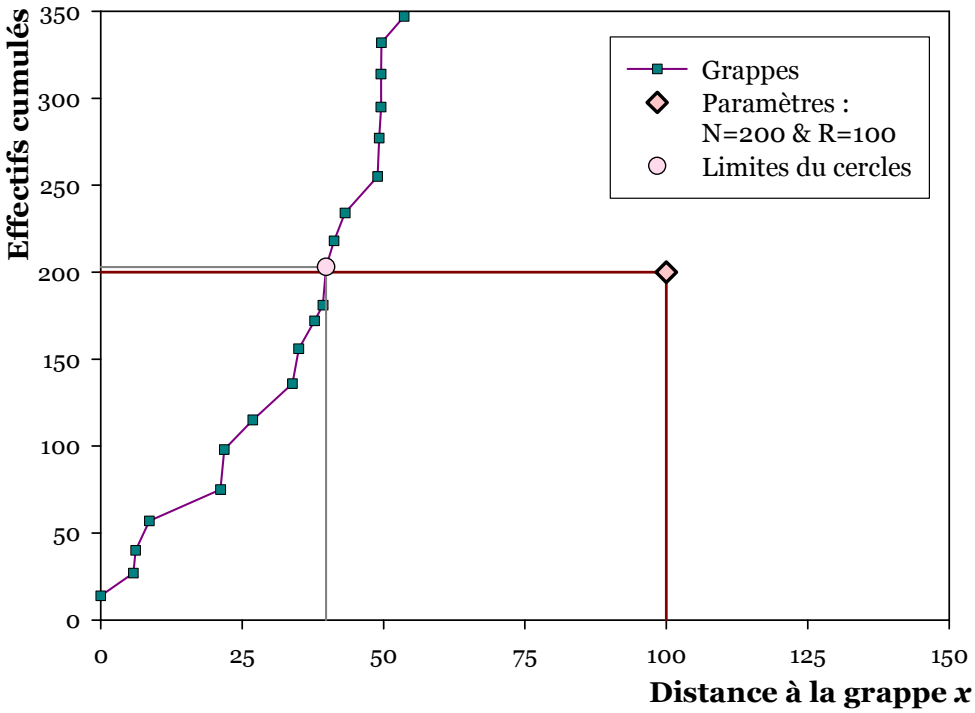


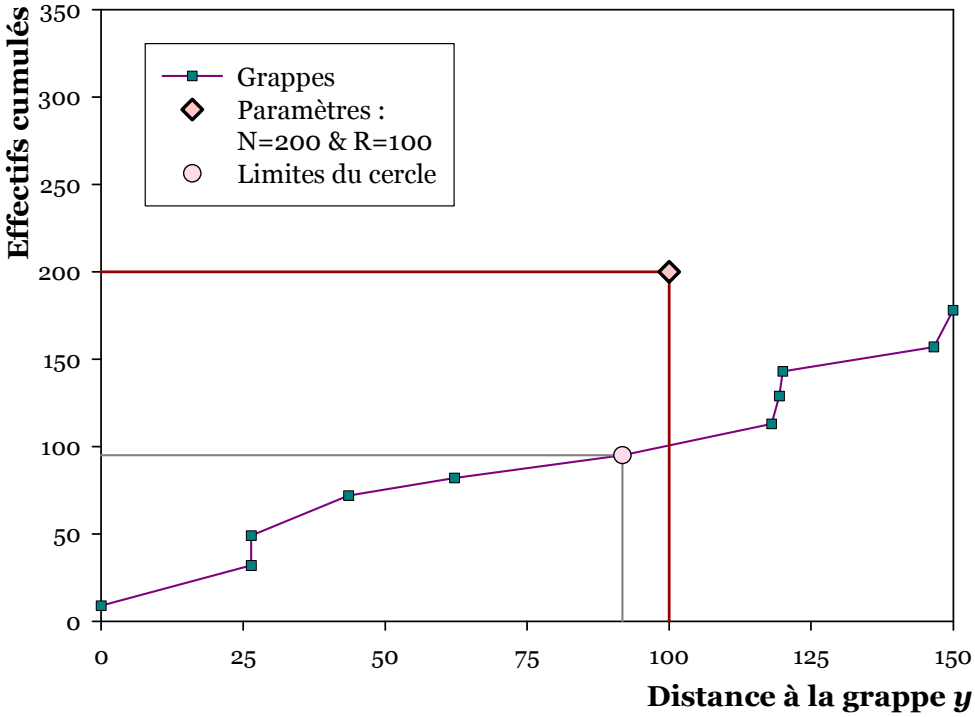
Note : carte obtenue par interpolation spatiale (krigeage ordinaire) du rayon des cercles.

Tableau 4.4
Quantiles du rayon des cercles de lissage pour N=500 (simulation exemple)

Quantile	50 %	75 %	80 %	85%	90 %	95 %	99 %
Valeur (km)	67	100	106	113	128	148	191

Figure 4.31
Sélection des grappes retenues pour l'estimation de la prévalence d'une grappe, méthode RN (deux exemples)





Nous pouvons envisager, pour ces grappes particulièrement isolées, qu'il est préférable de réduire la taille des cercles de lissage quitte à ce que les prévalences

soit estimées sur un nombre d'individus moindre. Nous allons donc réintroduire le paramètre R qui sera le rayon maximum des cercles de lissage (méthode RN). Si l'effectif N est atteint pour un rayon inférieur à R (grappe *x* sur la Figure 4.31), la prévalence estimée sera la même que celle obtenue avec la méthode N. Si l'effectif de personnes testées à l'intérieur du cercle de rayon R est inférieur à N (grappe *y*), alors la prévalence sera estimée sera celle obtenue par la méthode R.

Comme il s'agit de ne toucher que les grappes isolées, nous avons décidé d'utiliser comme paramètre R la valeur du 9^{e} décile du rayon des cercles de lissage lorsque la méthode N est utilisée (voir Tableau 4.4), soit 128 kilomètres pour notre simulation exemple. D'autres valeurs de R pourraient être utilisées. Avec une valeur plus élevée, nous affecterons les estimations de prévalence d'un nombre de grappes moindre. Le paramètre R a relativement peu d'impact sur la cartographie finale obtenue puisque les zones les moins enquêtées sont affectées. Nous avons réalisé plusieurs centaines de simulations d'EDS et regardé s'il existait une valeur de R (une fois N fixé) qui minimisait l'écart moyen. Si, pour la majorité des estimations, l'ajout du paramètre R permettait de réduire légèrement l'écart moyen entre prévalences réelles et prévalences estimées, cette réduction restait très légère. De fait, aucune valeur optimum pour le paramètre R ne se dégageait en procédant ainsi.

L'ajout ou non du paramètre R permet simplement de tester deux hypothèses concernant les zones faiblement enquêtées. En l'absence de ce paramètre, la prévalence de ces régions est fortement lissée à partir des valeurs des régions voisines. Lorsque R est appliquée, elle et plus fortement dépendante des quelques observations effectuées localement. En raison de la faible quantité de données dans ces régions, nous ne pourrons y interpréter les résultats estimés de manière précise. Lorsque R sera appliqué, nous ne pourrons déterminer si les variations locales que nous observerons correspondent à une réalité des variations locales de l'épidémie dans ces régions ou bien s'il s'agit de variations aléatoires dues à l'échantillonnage. Cependant, cela pourra constituer des pistes de recherche à investiguer. Un exemple plus concret sera donné à la section 4.15 page 255 à partir des résultats obtenus avec les données de l'EDS 2004 du Cameroun.

Nous parlerons de méthode NRU lorsque nous prendrons en compte simultanément en compte les paramètres N, R et U. Les approches dites N, R, NR, NU, etc. peuvent être considérées comme des cas particuliers de la méthode NRU, obtenues lorsque N ou R est égal à l'infini et U nul. Par exemple, la méthode N correspond à la méthode NRU avec R égal à l'infini et U égal à 0.

4.13 Élaboration d'un indicateur de qualité et cartes complémentaires

La qualité des estimations produites n'est pas constante à travers le pays. Dans les zones peuplées, l'information est précise du fait d'une bonne densité des grappes. Les variations infrarégionales du phénomène étudié sont alors visibles. A l'opposé, dans les zones où le nombre de personnes enquêtées est plus faible, seule une tendance régionale peut être mise en évidence. Enfin, pour les zones non enquêtées, les résultats doivent être interprétés avec prudence, les variations estimées résultant d'une interpolation à partir des zones voisines. Cette irrégularité de la qualité des estimations produites résulte, d'une part, de l'échantillonnage des EDS, conçues pour être représentatives des populations et non des territoires et, d'autre part, de l'approche méthodologique employée, à savoir un lissage adaptatif selon les zones (superficie des cercles variable).

Afin de faciliter la lecture et l'interprétation des cartes obtenues, il est donc pertinent de connaître les zones pour lesquelles les observations sont importantes. Nous pouvons alors représenter le nombre de personnes testées par grappe à l'aide de cercle dont la superficie est proportionnelle aux effectifs. En raison de la densité des grappes dans certaines zones, les cercles se superposent. Nous avons donc appliqué une couleur avec transparence. Les zones claires correspondent alors aux régions peu ou pas documentées et les zones foncées (du fait de la superposition des cercles) à celles où le nombre d'observations est important (voir la Figure 4.33.b page 252).

L'indicateur statistique utilisé le plus souvent en épidémiologie pour déterminer la qualité d'une mesure est l'intervalle de confiance. Nous pouvons calculer pour chaque grappe un intervalle de confiance à 90 ou 95% de la prévalence estimée. Cependant, cartographier l'amplitude de cet intervalle de confiance serait peu informatif. D'une part, cet indicateur ne prendra pas en compte la taille des cercles de lissage. Or, plus les cercles sont petits, meilleure est l'estimation. D'autre part, l'amplitude de ces intervalles est déterminée à la fois par l'effectif et le niveau de la prévalence estimée de la grappe. De ce fait, une représentation graphique de l'ampleur de l'intervalle de confiance traduira plus les variations spatiales de la prévalence estimée plutôt que la qualité des estimations réalisées.

Un indicateur plus pertinent devra prendre en compte à la fois le rayon des cercles de lissage et l'effectif de personnes testées sélectionnées, pour chaque grappe, pour l'estimation de la prévalence. Nous avions utilisé, pour la Figure 4.26 page 239, un indicateur de qualité globale correspondant au carré du rayon moyen divisé par la racine carrée de $N_{optimal}$. Nous pouvons donc élaborer un indicateur de qualité de même forme pour chaque grappe en prenant le rapport du carré du rayon du cercle de lissage de la grappe par la racine carrée de l'effectif de personnes testées sélectionnées pour l'estimation de la prévalence de la dite grappe (Équation 4.8).

Équation 4.8
Indicateur de qualité de la prévalence estimée des grappes

$$IQ_{grappe} = \frac{R^2_{grappe}}{\sqrt{N_{grappe}}}$$

IQ_{grappe} *: indicateur de qualité d' une grappe*
R_{moyen} *: rayon du cercle de lissage pour cette grappe*
$N_{optimal}$ *: nombre de personnes testées sélectionnées pour l' estimation de la prévalence*

Cet indicateur de qualité peut ensuite être cartographié par krigeage ordinaire (voir la Figure 4.33.a page 252). Une valeur faible de cet indicateur (couleurs foncées) correspondra à des zones où les estimations sont à la fois réalisées à partir d'un nombre suffisant d'individus et localisées (petit rayon du cercle de lissage). À l'inverse, les zones claires (valeurs élevées de l'indicateur de qualité) doivent être interprétées avec prudence, la faiblesse des observations induisant des petits effectifs et des rayons importants.

4.14 L'épidémie d'Alicante reconstituée

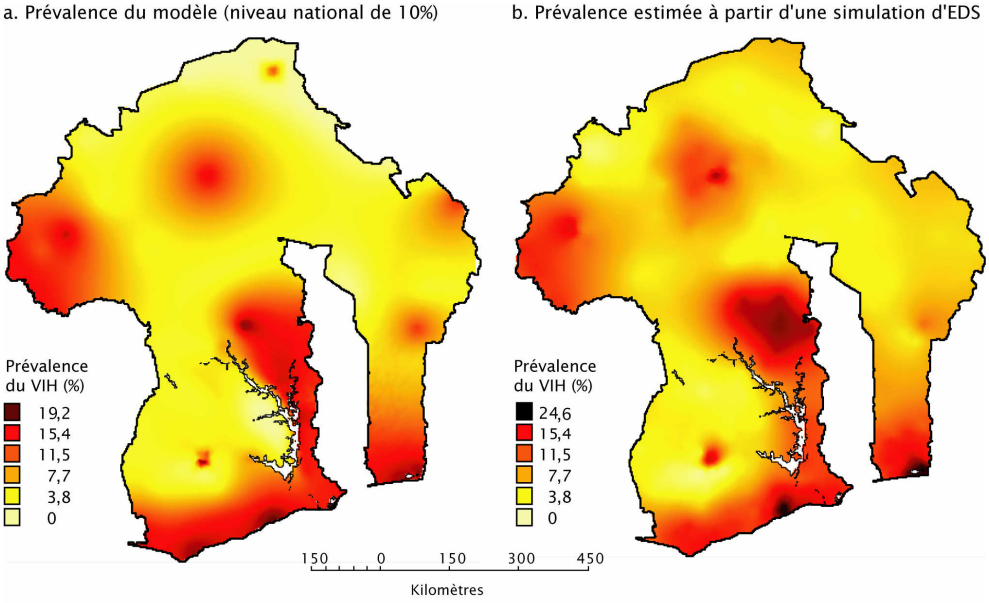
Les cartes a et b de la Figure 4.32 sont présentées avec la même échelle colorimétrique. Cependant, le maximum de la prévalence estimée (carte b) est supérieur à celui du modèle de départ (carte a), traduisant des variations plus marquées.

Globalement, les principales tendances du modèle ont été reconstituées. Apparaissent ainsi le gradient de la côte sud vers le nord et celui de la frontière ouest vers l'intérieur du pays. Les agglomérations A et I présentent un contraste plus marqué par rapport à l'épidémie de départ. Par contre, les deux petites agglomérations situées le long de la côte (B) ne se distinguent plus de leur voisinage sur la carte des prévalences estimées. Les agglomérations C et F sont clairement visibles. C présente toujours une prévalence concentrée au niveau de la ville, bien que moins marquée, et le gradient autour de F est également rendu, bien que plus irrégulier.

De part et d'autre du grand lac, où une rupture nette avait été introduite dans le modèle, la prévalence a été surestimée à l'ouest et sous-estimée à l'est. L'estimation ne prenant pas en compte les frontières naturelles et considérant la surface géographique comme un continuum, les prévalences estimées sont uniformisées aux petites échelles.

Figure 4.32

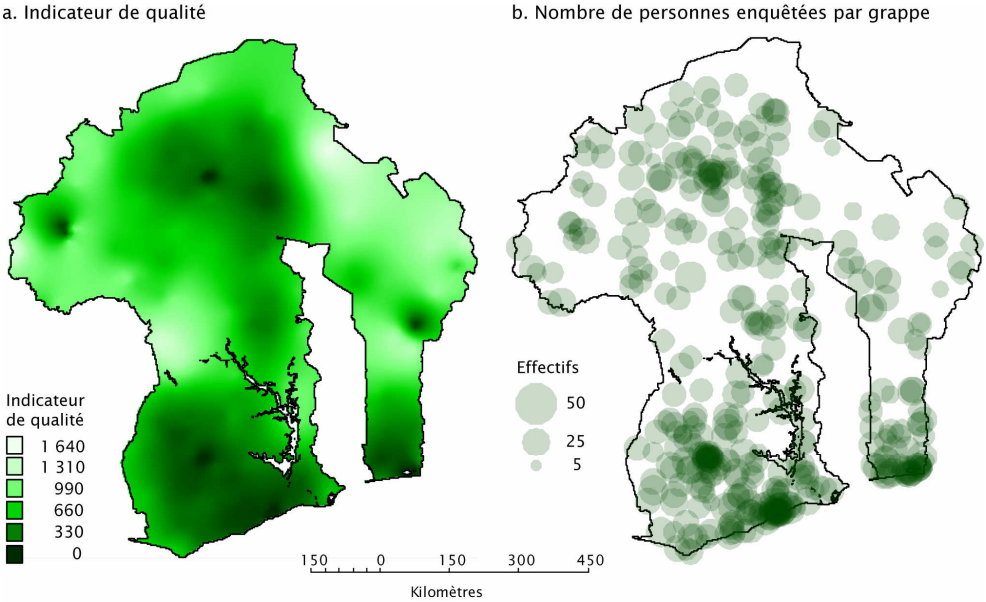
Prévalence du modèle et prévalence estimée par la méthode NRU pour la simulation exemple



Paramètres de la simulation : 8.000 personnes testées, réparties en 400 grappes avec une prévalence nationale de 10%.
Paramètres de l'estimation : N=250, R=128 km, U=6 (agglomérations A, I, C, F, E et H)

Figure 4.33

Indicateur de qualité et nombre de personnes testées par grappe pour la simulation exemple



Paramètres de la simulation : 8.000 personnes testées réparties en 400 grappe avec une prévalence nationale de 10%.
Paramètre de l'estimation : N=250, R=128 km, U=6.
Note : l'indicateur de qualité est calculé pour chaque grappe selon la formule $r^2/\sqrt{n}$ où r est le rayon du cercle de lissage et n le nombre de personnes sur lesquelles la prévalence estimée de la grappe a été calculée.

La zone de fortes prévalences autour de D est également reconstituée, avec des écarts plus accentués par rapport au modèle. La ville D elle-même ne se distingue plus par rapport à son voisinage. Dans la mesure où il s'agit d'une petite agglomération, peu enquêtée dans cette simulation donnée, elle n'a pas été retenue pour le paramètre U, et les prévalences estimées sont celles de la tendance locale (agglomération et voisinage confondus).

Plusieurs éléments ne transparaissent plus sur la Figure 4.32.b. Tout d'abord le pic local en milieu rural au nord du pays n'apparaît pas, du fait de l'absence de point de mesure dans cette zone (Figure 4.33.b).

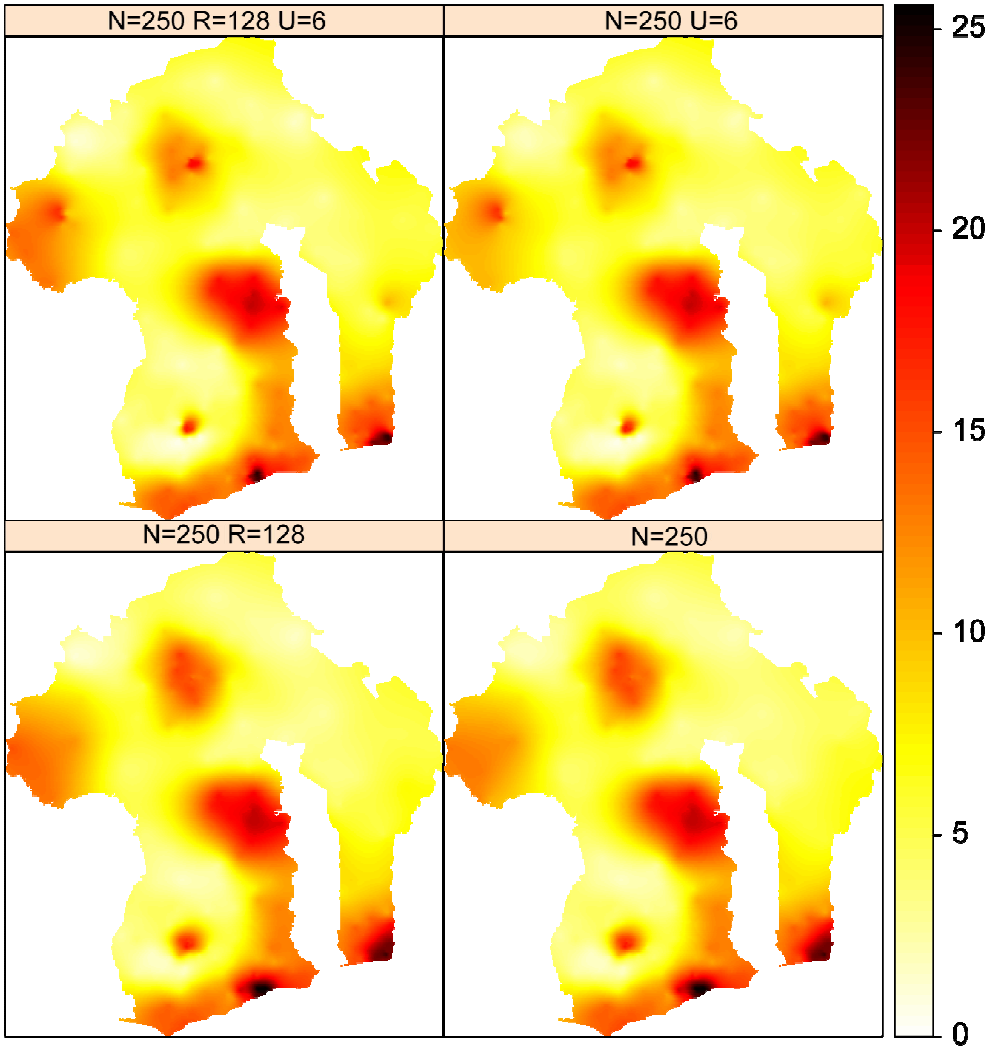
Les gradients autour de H et depuis la frontière ne sont guère visibles. Le nombre de grappes dans ces deux zones est faible (Figure 4.33.b) et les valeurs de l'indicateur de qualité élevées. Le nombre de personnes enquêtées n'est pas suffisant pour reconstituer les variations locales et seule une tendance régionale peut être estimée (effet d'uniformisation).

Les cartes a et b de la Figure 4.33 présentent de par leur construction des similitudes. Il est alors possible de distinguer les zones non enquêtées où les prévalences estimées sont la poursuite des tendances des zones voisines ; les zones faiblement enquêtées où seule des variations régionales sont reconstituées, les variations locales étant alors uniformisées ; et les zones où l'information est suffisamment importante pour révéler des différentiels locaux.

La Figure 4.34 permet de comparer les approches N, NR, NU et NRU et de voir l'effet sur la cartographie obtenue de l'ajout des paramètres U et R. L'ajout du paramètre R n'influe quasiment pas sur les cartes obtenues dans le cadre de la simulation exemple. Nous verrons plus loin son effet sur les cartes réalisées à partir des données de l'EDS 2003 du Burkina Faso et de l'EDS 2004 du Cameroun (section 4.15 page 255).

Le paramètre U, en revanche, a induit une concentration des pics épidémiques locaux au niveau des agglomérations urbaines retenues. L'épidémie estimée est ainsi plus concentrée au niveau de l'agglomération C tout en conservant la diffusion observée autour de F. Le pic de l'agglomération H qui, en raison de la faiblesse des observations locales, n'apparaît pas avec la méthode N simple, devient visible. Il ressort ainsi que la prise en compte des principales agglomérations urbaines (paramètre U) permet, pour celles de taille importante, de mettre à jour la variation du phénomène étudié à leur voisinage.

Figure 4.34
Comparaison des approches N, NR, NU et NRU (simulation exemple)



*
**

Globalement, cette approche méthodologique a permis de réduire les erreurs aléatoires et de dégager les tendances nationales et régionales, avec un lissage plus ou moins important selon la quantité de données locales, bien que ce soit au détriment d'une estimation fine des niveaux locaux. L'analyse des différentiels est valable à des échelles régionales voire infrarégionales lorsque la densité de données est importante. Cependant, il est nécessaire de tenir compte de l'espacement des grappes sur le territoire et il serait vain d'interpréter des différences se situant à des niveaux d'échelle inférieurs. Les estimations réalisées sont d'autant plus précises que le nombre total de personnes testées est important et que la prévalence du phénomène étudié se rapproche de 50%. À l'opposé, un effet d'uniformisation important s'observe dans les zones faiblement peuplées.

4.15 Application à deux pays : le Burkina Faso et le Cameroun

Nous avons appliqué cette approche aux données de prévalence du VIH (15-49 ans), fournies par les dernières EDS du Burkina Faso (2003) et du Cameroun (2004). La valeur seuil utilisée pour recoder les grappes urbaines a été 10 kilomètres. Les résultats restent néanmoins semblables pour des valeurs seuils de 15 ou 20 kilomètres. Le Tableau 4.5 présente les données ayant présidé à la sélection des agglomérations urbaines pour le paramètre U.

Nous avons sélectionné d'une part, les villes où le nombre de personnes testées est relativement important et réparties sur plusieurs grappes. Dans le cas présent, le choix a été de retenir les agglomérations où le nombre de personnes testées étaient supérieures à 150 au Burkina Faso et à 100 au Cameroun, réparties sur au moins 6 grappes.

D'autre part, des agglomérations où, même si le nombre de personnes testées est relativement faible, la prévalence observée dans l'EDS est proche de celle observée en clinique prénatale. C'est le cas de Bamenda, Bertoua, Kumba, Edéa et Ebolowa au Cameroun.

Enfin, des agglomérations où la prévalence observée dans l'EDS, bien qu'inférieure à celle mesurée en clinique prénatale, reste supérieure à celle estimée pour une valeur nulle de U. C'est le cas de Ouahigouya, Garoua et Maroua. Nous avons considéré en effet que, dans la mesure où nous avons montré que la prévalence observée en clinique prénatale pouvait être un indicateur local de la prévalence en population générale (section 3.4.2 page 183 et suivantes), il s'agissait d'hypothèses raisonnables.

Pour le paramètre N, nous avons utilisé 500 pour le Burkina Faso et 350 pour le Cameroun, conformément aux résultats du Tableau 4.2 page 241 et de la Figure 4.25 page 237 (valeurs arrondis à la dizaine la plus proche).

La valeur de R a été déterminée en prenant le neuvième décile du rayon des cercles de lissage lorsque seul N est appliqué (voir section 4.12 page 247). R vaut ainsi 117 kilomètres pour les EDS du Burkina Faso et du Cameroun (cette égalisé n'est qu'une simple coïncidence, la valeur inférieure de N au Cameroun étant compensée par la superficie plus grande du pays).

Pour l'habillage des cartes, nous avons utilisé les principales routes de chaque pays fournies par la base ArcAtlas de la société ESRI[37].

[37] http://arcdata.esri.com/data_downloader/DataDownloader?part=10200

Tableau 4.5

Sélection des agglomérations urbaines du Burkina Faso et du Cameroun pour le paramètre U

Aggl. urbaine	Population†	EDS‡				Surveillance sentinelle§			Inclue pour U
		Préval. du VIH (%)	IC 95 %	N	Nbre de grappes	Préval. du VIH (%)	IC 95 %	N	
Cameroun									
Douala	1 494 700	4,4	3,2-6,0	931	43	8,0	5,6-11,2	400	Oui
Yaoundé	1 248 200	8,5	6,8-10,6	870	45	7,2	5,1-10,0	471	Oui
Garoua	356 900	5,3	1,9-12,6	93	6	8,0	5,6-11,2	402	Oui
Bamenda	316 100	10,6	5,0-20,4	75	4	10,1	7,0-14,4	286	Oui
Maroua	271 700	6,2	2,3-14,5	80	3	7,3	4,7-11,0	300	Oui
Bafoussam	242 000	9,6	5,5-15,9	146	6	5,9	3,1-10,6	186	Oui
N'Gaoundéré	189 800	11,5	7,4-17,2	183	8	11,4	8,5-15,0	395	Oui
Bertoua	173 000	9,4	4,8-16,9	107	4	9,0*	5,3-14,7	166*	Oui
Loum	141 400	8,4	3,9-16,6	91	4	n. d.	n. d.	n. d.	Non
Kumba	125 600	7,2	3,2-14,9	96	5	9,8	3,7-22,2	51	Oui
Edéa	122 300	9,5	4,2-19,2	74	5	9,0*	4,5-16,8	100*	Oui
Kumbo	116 500	9,3	2,5-25,8	33	2	n. d.	n. d.	n. d.	Non
Foumban	113 100	5,1	1,3-15,1	59	3	7,3	3,0-15,8	82	Non
Nkongsamba	110 600	6,1	2,8-12,0	131	9	n. d.	n. d.	n. d.	Oui
Mbouda	101 100	5,3	0,3-28,1	19	1	n. d.	n. d.	n. d.	Non
Dschang	87 000	0,0	0,0-12,3	35	2	4,3*	1,9-8,9	164*	Non
Limbé	84 500	8,9	3,3-20,3	56	3	5,6	3,0-10,0	197	Non
Ebolowa	79 500	11,0	5,2-21,1	72	3	11,6	7,7-17,1	198	Oui
Burkina Faso									
Ouagadougou	709 736	3,8	2,1-6,5	346	18	4,0	2,3-7,0	321	Oui
Bobo-Dioulasso	309 711	3,3	1,8-5,7	402	19	4,3	2,9-6,3	630	Oui
Koudougou	72 490	4,3	2,0-8,5	190	7	n. d.	n. d.	n. d.	Oui
Ouahigouya	52 193	2,5	0,4-10,1	73	4	3,6	2,1-5,9	422	Oui
Banfora	49 724	4,5	2,0-9,2	167	10	n. d.	n. d.	n. d.	Oui
Pouytenga	35 720	0,0	0,0-16,6	25	2	n. d.	n. d.	n. d.	Non
Kaya	33 958	0,4	0,0-9,5	50	3	n. d.	n. d.	n. d.	Non
Dédougou	33 815	0,0	0,0-19,2	21	1	n. d.	n. d.	n. d.	Non
Tenkodogo	31 466	2,9	0,1-18,0	31	2	2,6	1,4-4,7	430	Non
Fada N'Gourma	29 254	0,0	0,0-10,9	40	3	1,3	0,5-3,0	455	Non
Dori	23 768	0,0	0,0-17,2	24	2	n. d.	n. d.	n. d.	Non
Gaoua	16 424	0,0	0,0-10,7	41	2	3,1	1,5-6,0	289	Non

n. d. : non disponible – IC : intervalle de confiance

Test exact de Fisher de comparaison de deux proportions : pour l'ensemble des agglomérations, aucune différence significative à 20 % n'est observée entre la prévalence observée dans l'EDS et celle provenant de la surveillance sentinelle des femmes enceintes, à l'exception de Douala où la différence est significative (p-value = 0,01203).

† Cameroun : population en 2001, résultats préliminaires, Cameroon Statistical Yearbook 2004 (NATIONAL INSTITUTE OF STATISTICS 2004) ; Burkina Faso : population en 1996, Recensement de 1996 cité par City Population (CITY POPULATION 2006).

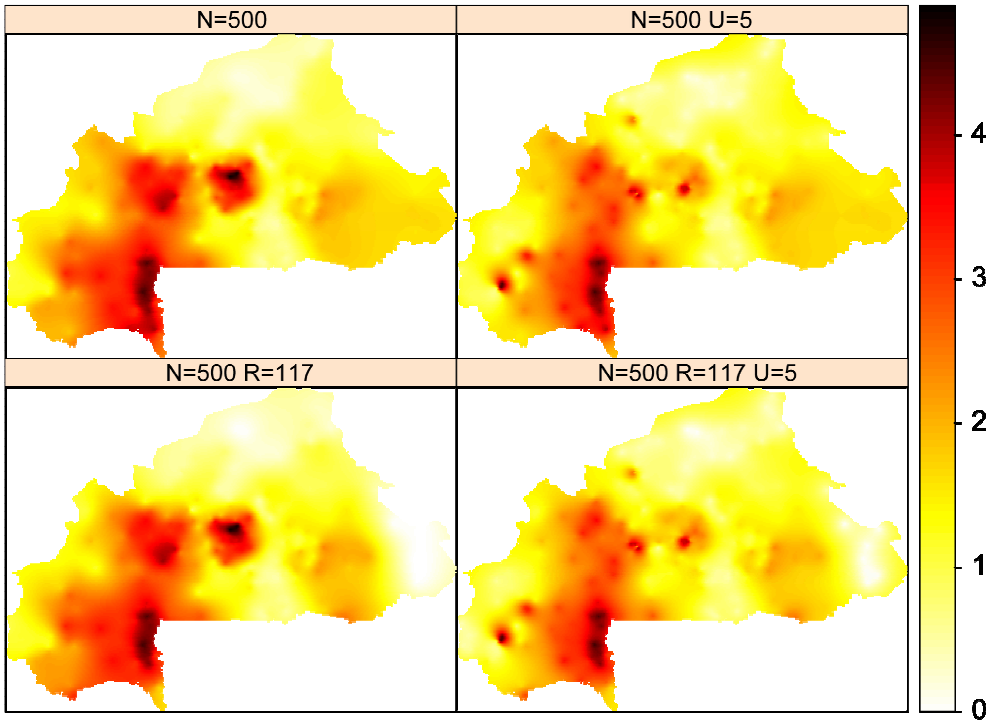
‡ Cameroun: EDS 2004; Burkina Faso : EDS 2003. Une grappe urbaine est considérée comme appartenant à une agglomération si elle est située dans un rayon de dix kilomètres. La pondération des individus est prise en compte pour les calculs (prévalence, intervalle de confiance, test de comparaison).

§ Cameroun : prévalence du VIH1 en 2002 (NATIONAL AIDS CONTROL COMMITTEE 2003) ; Burkina Faso : prévalence du VIH en 2003 (CNSL BURKINA FASO 2003).

* Résultats non disponibles en 2002; les données présentées datent de 2000(NATIONAL AIDS CONTROL COMMITTEE 2001).

Note : le taux d'urbanisation était de 52,8 % en 2003 au Cameroun (NATIONAL INSTITUTE OF STATISTICS 2004) et de 15,5 % en 1996 au Burkina Faso (INSTITUT NATIONAL DE LA STATISTIQUE ET DE LA DÉMOGRAPHIE 2000b).

Figure 4.35
Comparaison des approches N, NR, NU et NRU pour le Burkina Faso en 2003



Au Burkina Faso (Figure 4.36), les foyers d'infections se situent tout d'abord au niveau des agglomérations urbaines situées le long des axes routiers internationaux. Ouagadougou, la capitale, est concernée mais également Koudougou, Bobo-Dioulasso et Banfora sur l'axe reliant Ouagadougou à la Côte d'Ivoire, ou encore Ouahigouya sur un axe routier vers le Mali. L'ajout du paramètre U (Figure 4.35) montre que l'épidémie de Ouagadougou est relativement concentrée, ce qui est également mis en évidence par une cartographie réalisée selon les régions EDS (Figure 4.37).

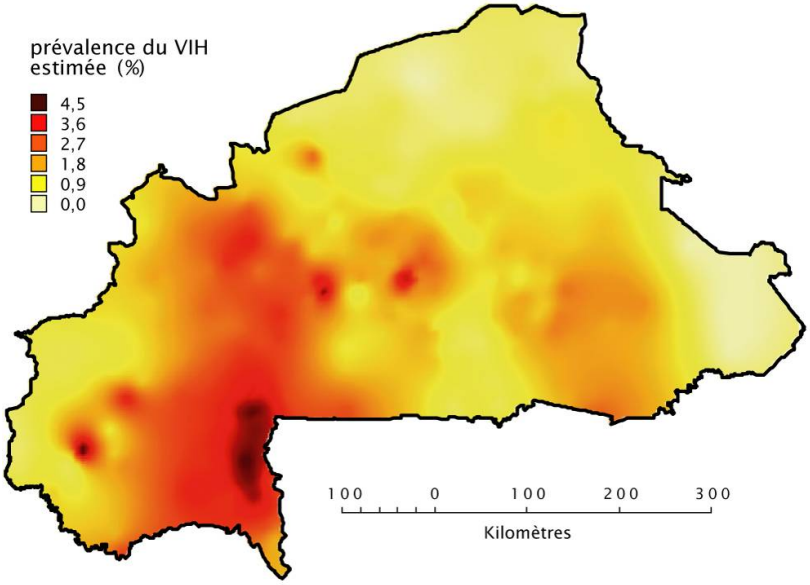
La région autour de Débougou et de Gaoua est particulièrement touchée. Outre sa proximité avec la Côte d'Ivoire et le Ghana, elle est connue comme une zone importante d'orpaillage, impliquant à la fois une présence d'hommes migrants venus seuls et un commerce sexuel non négligeable.

Le nord et l'ouest du pays, zones sahéliennes, présentent une prévalence faible. Étant peu habitées, et de ce fait peu enquêtées, seule une tendance régionale y est mise en évidence. Les variations locales, s'il y en a, ne transparaissent pas sur la carte. Par ailleurs, les pays frontaliers de ces régions ont des prévalences estimées du VIH faibles : 1,8 % pour le Mali, 1,1 % pour le Niger et 2,0 % pour le Bénin fin 2003 selon Onusida (UNAIDS 2006) tandis que celle de la Côte d'Ivoire est estimée à 7,0 %.

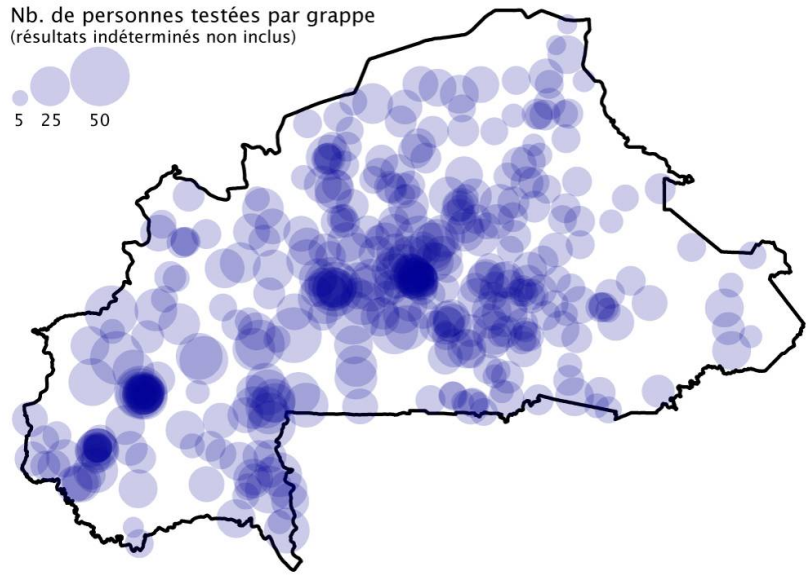
Figure 4.36

Prévalence du VIH estimée au Burkina Faso et cartes complémentaires

a. Prévalence estimée du VIH (15-49 ans)

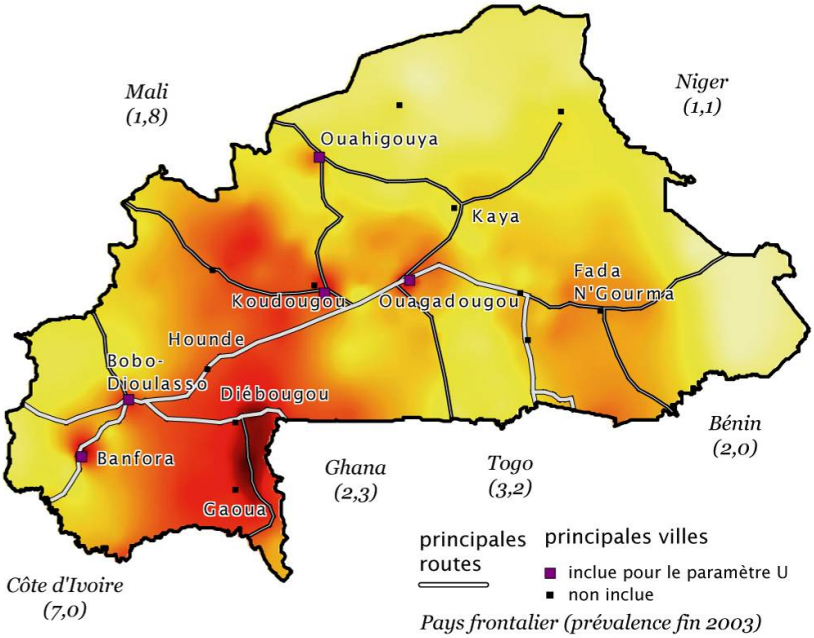


c. Nombre de personnes testées par grappe

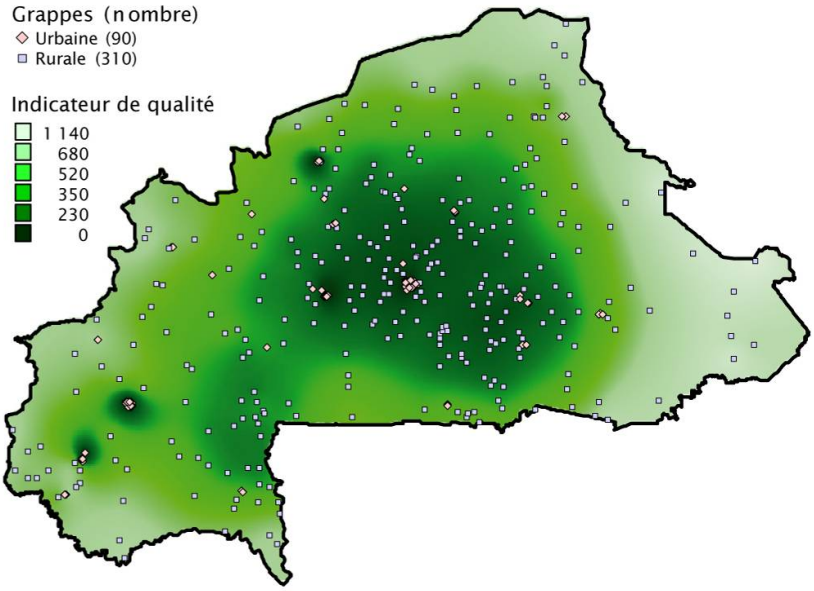


Note 1 : La prévalence nationale du VIH (15-49 ans) est de 1,8% dans l'EDS 2003 du Burkina Faso. 7.244 personnes ont été testées (résultats indéterminés exclus), réparties en 400 grappes.
Note 2 : Les paramètres utilisés pour l'estimation de la prévalence du VIH sont : N=500, R=117 km et U=5 (voir carte b pour les villes retenues).

b. Prévalence estimée du VIH, villes et routes

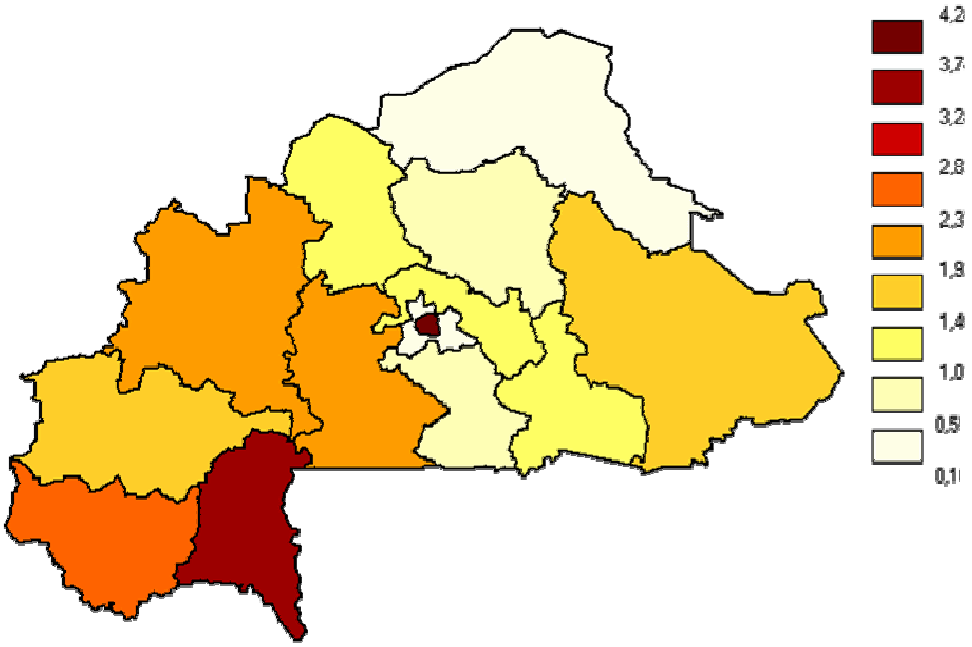


d. Indicateur de qualité et grappes enquêtées



Note 3 : L'indicateur de qualité est calculé pour chaque grappe selon l'expression $r^2/\sqrt{n}$ où r est le rayon du cercle de lissage et n le nombre de personnes testées dans ce cercle.
Sources : EDS 2003 du Burkina Faso pour les données VIH, DCW pour les frontières nationales, GRUMP pour les principales villes, ArcAtlas (ESRI) pour les principales routes, rapport ONUSIDA 2006 pour les prévalences fin 2003 des pays frontaliers.

Figure 4.37
Prévalence du VIH selon les régions EDS au Burkina Faso en 2003



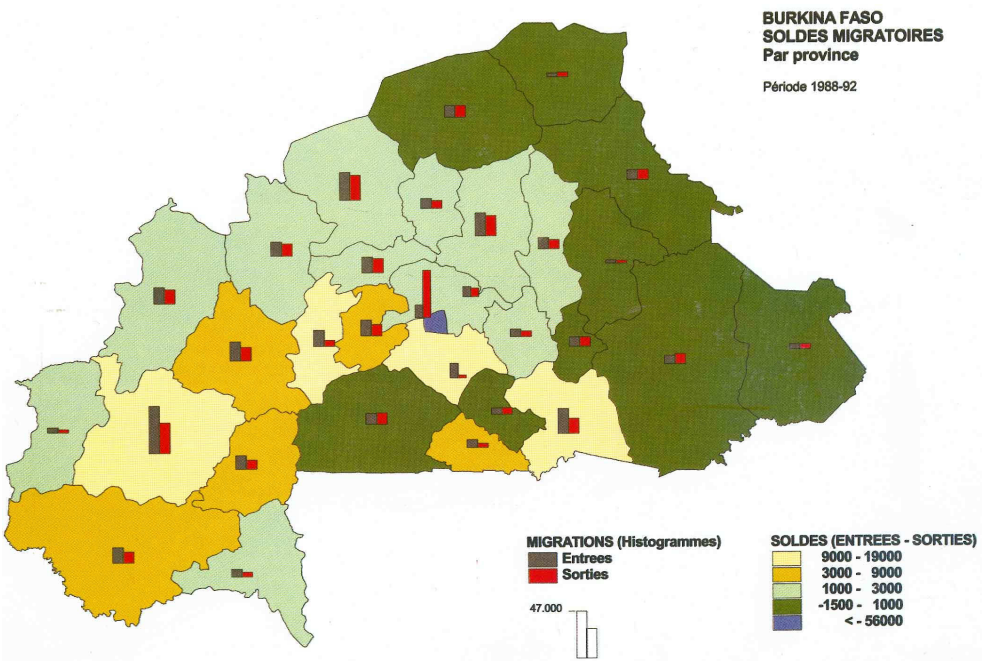
Source : EDS 2003 du Burkina Faso.

Enfin, cette carte présente des similitudes avec celle des zones migratoires (Figure 4.38) sur la période 1998-1992 (RÉSEAU MIGRATIONS ET URBANISATION EN AFRIQUE DE L'OUEST (REMUAO) 1997) et celle des retours de rapatriés burkinabés (Figure 4.39) depuis la Côte d'Ivoire fin 2002, début 2003 (SP/CONASUR 2004). Si ces similitudes doivent être interprétées avec prudence, ne constituant en rien une preuve de cause à effet, elles plaident néanmoins en faveur de la vraisemblance de la carte produite dans la mesure où plusieurs enquêtes ont montré que les migrations pouvaient être associées avec une prévalence du VIH plus élevée (LALOU 1994, QUINN 1994, DECOSAS 1995, LURIE 2003).

Le paramètre R influe peu sur la cartographie produite (Figure 4.35) excepté concernant la pointe Est du pays qui s'avère être faiblement habitée. Sur les sept grappes de cette zone, aucun résultat positif n'a été observé (Figure 4.4 page 206). La carte produite avec le paramètre R s'avère donc plus probable que celle sans ce dernier, puisqu'elle induit une prévalence proche de zéro.

Figure 4.38

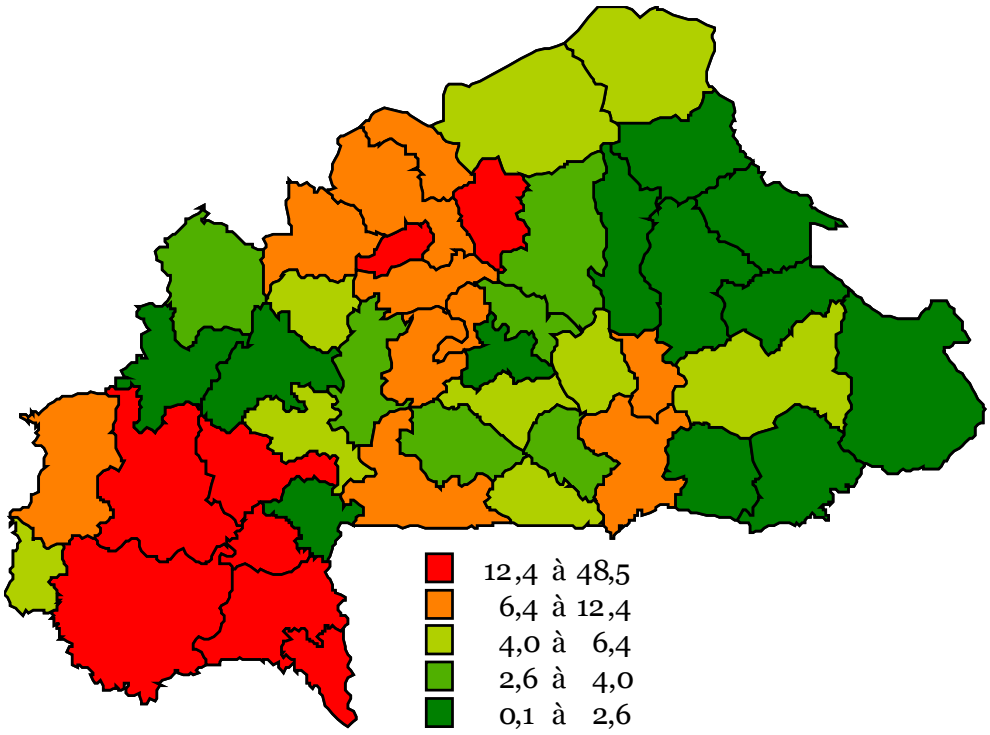
Soldes migratoires par province au Burkina Faso sur la période 1988-1992



Source : Enquête REMUAO (WANE 2000, carte A5).

Figure 4.39

Taux de rapatriés de Côte d'Ivoire (‰) en 2002 au Burkina Faso, par province



Sources : (SP/CONASUR 2004) pour les effectifs de rapatriés par province, (INSTITUT NATIONAL DE LA STATISTIQUE ET DE LA DÉMOGRAPHIE 2004) pour les effectifs de population en 2002.

Bien que nous ayons montré au chapitre précédent que le biais dû aux individus non testés n'influait pas significativement sur l'estimation de la prévalence nationale (section 3.3.4), nous pouvons nous demander si, à un niveau plus local, ce biais peut influer significativement sur la carte obtenue par la méthode des cercles de même effectif.

Nous pouvons, comme pour la prévalence du VIH, utiliser la méthode des cercles pour estimer le taux de non réponse au dépistage du VIH pour chaque grappe puis l'interpoler spatialement par krigeage ordinaire. Nous obtenons alors la Figure 4.40 ci-dessous. Il apparaît que les individus non testés se situent majoritairement en milieu urbain et, en particulier, à Ouagadougou, Koudougou, Bobo-Dioulasso et Banfora, qui sont également les villes les plus touchées du pays. Or, il s'avère que les individus non testés présentent une prévalence plus importante que les personnes enquêtées (voir Tableau 3.13 page 166). De ce fait, si le biais des non testés devait avoir un impact visible sur la carte obtenue, il s'agirait d'une augmentation des contrastes observés, les patterns restant les mêmes.

Figure 4.40

Taux de non testés pour le VIH, EDS 2003 du Burkina Faso, en pourcents

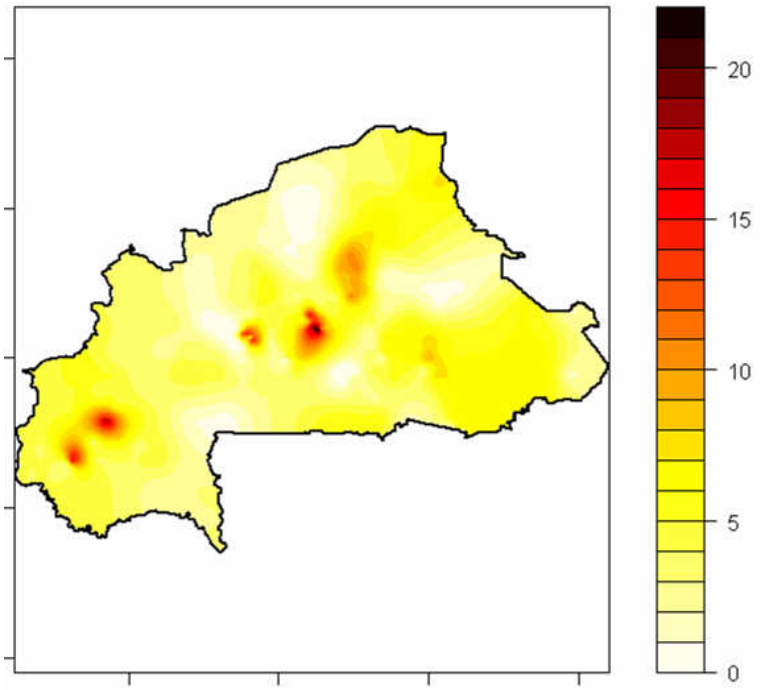
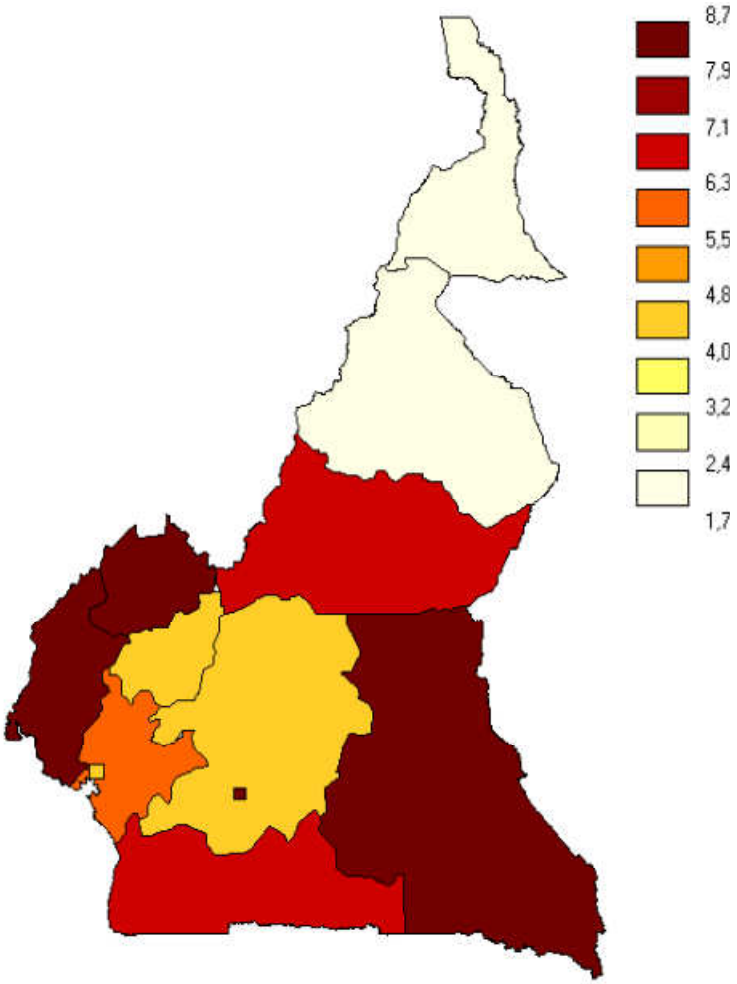


Figure 4.41
Prévalence du VIH selon les régions EDS au Cameroun en 2004



Source : EDS 2004 du Cameroun.

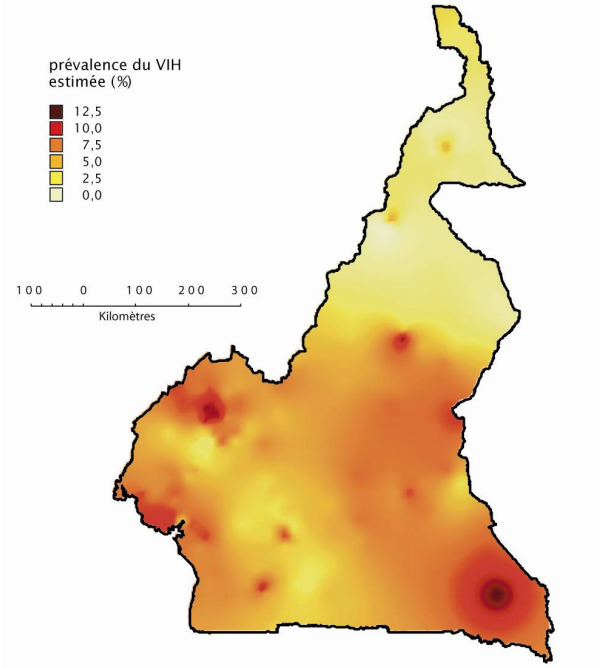
Au Cameroun (Figure 4.42), l'épidémie se concentre dans le sud-ouest du pays, particulièrement dense et urbanisé, Douala, Yaoundé et Bamenda/Bafoussam constituant un triangle historique d'importants mouvements migratoires internes (GUBRY 1983) comme le montre la Figure 4.44 et la Figure 4.45 ci-après. Dans le cadre d'une thèse sur mobilité et infection à VIH au Cameroun, Nathalie LYDIÉ a montré l'importance de celui-ci dans la dynamique épidémique du pays.

> *« Les trois têtes de ponts de ce triangle presque isocèle constituent, semble-t-il de bons relais spatiaux pour la propagation de l'épidémie. Ce réseau principal fonctionne d'autant mieux qu'il est épaulé d'un réseau secondaire bien structuré. » (LYDIÉ 2001, p. 196)*

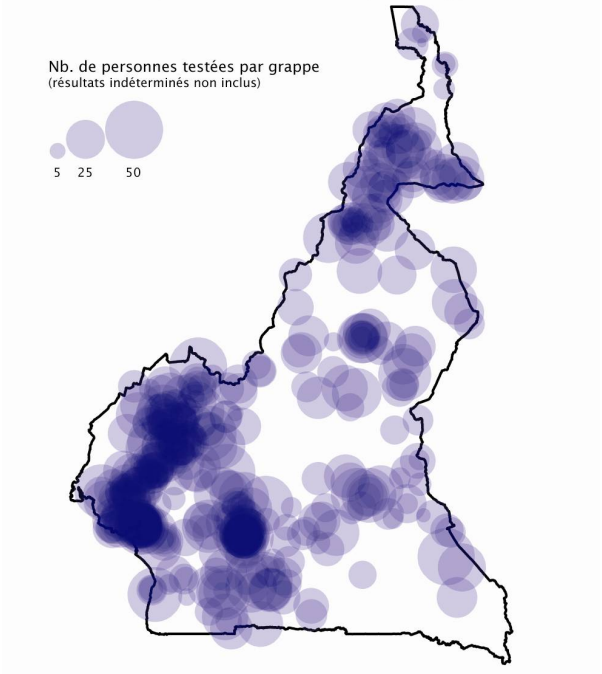
Figure 4.42

Prévalence du VIH estimée au Cameroun et cartes complémentaires

a. Prévalence estimée du VIH (15-49 ans)

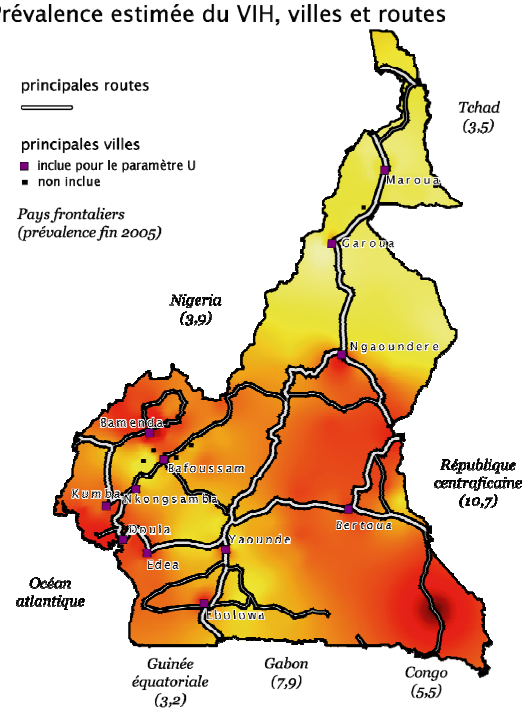


c. Nombre de personnes testées par grappe

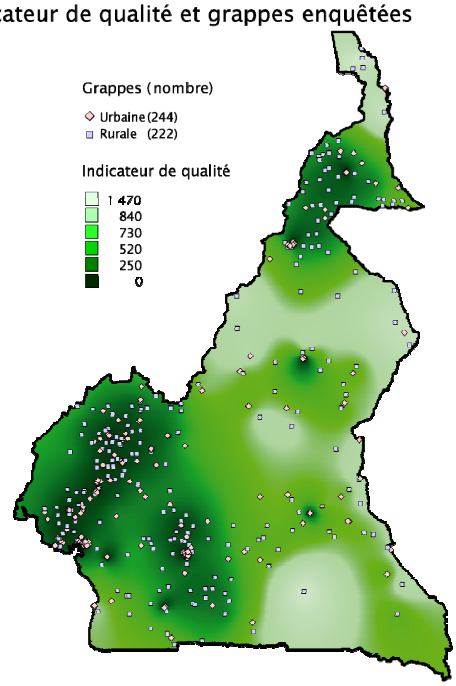


Note 1 : La prévalence nationale du VIH (15-49 ans) est de 5,5% dans l'EDS 2004 du Cameroun. 9.900 personnes ont été testées (résultats indéterminés exclus), réparties en 466 grappes.
Note 2 : Les paramètres utilisés pour l'estimation de la prévalence du VIH sont : N=350, R=117 km et U=12 (voir carte b pour les villes retenues).

b. Prévalence estimée du VIH, villes et routes

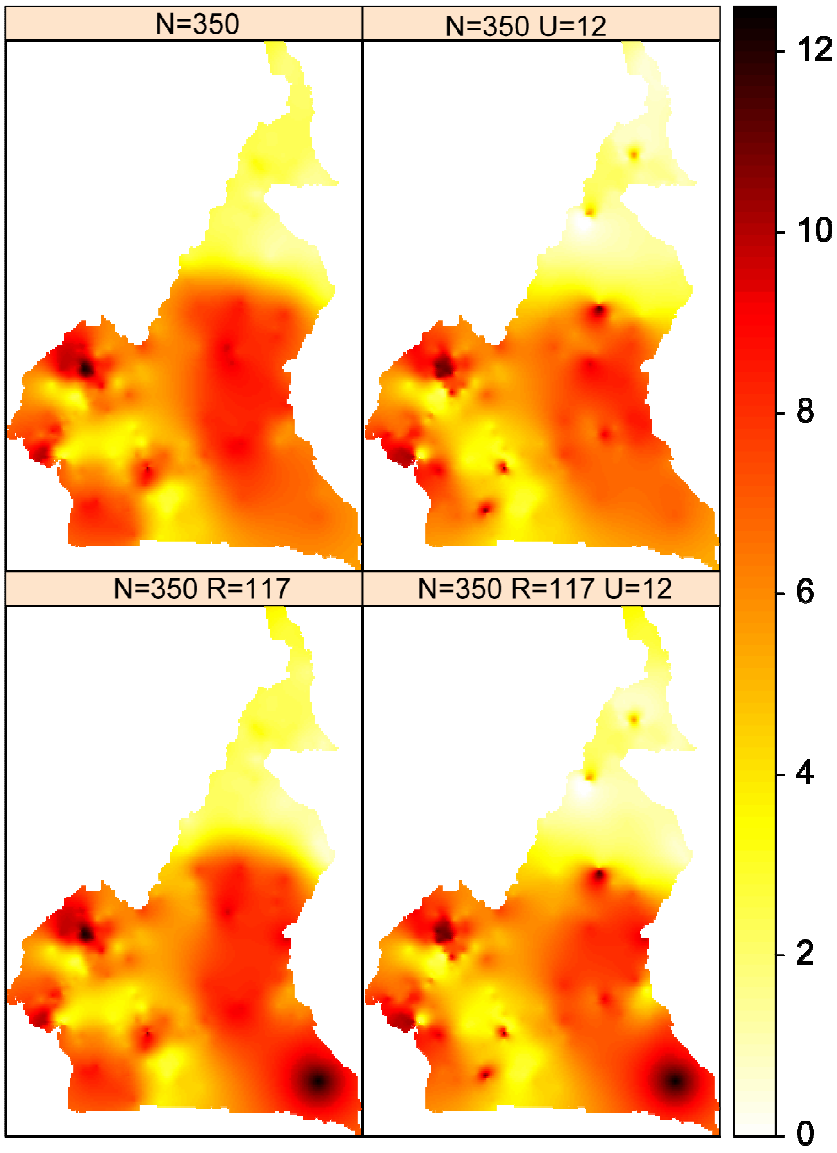


d. Indicateur de qualité et grappes enquêtées



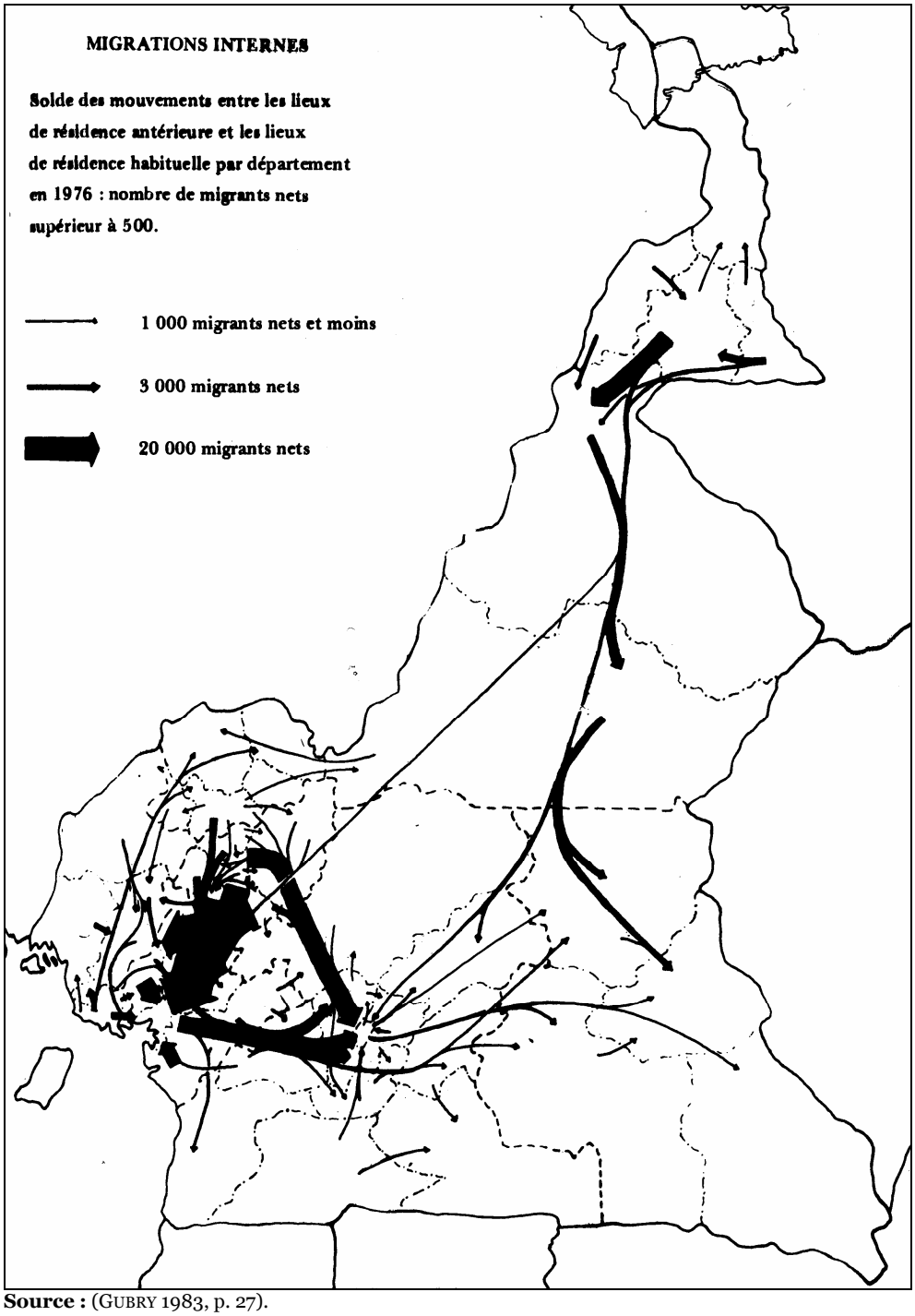
Note 3 : L'indicateur de qualité est calculé pour chaque grappe selon l'expression $r^2/\sqrt{n}$ où r est le rayon du cercle de lissage et n le nombre de personnes testées dans ce cercle.
Sources : EDS 2004 du Cameroun pour les données VIH, DCW pour les frontières nationales, GRUMP pour les principales villes, ArcAtlas (ESRI) pour les principales routes, rapport ONUSIDA 2006 pour les prévalences fin 2005 des pays frontaliers.

Figure 4.43
Comparaison des approches N, NU, NR et NRU pour le Cameroun en 2004



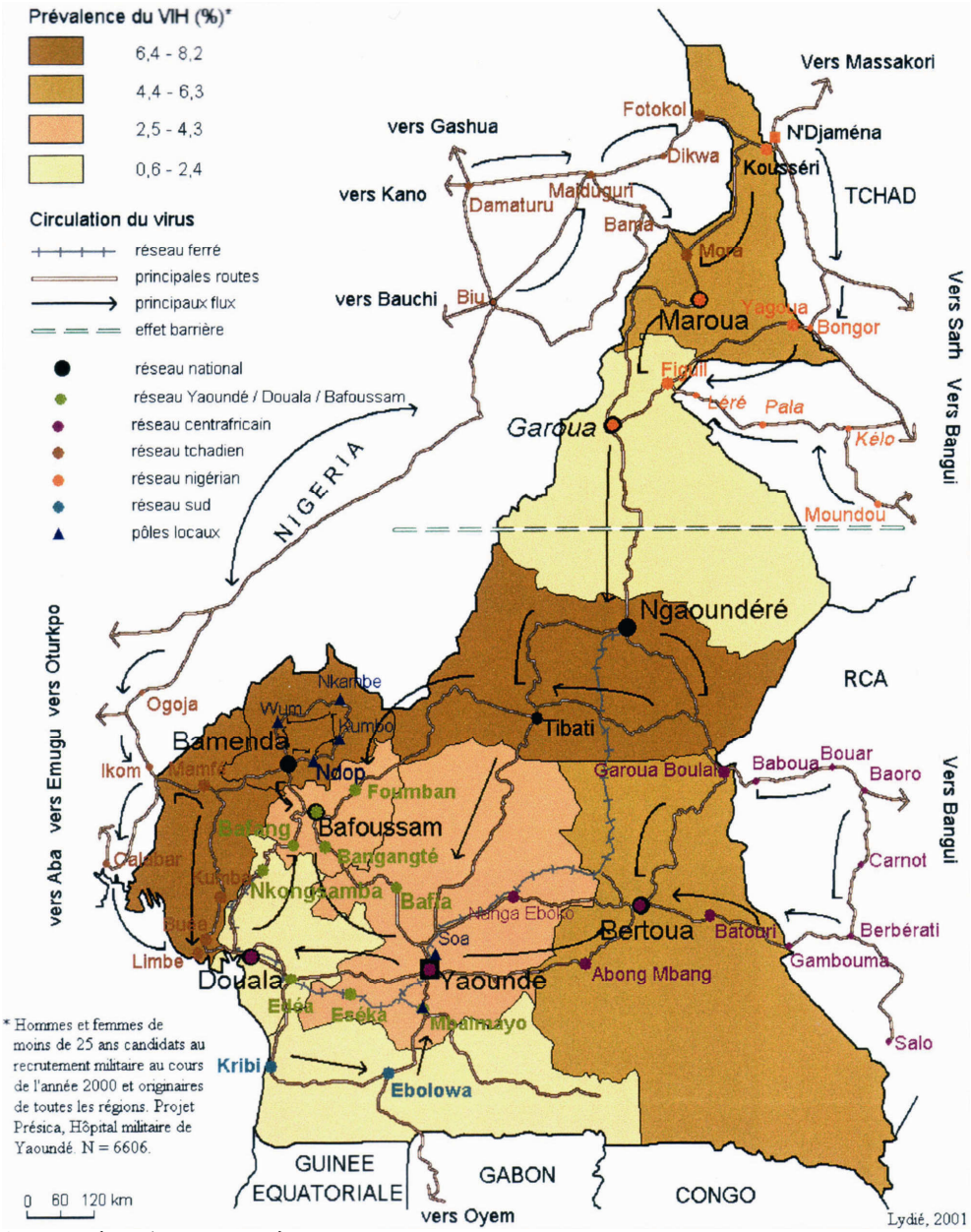
En dehors de cette zone, on retrouve des foyers épidémiques dans les villes situées le long des axes routiers : Maroua, Garoua, Bertoua, Ebolawa et surtout N'Gaoundéré qui présente les niveaux de prévalence les plus élevés du pays. Des liens entre mobilité et infection à VIH ont pu être établis au Cameroun (LYDIÉ 2001, LYDIE 2004). Ce n'est donc pas un hasard si les nœuds routiers transparaissent sur notre cartographie. Néanmoins, nous ne devons pas oublier que l'EDS enquête les individus sur le lieu de vie. Ces résultats traduisent donc la situation des personnes qui vivent dans ces nœuds routiers et non celle des personnes qui y transitent. S'ils sont un indicateur des liens entre réseaux migratoires et infection à VIH, ils sont insuffisants pour pouvoir mener une analyse fine sur la complexité de la dynamique épidémique.

Figure 4.44
Migrations internes au Cameroun en 1976



Source : (GUBRY 1983, p. 27).

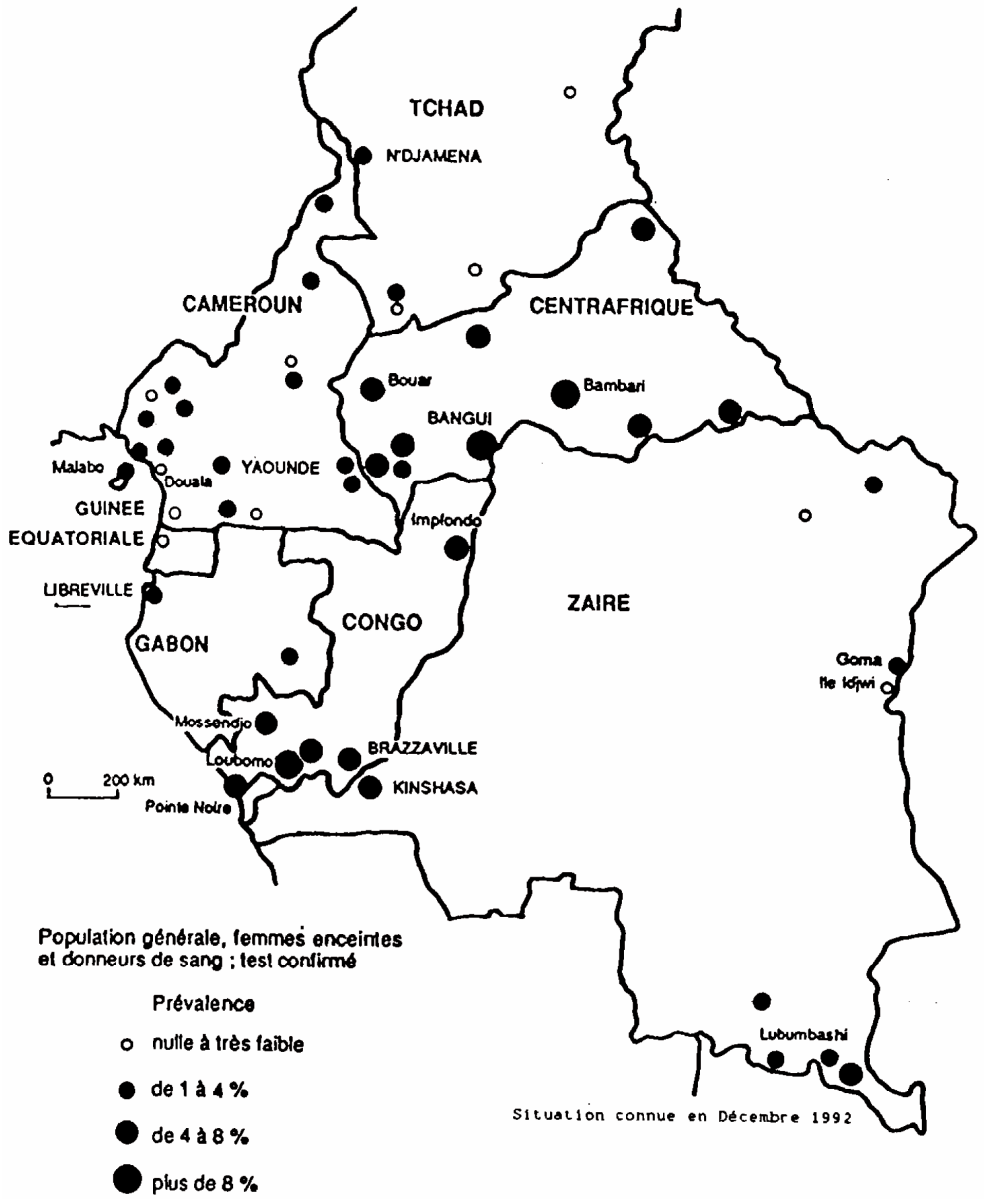
Figure 4.45
Mobilité et infections à VIH au Cameroun



Source : (LYDIÉ 2001, p. 202).

Une rupture nette est observée entre le nord et le sud de N'Gaoundéré, le nord sahélien présentant une épidémie plus faible. Cependant, la ligne de démarcation ne peut être dessinée précisément du fait du faible nombre de grappes dans cette zone et de la valeur élevée de l'indicateur de qualité (Figure 4.42.d).

Figure 4.46
Infection à VIH-1 en Afrique Centrale vers 1991-1992



Source : (RÉMY 1993).

Le Nigéria et le Tchad qui bordent le nord du pays présentent une prévalence du VIH relativement faible (3,9 % et 3,5 % fin 2005 selon Onusida (UNAIDS 2006)) par rapport aux pays limitrophes du sud-est (10,7 % en Afrique Centrale, 5,5 % au Congo et 7,9 % au Gabon). Cette zone frontalière semble présenter des prévalences relativement élevées. Cependant, bien que ce résultat soit plausible, il doit être interprété avec prudence vu le faible nombre de personnes testées et des valeurs élevées de l'indicateur de qualité dans cette zone (Figure 4.42 c et d). Le pic épidémique qui apparaît lorsque le paramètre R est appliqué (Figure 4.43) est dû à deux grappes proches ayant enregistré respectivement cinq et sept cas positifs (voir

Figure 4.4 page 206). La troisième grappe de la zone a enregistré, pour sa part, un cas positif. Les données à notre disposition ne nous permettent pas de conclure quant à la situation réelle de l'épidémie dans cette zone. Cependant, ce résultat constitue une piste de recherche qui mériterait d'être explorée afin de clarifier la situation de cette région.

Les résultats que nous obtenons pour le Cameroun avec la méthode des cercles de même effectif sont particulièrement proches de ceux présentés en 1993 par Georges RÉMY concernant l'Afrique Centrale (RÉMY 1993) et que nous avons reproduits sur la Figure 4.46. Par ailleurs, le foyer épidémique que nous observons dans le Sud-Est du pays y transparaissait déjà, les données de la Figure 4.46 suggérant une diffusion de part et d'autres de la frontière avec la République Centrafricaine. La dernière EDS réalisée dans ce pays date malheureusement de 1993/1994. Nous n'avons donc pas les données adéquates pour procéder à une cartographie comparable de l'épidémie de ce pays.

*
**

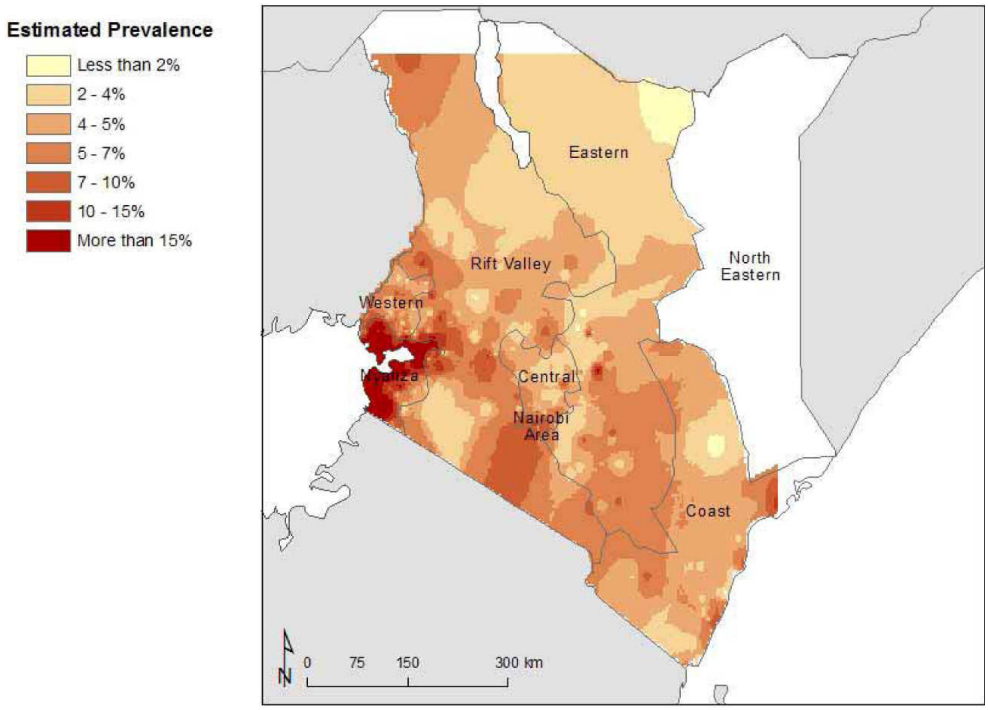
Les résultats obtenus à partir des EDS du Burkina Faso et du Cameroun s'avèrent être cohérents avec les autres informations à notre disposition concernant l'épidémie de ces deux pays. Le gain d'information par rapport à une cartographie simple selon les régions EDS (Figure 4.37 et Figure 4.41) est évident. Notre approche est avant tout descriptive, les données n'étant pas suffisamment fines pour permettre une analyse démonstrative sur les déterminants de la diffusion du VIH. Cependant, elle suggère une diffusion de l'épidémie liée aux axes routiers et migratoires, mais également, pour le milieu rural, à des zones présentant des activités économiques particulières (par exemple la région de Diébougou). Ce résultat trouve écho dans les travaux de Georges RÉMY :

> « *Les villes, notamment les plus grandes, sont spécialement exposées à l'infection à toutes les étapes de sa dynamique… Mais des sites ruraux sont également vulnérables. Ils se distinguent par leur participation à des activités économiques variées, à caractère monétaire : centres miniers, étapes routières, périmètres agro-industriels, marchés. Par contre, les sites les moins atteints, dans leur contexte régional, présentent tous un caractère "villageois" ; ils sont éloignés des pôles économiques, des axes de communication.*
>
> *Saisis à l'échelle nationale ou régionale, les déterminants de la dynamique de l'infection renvoient à une économie active, à la mobilité humaine qui l'accompagne, aux spécificités ou aux changements que l'une et l'autre introduisent dans l'organisation et le fonctionnement des collectivités. » (REMY 1999)*

4.16 Application au Kenya : comparaison avec les travaux de MONTANA *et al.*

À notre connaissance, seuls Livia MONTANA, Melissa NEUMAN et Vinod MISHRA ont également procédé à une cartographie infrarégionales de la prévalence du VIH à partir de données EDS. Ils ont publié en janvier 2007 un DHS Working Paper consacré à une modélisation spatiale de la prévalence du VIH au Kenya[38].

Figure 4.47
Prévalence du VIH estimée au Kenya en 2003 par MONTANA et ses collaborateurs



Source : (MONTANA 2007, Map 3 p. 21).

Les auteurs ont estimé la probabilité de chaque individu d'être séropositif à l'aide d'une régression logistique incluant des caractéristiques sociodémographiques (âge, milieu de résidence, région, instruction, emploi, ethnie, quintile de bien-être, religion, statut matrimonial, naissance dans les cinq dernières années pour les femmes, migration saisonnière pour les hommes), des variables de comportement sexuel (âge au premier rapport, rapport à risque au cours de l'année, multi-partenariat, rapports monnayés, utilisation du préservatif), d'autres variables comportementales (symptômes d'IST, consommation d'alcool et de tabac, perception des risques d'infection au VIH, désir de se protéger, participation aux

[38] MONTANA L., NEUMAN M. et MISHRA V., *Spatial Modeling of HIV Prevalence in kenya*, coll. *DHS Worling Papers*, 2007.

décisions du ménage pour les femmes, accès aux média, circoncision pour les hommes) et des variables spatiales (distance à une route importante, distance au lac Victoria, densité de population).

La prévalence de chaque grappe a ensuite été calculée en faisant la moyenne des probabilités prédites pour chaque individu et la cartographie finale a été générée par interpolation selon une pondération inverse de la distance (Figure 4.47 ci-dessus).

Afin de comparer leurs résultats, nous avons appliqué notre méthodologie aux données de l'EDS 2003 du Kenya. La situation géographique générale du Kenya est représentée sur la Figure 4.48. La répartition spatiale des grappes (Figure 4.49) met en évidence une concentration dans le sud du pays, du lac Victoria à la côte est. À l'inverse, le nombre de grappes est très faible dans le nord du pays. Les grappes urbaines se concentrent dans trois zones : Kisii-Kisumu-Kakamega-Kitale à l'ouest, Nairobi-Nyeri-Embu-Meru au centre et Mombasa-Kilifi-Malindi sur la côte. Cependant, peu d'agglomérations urbaines se détachent nettement. Seules les villes les plus importantes du pays ont été enquêtées sur au moins 3 grappes ou 40 personnes testées, à savoir la capitale Nairobi avec 265 personnes testées réparties sur 19 grappes et Mombasa avec un effectif de 98 individus répartis sur 7 zones d'enquêtes. Nous n'avons donc retenu que ces deux villes pour les approches NU et RNU.

Pour le paramètre N, nous avons utilisé la valeur de 250 conformément au Tableau 4.2 page 241. Nous avons retenu 136 kilomètres pour le paramètre R, correspondant à la valeur du 9^{e} décile des rayons des cercles de lissage lorsque l'approche N est appliquée.

Dans le cas présent, le paramètre U n'apporte guère un surcroit d'information. Par ailleurs, lorsque nous appliquons le paramètre R, un pic épidémique apparaît dans le nord-ouest du pays (Figure 4.51). Or, ce dernier n'est dû qu'à la présence d'une grappe atypique isolée où trois personnes ont été testées positives tandis qu'aucun autre cas positif n'a été observé dans le nord du pays (Figure 4.50). Il nous semble donc préférable de ne pas appliquer de rayon maximum au cercle de lissage. Les données sont trop parcellaires dans ces régions pour pouvoir poser une hypothèse réaliste quant aux variations locales de l'épidémie. Tout ce que nous pouvons dire, c'est que, globalement, le nord est relativement peu touché, l'EDS réalisée en 2003 au Kenya n'ayant pas suffisamment investigué cette partie du pays pour permettre d'en dégager des estimations plus fines. Cette limite interprétative est visible sur la carte de l'indicateur de qualité qui présente une distinction marquée entre le nord et le sud (Figure 4.52).

Vu la faiblesse du nombre total de personnes testées (à peine plus de 6 000) et la concentration marquée des zones d'enquête dans certaines régions, l'approche N s'avère ici la plus appropriée.

Figure 4.48

Carte de situation du Kenya

Source : (UNITED NATIONS CARTOGRAPHIC SECTION 2004a).

Figure 4.49
Répartition des grappes par milieu de résidence selon l'EDS 2003 du Kenya

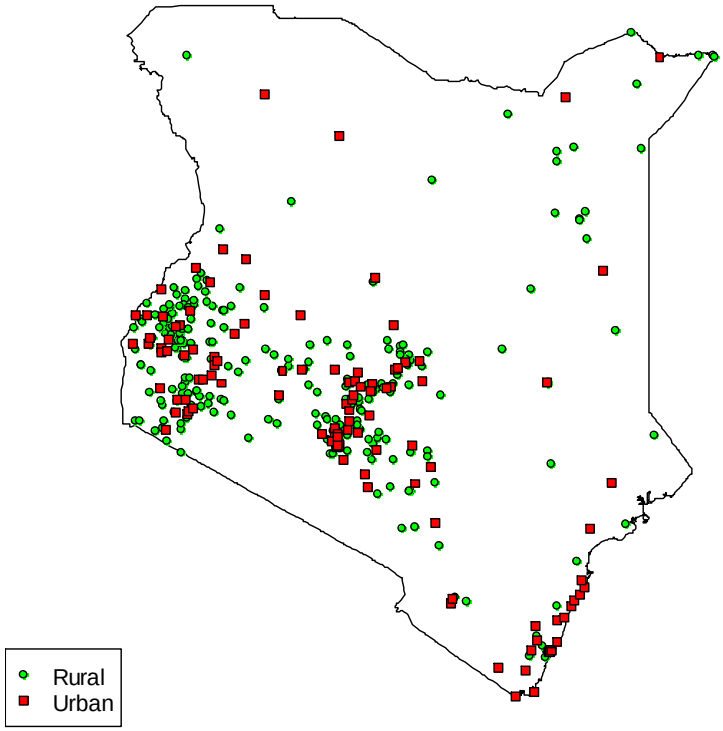


Figure 4.50
Nombre de personnes testées séropositives au VIH selon l'EDS 2003 du Kenya

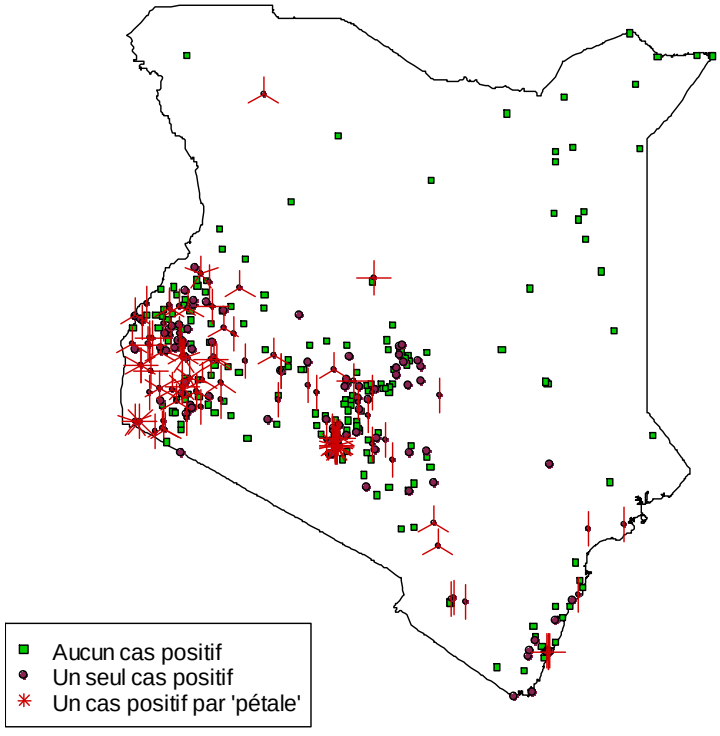


Figure 4.51
Comparaison des approches N, NR, NU et NRU sur les données de l'EDS 2003 du Kenya

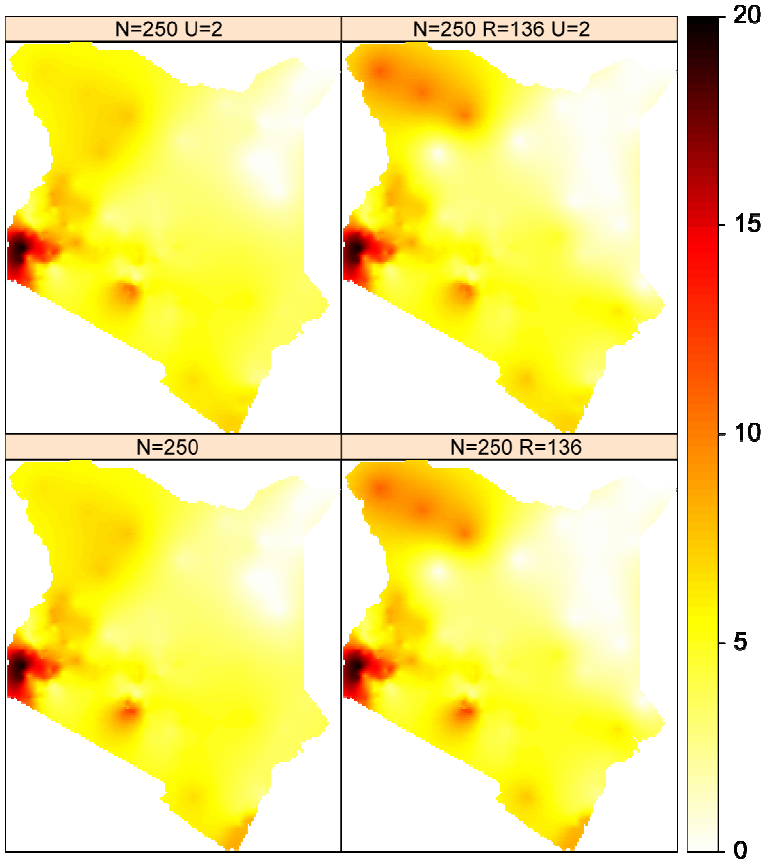


Figure 4.52
Indicateur de qualité pour l'approche N=250 (EDS 2003 du Kenya)

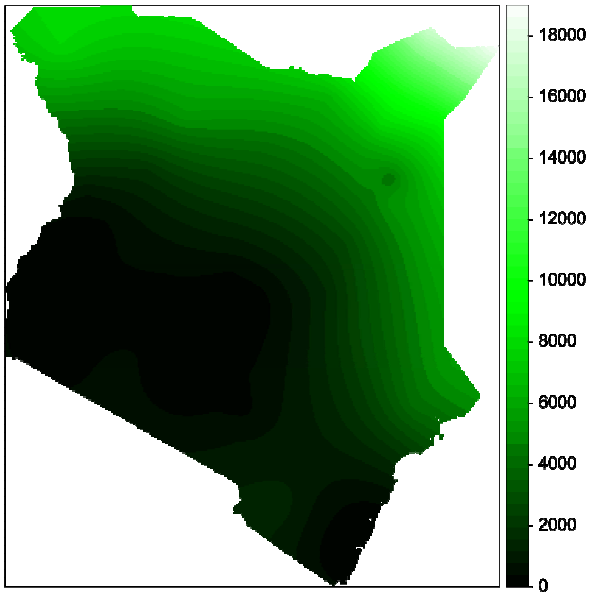
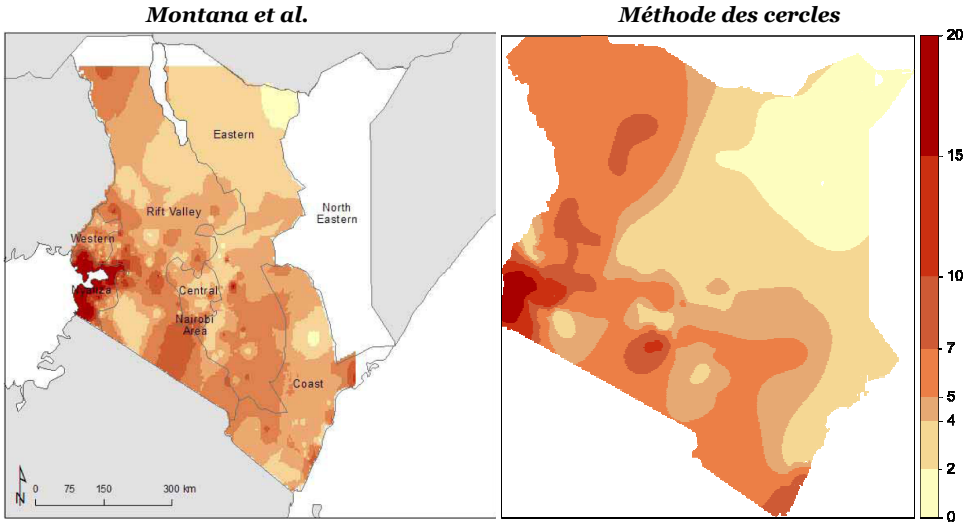


Figure 4.53
Comparaison des résultats de MONTANA et ses collaborateurs et de l'approche par les cercles de mêmes effectifs (EDS 2003 du Kenya)



Nous avons représenté côte à côte avec la même échelle colorimétrique les cartes produites par Montana et ses collaborateurs et la méthode des cercles de même effectifs (Figure 4.53)[39].

Si nous comparons ces deux cartes avec celle du nombre de personnes dépistées positives par grappe (Figure 4.50), celle élaborée par MONTANA et ses collaborateurs est plus proche des observations effectuées. Dans les zones fortement enquêtées, elle s'avère donc plus précise. Elle affiche ainsi une petite poche épidémique au niveau de Lamu sur la côte Est tandis que la méthode des cercles de même effectif y applique une prévalence faible. Cependant, elle est de ce fait plus sensible aux erreurs aléatoires potentielles. La grappe atypique située près de Lodwar dans le Nord-Ouest du pays est traitée différemment par les deux approches. Le modèle par régression logistique induit un pic épidémique décalé vers l'Ouest tandis que la méthode des cercles le décale vers l'Est.

Si les contours de ces deux cartes ne se superposent pas, les principaux patterns épidémiques restent néanmoins les mêmes. L'approche par les cercles de même effectif, plus aisément applicable et nécessitant considérablement moins de données, permet donc de dégager des variations spatiales principales des épidémies de VIH cohérentes avec celles produites via le recours à des méthodes statistiques plus complexes.

[39] Les deux cartes ne sont pas représentées avec la même projection. Par ailleurs, le lac Victoria et le lac Turkana n'ont pas été détourés sur la carte produite avec la méthode des cercles de même effectif.

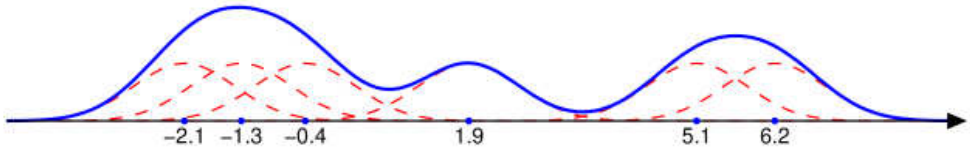
4.17 Perspectives de développements futurs

Peu de travaux cartographiques portant sur les prévalences du VIH ou des données issues d'enquêtes démographiques et de santé ont été publiés. Dans une étude sur les variations spatiales de la mortalité maternelle en Guinée, KEITA et TOURE ont eu recours à la méthode connue en anglais sous le nom de *Kernel density estimation* (estimation de la densité par la méthode du noyau) pour déterminer les variations spatiales de plusieurs indicateurs mesurés par les EDS[40] (KEITA 2006).

Cette méthode consiste à répartir sur une grille chaque observation (courbes pointillés en rouge sur la Figure 4.54) selon une courbe ayant la forme d'un cône (par exemple une loi normale) puis à sommer pour chaque point de la grille la part de chaque observation (courbe bleue). Il est possible de modifier le degré de lissage en variant le rayon du cône[41].

Avec cette méthode, nous pourrions alors calculer une surface de densité des cas positifs ainsi qu'une surface de densité du nombre de personnes testées. Le rapport de ces deux surfaces fournirait alors les variations spatiales de la prévalence.

Figure 4.54
Exemple de lissage à une dimension selon la méthode du noyau



Source : http://en.wikipedia.org/wiki/Image:Parzen_window_illustration.png.

Si cette méthode s'avère efficace lorsque la position de chaque observation est connue avec précision et que l'ensemble des cas a été recensé de manière exhaustive, ses résultats sont plus aléatoires dès lors qu'elle est appliquée à des données issues d'EDS. D'une part, du fait de la répartition inégale des grappes, il est nécessaire d'avoir recours à des cercles de grande taille, d'où une perte de précision dans les zones correctement enquêtées. Par ailleurs, les incertitudes sont amplifiées par le rapport des deux densités estimées, d'où l'apparition possible de valeurs aberrantes.

[40] Plus précisément, ces auteurs ont procédé au rapport de deux surfaces de densité obtenues par cette méthode (voir les paragraphes suivants).

[41] Pour plus de détails sur cette méthode, voir DI SALVO M., GADAIS M. et ROCHE-WOILLEZ M., *L'estimation de la densité par la méthode du noyau : méthodes et outils*, Lyon (FR), CERTU, coll. *Rapports d'études du CERTU*, 2005.

Si peu de modèles spatiaux ont été développés concernant le VIH dans les pays en développement, des recherches ont été menées dans des pays du Nord ou sur d'autres pathologies. La technique d'interpolation spatiale par krigeage a d'ailleurs été dans plusieurs études épidémiologiques (CARRAT 1992, OLIVER 1992, OLIVER 1998, KLEINSCHMIDT 2001). Plusieurs études ont été menées sur les infections à chlamydia ou à gonorrhée (BECKER 1998, SHAHMANESH 2000, ELLIOTT 2002, ZENILMAN 2002). En Caroline du Nord (USA), les prévalences des infections à chlamydia, des gonorrhées, de la syphilis et du VIH ont pu être estimées et cartographiées (LAW 2004). Les auteurs disposaient de densités de population très précises et ont pu positionner chaque cas observés à partir des adresses de résidence des individus. En Afrique, plusieurs travaux ont porté sur l'analyse spatiale du paludisme (KLEINSCHMIDT 2000, KLEINSCHMIDT 2001, MAUNY 2004, GEMPERLI 2006, CRAIG 2007). Outre le fait que ces différentes études portaient sur des effectifs relativement importants, elles ont toutes eu recours à des modèles prédictifs complexes incluant, outre des données de surveillance, des facteurs environnementaux et climatiques, bien documentés concernant cette pathologie.

Les approches statistiques développées sont nombreuses et variées. Cependant, elles requièrent toutes une information abondante et précise. Elles nécessitent par ailleurs une bonne connaissance préalable du phénomène étudié afin de poser les hypothèses adéquates dans leurs modèles.

L'approche que nous avons développée ici par le recours à des cercles de même effectif se situe, pour sa part, dans une démarche de statistiques imparfaites. Il s'agit d'extraire un maximum d'information, à partir de données parcellaires, en essayant de réduire nos hypothèses à leur minimum. La méthode des cercles est aisée à mettre en œuvre, notamment avec prevR, et ne nécessite que peu de données. Outre les résultats d'enquêtes, seuls les frontières du pays et la localisation des principales villes sont nécessaires, données qui sont disponibles de manière standardisée et gratuite sur internet. Elle permet de reconstruire les principales variations des épidémies et s'avère ainsi un outil descriptif pertinent. L'un de nos objectifs était d'ailleurs d'élaborer un outil simple de suivi épidémiologique. Par ailleurs, cette méthodologie n'est pas spécifique au VIH et peut s'appliquer à toute proportion mesurée dans une enquête de type EDS.

Cependant, les résultats sont fortement limités par la qualité et la quantité des données initiales présentes dans les EDS. Celles-ci sont insuffisantes pour réaliser une analyse précise des facteurs influençant la structure spatiale des épidémies. Il est nécessaire pour cela d'élaborer des enquêtes de plus grande ampleur.

Nous avons fait l'hypothèse d'un raisonnement par analogie en extrapolant les résultats de notre pays modèle concernant la valeur optimale de N. Si cette modélisation permet de vérifier la capacité de l'approche par des cercles de mêmes effectifs à traduire correctement les variations régionales de la prévalence, elle ne peut servir de test de vérification concernant le choix de la meilleure valeur de N,

puisque ces valeurs optimales de N ont été calculées à partir d'elle. Il nous faudrait, pour vérifier cette hypothèse, envisager un test ne faisant pas intervenir Alicante.

Une possibilité consisterait à simuler des enquêtes démographiques et de santé à partir de données réelles de recensement. Du fait de leur exhaustivité, nous pourrions alors comparer les variations réelles d'un phénomène à un niveau local et les variations reconstruites après simulation d'une EDS, estimation et interpolation spatiale. Cependant, nous n'avons pu procéder à une telle vérification sur des données africaines en raison des difficultés d'accès à des données de recensement géoréférencées[42].

Le développement et la multiplication des EDS (et enquêtes apparentées) incluant le dépistage du VIH permettent d'envisager plusieurs pistes de recherche à approfondir.

D'une part, il apparaît que les estimations dans les zones frontalières sont plus fréquemment d'une qualité moindre. En effet, nous ne disposons d'informations que d'un seul côté de la frontière et, de ce fait, les variations de l'épidémie peuvent être mal rendues par la méthode des cercles. Nous pourrions envisager de procéder à l'estimation des prévalences et à une interpolation spatiale en fusionnant les données provenant de plusieurs enquêtes. Si elles sont relativement proches dans le temps, du fait de l'inertie relative de la prévalence[43], la fusion de données diachroniques sera une hypothèse acceptable. Cependant, il restera à régler la question de la pondération à appliquer aux grappes de chaque pays, chaque enquête étant échantillonnée séparément. Ainsi, la réalisation d'une AIDS Impact Survey en 2005 en Côte d'Ivoire laissait supposer la possibilité d'une telle cartographie en fusionnant ces données avec celles de l'EDS 2003 du Burkia Faso. Mais, les coordonnées géographiques des zones d'enquêtes n'ont pas été collectées pour cette enquête. Alors que la collecte par GPS des longitudes et latitudes de chaque grappe étaient devenues un élément presque standard des EDS à la fin des années 1990 et au début des années 2000, plusieurs EDS récentes ne l'ont pas inclue bien que ce fut le cas pour les enquêtes précédentes dans le pays. C'est le cas

[42] Nous disposions, des suites d'une autre recherche, des données du recensement français de la population de 1999, agrégées par zone IRIS (découpage fin du territoire français par l'INSEE), pour l'Île-de-France. Nous y avons simulé une enquête avec 8 000 personnes enquêtées réparties en 400 grappes et comparé la carte du taux d'actifs allant travailler en transport en commun observé dans le recensement et celle obtenue par l'approche NR (cartes non reproduites) en utilisant la valeur optimale de N déduite du modèle. Il s'est avéré que les tendances spatiales du phénomène étaient correctement reproduites à l'exception des variations linéaires (le long des lignes de RER) dans la partie extérieure de la grande couronne.

[43] En tant que variable de type *stock*, la prévalence du VIH évolue plus lentement que l'incidence ou la mortalité qui représentent les *flux* des épidémies. Une certaine durée est ainsi nécessaire pour que des changements d'incidence ou de mortalité modifient amplement les niveaux de prévalence.

notamment de l'EDS 2005 du Sénégal ou de l'EDS 2006 du Mali selon le site de MEASURE DHS.

D'ici quelques années, certains pays disposeront de deux EDS avec dépistage du VIH. Outre la possibilité de mesurer si l'épidémie a augmenté ou diminué durant la période inter-enquête, une comparaison des cartographies obtenues aux deux dates permettra d'appréhender si la géographie globale de l'épidémie s'est modifié pendant cette période. Par ailleurs, diverses techniques d'interpolation spatio-temporelle ont été développées (LI 2002) dont le krigeage spatio-temporelle (LU 2006). Bien que ces techniques soient relativement complexes à mettre en œuvre, elles constituent une piste à explorer pour mieux comprendre les dynamiques des épidémies de VIH.

*
**

Les EDS sont conçues pour être représentatives de la population nationale et non des territoires. De ce fait, la répartition spatiale des personnes enquêtées n'est pas aléatoire mais reflète les densités de population. L'approche développée ici repose sur deux éléments. D'une part une estimation de la prévalence de chaque grappe à partir de cercles de même effectif, afin de compenser les aléas du tirage au second degré (sélection des ménages éligibles de chaque grappe). D'autre part, une interpolation spatiale pour estimer la surface de variation spatiale de la prévalence et compenser ainsi le tirage au premier degré (sélection des grappes).

Nous avons fait le choix d'une méthodologie reposant sur des cercles de même effectif plutôt que des cercles de même rayon. Cette seconde alternative aurait été appropriée dans le cas d'enquêtes échantillonnées sur les territoires (sélection des zones d'enquêtes avec une probabilité égale à leur superficie). Cependant, pour les EDS qui sont échantillonnées avec une probabilité de sondage des zones d'enquête égale à leur population[44], nous aurions alors estimé la prévalence sur un nombre de personnes très élevé dans les zones fortement peuplée et procédé ainsi à une perte d'informations. Dans les zones peu denses, les effectifs auraient été trop faibles pour compenser les erreurs aléatoires de l'échantillonnage.

Le recours à une fenêtre mobile de même effectif, inspirée de techniques d'analyse en composantes d'échelle, permet de mettre à jour les tendances régionales tout en minimisant l'effet des erreurs aléatoires, bien que ce soit au prix de la perte d'une information sur les résidus locaux. L'interpolation spatiale devient alors possible à partir des données parcellaires des EDS. Nous nous heurtons néanmoins aux limites inhérentes à ce type d'enquête et nos résultats sont fortement dépendants de la qualité et de la quantité des données initiales. Quelque soit l'approche

[44] Plus précisément le nombre de ménages au dernier recensement de la population.

employée, il reste impossible de descendre à des niveaux locaux inférieurs aux « mailles » de la répartition des grappes.

Les cartes obtenues sur les données de prévalence du VIH au Burkina Faso et au Cameroun sont plausibles et cohérentes au regard des informations que nous avons sur ces épidémies. Par ailleurs, au Kenya, les tendances mises à jour par notre approche correspondent à celles générées par une modélisation à l'aide de régressions logistiques.

La technique des cercles de même effectif reste simple à mettre en œuvre et ne requiert que peu d'hypothèses. L'interprétation est facilitée par la génération de cartes complémentaires et d'un indicateur de qualité. Néanmoins, la cartographie produite ne doit pas être interprétée « au pixel prêt ». Il importe de garder en mémoire les différentes hypothèses posées et discutées au cours de l'analyse (notamment concernant le choix des agglomérations urbaines retenues et l'utilisation ou non du paramètre R).

Malgré ses limites, cette approche permet d'obtenir des cartes plus précises que celles obtenues à partir des prévalences calculées par région (Figure 4.37 page 260 et Figure 4.41 page 263). Elles permettent, de fait, une connaissance plus fine de la répartition spatiale des épidémies. Bien que les données soit insuffisantes pour permettre une analyse fine des déterminants spatiaux, les résultats obtenus fournissent des pistes de recherche à explorer (par exemple, la situation de la région de Diébougou au Burkina Faso ou celle dans la pointe Sud-Est du Cameroun).

Chapitre 5
Échelles, niveaux et tendances

Au cours des deux précédents chapitres, nous avons abordé en filigrane les notions d'échelles d'analyse, de niveaux ou bien encore de tendances. Il s'agit des différentes dimensions d'une mesure. Elles traduisent divers aspects de nos énoncés et véhiculent des significations distinctes.

La notion d'échelle renvoie à celle d'ensemble population-espace-temps associé à un énoncé. Ainsi, comme nous l'avons montré dans le Chapitre 2, la validité d'un indicateur est limitée à l'échelle, qu'elle soit géographique, temporelle ou populationnelle, à laquelle il a été calculé. Nous ne pouvons directement extrapoler nos observations à une autre échelle sans poser des hypothèses anticipatrices adéquates.

Les notions de niveaux et de tendances renvoient, quant à elles, aux diverses manières de poser une conditionnalité sur des concepts opératoires. Elles permettent de préciser ce qui est porteur de sens. Une mesure peut ainsi être valide en ce qui concerne les tendances d'un indicateur, bien qu'elle ne le soit pas concernant son niveau.

Une mesure de niveau se traduira ainsi par une égalité entre une valeur chiffrée et un concept opératoire. Du fait que nous devons prendre en compte la précision de nos mesures ou de nos estimations, un indicateur de niveau se traduira plus exactement par un intervalle au sein duquel se situe notre indicateur. Plus cet intervalle sera restreint et plus précise sera notre mesure.

Une mesure de tendances s'exprime, pour sa part, par une série de valeurs chiffrées qui vont traduire l'évolution d'un indicateur dans l'espace et/ou dans le temps. Nous parlerons alors de croissance, de décroissance ou de stagnation. Dans une vision plus mathématique, nous pourrions dire que les niveaux correspondent aux valeurs prises par une fonction tandis que les tendances correspondent à sa dérivée. S'il est possible de déduire les tendances d'un indicateur lorsque nous avons une connaissance précise de ses niveaux, la réciproque n'est pas vraie, la connaissance des tendances ne permettant pas à elle seule d'en déterminer les niveaux.

Enfin, évoquons la notion de seuil bien qu'elle ne nous concerne pas directement ici. Une mesure de ce type nous informera de la position d'un indicateur par rapport à une valeur seuil. Un exemple nous est fourni avec la mesure de la charge virale plasmatique chez une personne infectée par le VIH. Lorsque la charge virale est indétectable, cela ne signifie pas pour autant qu'elle est nulle mais simplement qu'elle se situe en-dessous d'un certain seuil (50 copies d'ANR VIH/mL[1]). Nous n'avons alors aucune information précise sur le niveau de celle-ci ou de ses tendances si ce n'est qu'elle se maintient en dessous de ce seuil.

Nous reviendrons dans un premier temps sur les enquêtes nationales en population générale qui fournissent actuellement la meilleure estimation de la prévalence nationale (section 5.1). Nous verrons également ce qu'il en est pour des prévalences distribuées selon le milieu de résidence, la région, une caractéristique sociodémographique ou portant sur une sous-population spécifique.

Puis, nous synthétiserons nos conclusions précédentes concernant la surveillance sentinelle des femmes enceintes (section 5.2). Que pouvons-nous dire d'une comparaison entre surveillance sentinelle et cartographie à partir des EDS ? Qu'en est-il des tendances observées chez les femmes enceintes ?

Cela nous amènera à discuter du logiciel EPP (Estimation and Projection Package) développé par l'ONUSIDA pour ses estimations biennales par pays (section 5.3). Les évolutions successives d'EPP lui ont permis de contourner certaines de ses limites initiales.

[1] YENI P. et GROUPE DES EXPERTS « PRISE EN CHARGE MÉDICALE DES PERSONNES INFECTÉES PAR LE VIH », *Prise en charge médicale des personnes infectées par le VIH - Recommandations du groupe d'experts*, Paris (FR), Ministère de la Santé et des Solidarités, Médecine-Sciences Flammarion, 2006. Cette valeur seuil peut varier en fonction de la technologie utilisée.

5.1 Les EDS : meilleure estimation du niveau

5.1.1 Prévalences nationales, régionales et par milieu de résidence

Nous avons montré en détail dans le Chapitre 3 (section 3.3 page 142 et suivantes) que les Enquêtes Démographiques et de Santé, malgré les différentes sources de biais, fournissent de bonnes estimations du niveau national de la prévalence du VIH. Les prévalences ajustées que nous avons calculées, sous différentes hypothèses maximisant les biais, se situaient, pour les EDS 2003 du Burkina Faso et du Kenya et l'EDS 2004 du Cameroun, au sein des intervalles de confiance à 95 % des prévalences observées. La prévalence nationale calculée à partir d'une EDS est de fait un bon indicateur du niveau de l'épidémie, à la condition de prendre en compte sa précision, dont une estimation raisonnable nous est fournie par l'intervalle de confiance à 95 %. Une erreur encore courante consiste à interpréter la prévalence obtenue en lui attribuant une précision excessive. Cela n'a pas de sens, en l'occurrence, de préciser les centièmes de décimales[2] et tout résultat se doit d'être accompagné de son intervalle de confiance.

La majorité des EDS est conçue de manière à être représentative non seulement au niveau national mais également selon les régions et le milieu de résidence. La manière dont ces enquêtes sont échantillonnées leur confère effectivement cette représentativité. Notre propos précédent reste donc valable pour les prévalences calculées par région ou par milieu de résidence. Encore une fois, il importe de préciser, pour chaque prévalence, son intervalle de confiance. Celui-ci sera nécessairement d'une plus grande amplitude que celui de la prévalence nationale en raison d'effectifs plus faibles. Les prévalences régionales seront donc d'une précision moindre.

Nous pouvons néanmoins nous demander dans quelle mesure les différents biais évoqués au Chapitre 3 peuvent impacter différemment chaque région. La proportion d'individus infectés non observables du fait de la fenêtre sérologique des procédures de dépistage peut être considérée constante à travers le territoire, le même procédé étant utilisé pour l'ensemble de l'enquête.

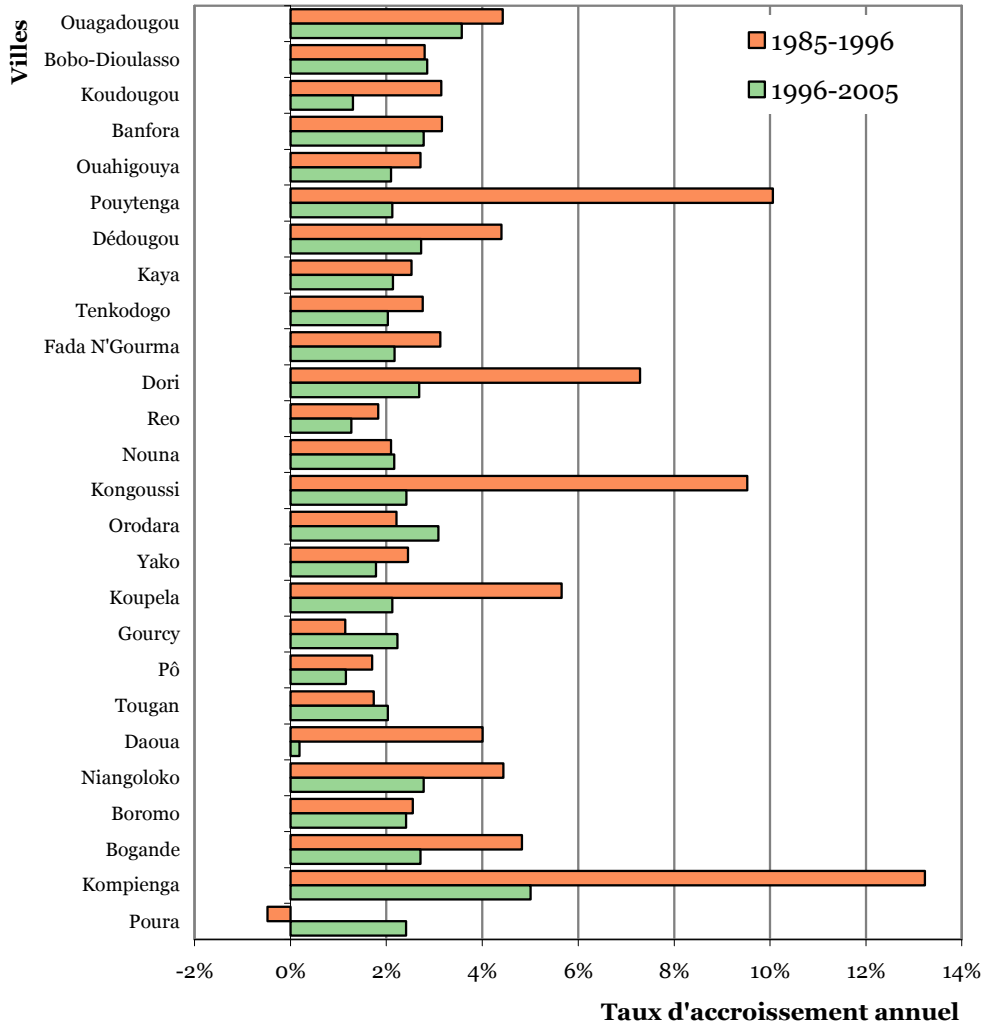
Concernant l'ancienneté de la base de sondage, le principal effet sur les estimations régionales dépendra de l'évolution du taux d'urbanisation de chaque région. Celui-ci n'est généralement pas directement disponible dans les publications des instituts

[2] C'est à dessein que nous avons, dans la section 3.3, transcrit les prévalences nationales avec une précision à deux chiffres. En effet, vu le faible impact de nos ajustements, deux chiffres après la virgule se sont avérés nécessaires pour représenter les écarts entre prévalences observées et prévalences ajustées.

nationaux de la statistique. Cependant, nous pouvons rapprocher les taux d'accroissement annuel des principales villes du Burkina Faso (Figure 5.1) de ceux des régions correspondantes sur la même période (Figure 3.12 page 153). La situation est similaire au Cameroun et au Kenya.

Figure 5.1

Taux d'accroissement annuel des principales villes du Burkina Faso (1985-2005)



Source : (INSTITUT NATIONAL DE LA STATISTIQUE ET DE LA DÉMOGRAPHIE 2006), nos calculs.

Le Tableau 5.1 permet de comparer le taux d'accroissement annuel, sur la période 1996-2005, des quinze principales villes du Burkina Faso avec celui de leurs régions respectives. Un taux d'accroissement de la ville supérieur à celui de sa région traduira une augmentation du taux d'urbanisation de celle-ci[3]. Pour la

[3] Pour être parfaitement rigoureux, il faudrait prendre en compte, pour une région donnée, l'ensemble des villes qui s'y trouvent.

plupart des villes présentées ici, leur taux d'accroissement annuel est sensiblement le même que celui de leur région. Il s'avère inférieur dans certains cas et seule la ville de Dédougou présente un taux sensiblement supérieur. À moins que la base de sondage ne soit particulièrement ancienne et qu'une région ait vu son taux d'urbanisation évoluer de manière très importante, nous pouvons supposer que l'impact de l'ancienneté de la base de sondage sur les estimations des prévalences régionales devrait être du même ordre de grandeur que celui estimé sur la prévalence nationale. Dans la majorité des cas, cela n'affecterait donc pas significativement les estimations par région.

Tableau 5.1

Taux d'accroissement annuel (1996-2005) des quinze principales villes du Burkina Faso et de leurs régions respectives

Ville	Population 1996	Tx d'accr. 1996-2005	Région	Tx d'accr. rég. 1996-2005
Ouagadougou	750 398	3,6%	Centre	3,6%
Bobo-Dioulasso	309 771	2,9%	Hauts-Bassins	3,0%
Koudougou	72 490	1,3%	Centre Ouest	1,7%
Banfora	49 724	2,8%	Cascades	2,6%
Ouahigouya	52 193	2,1%	Nord	2,1%
Pouytenga	35 720	2,1%	Centre Est	2,1%
Dédougou	33 815	2,7%	Boucle du Mouhoun	2,3%
Kaya	33 958	2,1%	Centre Nord	2,2%
Tenkodogo	31 466	2,0%	Centre Est	2,1%
Fada N'Gourma	29 254	2,2%	Est	2,9%
Dori	23 768	2,7%	Sahel	2,8%
Reo	22 534	1,3%	Centre Ouest	1,7%
Nouna	19 105	2,2%	Boucle du Mouhoun	2,3%
Kongoussi	17 893	2,4%	Centre Nord	2,2%
Orodara	16 581	3,1%	Hauts-Bassins	3,0%

Source : (INSTITUT NATIONAL DE LA STATISTIQUE ET DE LA DÉMOGRAPHIE 2006), nos calculs.
Tx d'accr. : taux d'accroissement annuel. Tx d'accr. rég. : taux d'accroissement annuel de la région.

Concernant les individus ne résidant pas au sein d'un ménage ordinaire, le biais sera dépendant de la situation propre à chaque pays. De manière générale, les individus hors ménage se répartissent dans les différentes régions du pays et, du fait de leur faible poids dans la population générale, ils n'influent que peu sur les prévalences régionales. Cependant, concernant les camps de réfugiés ou les camps militaires de très grandes tailles, rassemblant plusieurs milliers d'individus, ces populations hors ménage peuvent alors représenter une proportion non négligeable de la population d'une région, notamment s'il s'agit d'une région faiblement peuplée.

Au Kenya, en 2003, selon l'UNHCR (UNHCR 2004a), 89 378 individus vivaient dans le camp de Kakuma et 134 716 dans ceux de Dadaab (Figure 3.9 page 146), dont la moitié serait âgée de 15 à 49 ans soit respectivement 44 689 et 67 358.

Le camp de Kakuma est situé dans la région Rift Valley dont la population âgée de 15 à 49 ans était, en 2003, de 3 737 367 (Tableau 3.7 page 156). Les adultes vivant en camps de réfugiés ne représentaient donc que 1,20 % de la population de la région. La prévalence du camp de Kakuma était estimée à 5,0 % en 2002 (Figure 3.9 page 146) tandis que celle observée dans l'EDS pour cette région était de 5,29 % (Tableau 3.8 page 157). Si nous prenons en compte le camp de réfugiés de Kakuma, nous calculons alors une prévalence ajustée de 5,2865 %[4] !

La région North Eastern, où se situent les camps de Dadaab, est en revanche peu peuplée : seulement 591 377 adultes en 2003. Les trois camps de Dadaab représentent alors 11,4 % de la population de la région. Une prévalence de 1,4 % a été observée en 2005 dans les camps tandis que l'EDS n'a enregistré aucun cas positif sur les 110 personnes testées dans cette région. Nous calculons alors une prévalence ajustée de 0,16 % à comparer à l'intervalle de confiance à 95 % de la prévalence observée (entre 0 et 4,2 %).

Les camps de réfugiés pourront donc avoir éventuellement un impact sur l'estimation de la prévalence d'une région si leur part dans la population régionale est importante et s'ils présentent une prévalence significativement différente de celle du reste de la région.

Concernant les taux de participation au test de dépistage du VIH, nous pouvons rapprocher et comparer les prévalences observées par région (Tableau 3.8 page 157) et celles prédites par régressions logistiques (Tableau 3.14 page 170). Nous avons calculé pour chaque prévalence observée son intervalle de confiance à 95 % (tableau non reproduit). Pour l'ensemble des régions des EDS du Burkina Faso, du Cameroun et du Kenya, la prévalence prédite se situe au sein de l'intervalle de confiance à 95 %. Cela est toujours vérifié lorsque nous réduisons l'ampleur de ces intervalles en prenant une confiance de 75 %. Nos conclusions concernant l'impact des taux de non testés sur les estimations des prévalences nationales (section 3.3.4) restent donc valables concernant les prévalences régionales.

Il importe cependant de vérifier au préalable les taux de couverture du test de dépistage par région. Il se peut que, pour diverses raisons, ce dernier soit particulièrement bas pour une région donnée. Ainsi, dans le cadre de l'EDS 2004 du Malawi, la région de Lilongwe présente des taux de participation particulièrement bas : 39 % pour les femmes et 38 % pour les hommes (NATIONAL

[4] Nous avons affiché quatre décimales afin de montrer à quel point l'impact du camp de réfugiés de Kakuma est négligeable.

STATISTICAL OFFICE 2005). Bien que le rapport final de l'enquête ne précise pas les raisons de ces faibles taux, les auteurs ont procédé à un ajustement, selon une approche par régressions logistiques déjà utilisée par ORC Macro (MISHRA 2006) et comparable à celle que nous avons utilisée à la section 3.3.4. Si l'impact est limité concernant les autres régions (prévalence ajustée de 13,1 % pour l'ensemble du Malawi, excepté Lilongwe, contre 13,2 % pour la prévalence observée), la prévalence prédite par le modèle pour Lilongwe s'avère de 10,3 % tandis que celle observée est de 3,7 % !

*
**

Nous n'avons pas détaillé ici la question des prévalences par milieu de résidence. Cependant, en procédant selon un raisonnement similaire, nous arriverions aux mêmes conclusions. Les EDS, par leur mode d'échantillonnage, fournissent de bons indicateurs du niveau des prévalences nationales, régionales et par milieu de résidence, à la condition de tenir compte systématiquement de la précision des estimations (traduite par les intervalles de confiance à 95 %) et sous réserve qu'une région ne présente pas un taux anormalement bas de participation à l'enquête ou bien que les populations hors ménage (en particulier celle des grands camps de réfugiés) y soient particulièrement importantes et que les prévalences de ces dernières soient d'un autre ordre de grandeur par rapport à celle de la population des ménages ordinaires.

5.1.2 Distributions sociodémographiques des prévalences

Différents modèles épidémiologiques, démographiques ou économiques nécessitent, parmi leurs paramètres d'entrée, des estimations des distributions des prévalences du VIH par âge, sexe ou d'autres variables sociodémographiques. C'est le cas, par exemple, du logiciel Spectrum utilisé par l'ONUSIDA pour ses projections sur le VIH/SIDA et ses estimations de l'impact socio-économiques des épidémies (STOVER 2007). Les modèles économiques développés pour analyser les coûts liés à la généralisation des traitements antirétroviraux nécessitent d'estimer le nombre d'individus susceptibles d'être mis sous traitement (FREEDBERG 2003), estimation elle-même réalisée à partir, entre autres, de données de prévalences.

Nous avons montré précédemment que la surveillance sentinelle des femmes enceintes était inappropriée (section 3.4.1), certains patterns pouvant être inversés chez les femmes enceintes par rapport à ceux observés parmi l'ensemble des femmes. C'est le cas par exemple du niveau d'instruction. Dans certaines cliniques prénatales, les femmes instruites présentaient une prévalence du VIH plus élevée tandis que des enquêtes locales en population générale montraient l'inverse (FYLKESNES 2001, GREGSON 2001, GREGSON 2002b). Par ailleurs, la prévalence des

hommes ne peut être déduite à partir des données de surveillance sentinelle (section 3.4.2).

Les enquêtes en population générale sont, pour leur part, tout à fait indiquées pour calculer la distribution des prévalences du VIH selon certaines variables démographiques. Plusieurs études ont d'ailleurs analysé les déterminants socioéconomiques de l'infection à VIH à partir d'échantillons représentatifs nationaux (AKWARA 2005, DE WALQUE 2006, LACHAUD 2007). Cependant, comme à chaque fois, les estimations seront dépendantes des effectifs, de la qualité des données et des taux de non réponse.

Cela pourra paraître trivial mais, avant toute estimation, il est nécessaire de vérifier que nous disposons, pour chaque catégorie de la variable analysée, d'effectifs suffisants pour pouvoir estimer une prévalence suffisamment précise d'un point de vue statistique. Certaines variables présentant de nombreuses modalités, telles que la religion ou l'ethnie, nécessiteront ainsi, le plus souvent, un recodage préalable. Nous ne détaillerons pas plus. Il s'agit, somme toute, des règles statistiques usuelles d'utilisation d'une base de données quantitatives.

Par ailleurs, il importera de vérifier qu'il n'y ait pas eu de taux de participation au dépistage du VIH particulièrement faible pour une ou plusieurs modalités de la variable choisie. Sur ce point, cette vérification est équivalente à celle réalisée pour l'estimation des prévalences régionales. Pour les variables sociodémographiques usuelles (âge, niveau d'instruction, quintile de bien-être…), ces taux de participation sont publiés dans le rapport final de chaque enquête et sont généralement du même ordre d'une catégorie à une autre. Comme pour les prévalences régionales, le biais dû aux personnes non testées restera dans la majorité des cas limité.

Lorsqu'il s'agit de déterminer les catégories les plus touchées ou d'établir un pattern relatif (niveau des catégories les unes par rapport aux autres), la présence de biais n'est pas problématique tant que l'impact de ces derniers est le même d'une catégorie à une autre. Pour des variables standards telles que le sexe, le niveau d'instruction ou le statut matrimonial, l'effet des différents biais peut être considéré comme généralement faible. Par contre, dès lors qu'il s'agit d'étudier les liens entre prévalence du VIH et des variables telles que la mobilité, nous sommes confrontés aux limites propres au type d'enquêtes que sont les EDS. En effet, la variable explicative s'avère, dans le cas présent, un déterminant des biais. Ainsi, les personnes mobiles sont beaucoup moins susceptibles d'avoir été enquêtées dans le cadre d'une EDS que les non migrants. Dès lors que les variables mises en relation avec la prévalence du VIH ont un impact direct sur le recrutement des individus enquêtés, les résultats obtenus à partir d'une Enquête Démographique et de Santé doivent être interprétés avec prudence. L'étude des relations complexes entre infection à VIH et migrations requiert ainsi l'élaboration d'enquêtes spécifiques à même de rendre compte de la diversité des profils migratoires.

La variable âge est utilisée dans de nombreuses modélisations. Elle est déterminante pour comprendre nombre de phénomènes démographiques et épidémiologiques. Or, dans des pays où l'état civil n'est pas efficient, les populations ne connaissent pas nécessairement leur âge avec précision. Ce phénomène est connu de longue date. Deux tendances sont fréquentes : l'attraction aux âges ronds et le rajeunissement ou le vieillissement systématique de certaines classes d'âges (WUNSCH 1984, p. 181). Plusieurs indices ont été élaborés pour rendre compte de cette attraction à certains âges, notamment l'indice de Myers ou l'indice combiné des Nations Unies (ONU 1984, GENDREAU 1985). Lorsque les individus ne connaissent pas leur âge, les enquêteurs ont recours à différentes techniques pour estimer celui-ci, telles que le recours à un calendrier historique, un positionnement par rapport aux membres de la famille ou bien une estimation à partir de l'histoire génésique des individus en dernier recours (AYAD 1997 p. 4). Du fait de ces décalages, il est préférable d'utiliser les groupes d'âges quinquennaux plutôt que les âges annuels. Outre le fait qu'une partie des erreurs sera compensée, les prévalences par âge seront calculées sur des effectifs plus importants et donc de meilleure qualité.

En 1990, une étude a été menée sur la qualité des données de la première série d'Enquêtes Démographiques et de Santé menées dans les années 1980 (INSTITUT FOR RESOURCE DEVELOPMENT 1990). Si la structure par groupe quinquennal d'âges présentait quelques distorsions, ces dernières étaient inférieures à celles observées dans la série des Enquêtes Mondiales de Fécondité menées pendant les années 1970. Il est vraisemblable que la qualité des données d'âges se soit améliorée dans les dernières EDS par rapport à celles conduites dans les années 1980. En effet, les plus jeunes générations connaissent aujourd'hui mieux leur âge que leurs aînés. Les pyramides des âges des populations enquêtées dans les EDS en Afrique subsaharienne présentent par ailleurs un profil usuel pour ce continent (AYAD 1997, p. 61).

L'âge renseigné dans les questionnaires individuels est en général de meilleure qualité que celui de l'enquête ménage puisque la question a été posée directement à l'individu concerné. Or, l'âge renseigné lors du recensement des individus d'un ménage détermine l'éligibilité pour le questionnaire individuel et le dépistage du VIH. Sont enquêtés le plus souvent les femmes de 15-49 ans et les hommes de 15-59 ans[5]. Les décalages d'âges autour de ces valeurs limites peuvent avoir un impact plus ou moins important. Les femmes de 45-49 ans sont ainsi le plus souvent sous-représentées. Les résultats pour ce groupe d'âges seront donc à interpréter avec prudence. De même, certaines analyses sont conduites en détaillant avec précision les âges entre 15 et 24 ans. Encore une fois, une analyse préalable de la qualité des données de chaque enquête et des biais d'enregistrement d'âges sera nécessaire

[5] Cela peut varier d'une enquête à une autre.

avant de travailler à ces échelles. Des simulations ont été réalisées, sous diverses hypothèses, pour essayer de déterminer l'impact des effets d'exclusion, en raison des décalages aux âges limites, sur trois indicateurs : l'indice conjoncturel de fécondité (ICF), le quotient de mortalité infanto-juvénile (avant cinq ans exact) et la prévalence de l'usage d'une méthode contraceptive chez les femmes, à partir de 22 EDS réalisées entre 1986 et 1993 (INSTITUT FOR RESOURCE DEVELOPMENT 1990, p. 21-23). Il en ressort que l'impact sur les estimations de ces trois indicateurs reste faible, inférieur à 1 % dans la majorité des cas et systématiquement inférieur à 5 %[6].

Si les variables sociodémographiques usuelles (sexe, âge quinquennal, niveau d'instruction, indice de bien-être…) sont relativement bien renseignées dans les EDS, la situation est plus complexe concernant les variables comportementales et notamment celles portant sur la sexualité. Il s'agit de variables déclaratives portant sur des domaines intimes et socialement normés. Ce type de biais n'est pas spécifique aux EDS mais fréquent dans toutes les enquêtes abordant la sexualité. À la fin des années 1980, il apparaissait dans les premières EDS menées en Afrique subsaharienne que, si les âges au premier mariage et à la première naissance étaient relativement bien renseignés, l'âge à la première relation sexuelle pouvait présenter plusieurs inconsistances (BLANC 1990). Certains résultats suggèrent que la question concernant l'âge au premier rapport sexuel peut être mal interprétée par les personnes enquêtées qui déclarerait alors préférentiellement l'âge de leur premier rapport avec leur conjoint ou partenaire actuel plutôt que leur première relation sexuelle (MEEKERS 1995). Se pose par ailleurs la définition d'un acte sexuel. Si pour une majorité cela sera entendu comme une relation pénétrative, la pratique du coït n'est pas la seule pouvant définir un rapport sexuel. Certains enquêtés pourront alors déclarer l'âge auquel ils ont connu leur première pratique sexuelle avec un partenaire bien qu'elle soit non pénétrative. D'autres évoqueront leur premier coït bien qu'ils aient déjà entamé leur vie sexuelle. Il reste néanmoins difficile de tirer des conclusions précises sur la validité des données sur la sexualité (DARE 1994). Au Zimbabwe où les hommes et les femmes déclarent un âge au premier rapport sexuel plus élevé que dans d'autres pays africains, ce résultat pourrait être s'expliquer en partie par une forme de déni de la sexualité pré-maritale (ZABA 2002). Ce phénomène serait accentué si l'adolescent est toujours scolarisé dans la mesure où des sanctions existent contre toute activité sexuelle pendant la scolarisation.

Si les déclarations sur la fréquences des rapports sexuels entre conjoints correspondent assez bien (FERRY 1995), il semble que les femmes auraient tendance à sous-estimer le nombre de leurs partenaires sexuels (BUVE 2001b). Selon une autre étude, les femmes sont plus discrètes concernant leur sexualité non maritale que les hommes, bien que cette tendance ne soit pas universelle (NNKO

[6] Il s'agit de l'écart relatif entre la valeur ajustée et la valeur observée de chaque indicateur.

2002). Les femmes célibataires ont plus tendance à sous-déclarer leurs relations que les femmes mariées, certains types de relations étant plus susceptibles que d'autres d'être sous-déclarés. Les hommes célibataires ont pour leur part tendance à exagérer le nombre de leurs relations. Par ailleurs, la qualité des déclarations des répondants peut varier au cours du temps du fait que les personnes enquêtées acceptent plus ou moins de déclarer leurs partenaires pré- ou extra-maritaux du fait des contraintes sociales (CURTIS 2004). Une étude récente, sur le statut sérologique au sein des couples à partir des EDS, montre une proportion importante de couples séro-différents où la femme est séropositive. Cette proportion ne peut être expliquée exclusivement par d'autres facteurs que la sexualité extraconjugale. Il en ressort alors une contradiction avec les très faibles déclarations de sexualité hors couple par les femmes. L'auteur suggère que la sexualité extra-maritale parmi les femmes mariées doit être plus importante que ne le montre les données et que le schéma épidémiologique usuellement avancé, selon lequel ce sont les hommes infidèles qui propagent l'épidémie des groupes à risques vers la population générale, doit être revisité (DE WALQUE 2007). Au final, une étude fine des déclarations relatives aux comportements sexuels à partir des EDS relèvera différentes incohérences ou contradictions, plus ou moins marquées (GERSOVITZ 2005).

Si les enquêtes en population générale restent une source importante et valable d'information sur les comportements sexuels, des précautions doivent être prises pour le suivi des évolutions comportementales. Les résultats doivent être interprétés avec prudence en tenant compte du contexte de chaque pays et du déroulement de l'enquête (CURTIS 2004).

*
**

Les EDS sont une source pertinente pour calculer la distribution des prévalences du VIH selon les caractéristiques sociodémographiques usuelles (âges quinquennaux, sexe, statut matrimonial…). Cependant, concernant l'analyse des déterminants comportementaux ou migratoires de la prévalence, une étude préalable de la qualité interne des données de chaque enquête est nécessaire, ce type de variables pouvant être soumis à de nombreux biais, notamment de sélection ou d'erreurs de déclaration.

5.1.3 Population générale et populations spécifiques

Dès le début des épidémies de VIH, certaines sous-populations spécifiques ont présenté des taux de prévalence élevés, en particulier les hommes ayant des rapports sexuels entre hommes (HSH) et les consommateurs de drogues injectables (CDI) dans les pays du Nord ou les travailleuses du sexe (TS) dans les pays du Sud.

Pour les pays à épidémie concentrée, l'ONUSIDA réalise des estimations séparées selon ces trois sous-groupes de population (à l'aide du logiciel EPP ou de la méthode Workbook, voir Encadré 1.7 page 64) avant d'estimer la prévalence nationale.

Bien que, dans les pays à épidémie généralisée, les estimations de la prévalence nationale ne requièrent pas une connaissance des prévalences au sein de chacune de ces sous-populations, de nombreuses enquêtes ont été menées parmi les groupes dits à risques (TS, patients atteints d'IST, plus rarement HSH[7]…) ou auprès de populations « passerelles » (chauffeurs routiers, militaires, clients des TS, réfugiés…), nommées ainsi en raison de leur « potentialité » supposée à relayer l'épidémie de certains groupes à risques vers la population générale. L'étude de ces populations est donc nécessaire pour mieux comprendre la dynamique épidémique et élaborer des programmes d'action ciblés.

Trois problématiques majeures se posent lors de l'étude de ces populations : la définition des individus appartenant à la population considérée, l'estimation de la taille de la population et la représentativité des personnes enquêtées. Définir les contours d'une population est complexe.

La prostitution peut prendre de nombreuses formes. « *Il est bien sûr possible d'établir que la "prostitution" renvoie généralement à des services sexuels hétérosexuels offerts par des femmes. […] Les caractéristiques traditionnellement attachées à ces pratiques sexuelles, promiscuité, absence de liens affectifs et compensation monétaire, ne permettent pas d'identifier les pratiques associées au commerce des services sexuels d'une société à l'autre. Celles-ci sont marquées par l'hétérogénéité. Au Nigéria, parmi certaines populations, par exemple, l'homme qui veut se marier doit verser le prix de l'épouse et acquiert ainsi des droits sur la sexualité de sa femme. Celle-ci peut développer une relation avec un (ou plusieurs) amant si celui-ci verse un montant annuel au mari. Qui plus est, cet amant fait régulièrement des dons à la femme. Malgré le caractère économique marqué de la relation, cette pratique n'est pas considérée comme de la "prostitution".* » (PARENT 1994)

Plusieurs définitions de l'homosexualité[8] existent, en termes de désirs, de pratiques ou bien encore d'identités sociales. Le paradigme LBGT[9] (Lesbiennes Gays

[7] À notre connaissance, une seule étude épidémiologique a été publiée à ce jour concernant les HSH en Afrique : WADE A. S., KANE C. T. *et. al.*, « HIV infection and sexually transmitted infections among men who have sex with men in Senegal », *AIDS*, n°19(18), 2005.
La problématique des HSH sur ce continent est récente. Plusieurs projets de recherche se développent actuellement sur cette question, mais ils restent marginaux.

[8] Nous ne parlerons ici que d'homosexualité masculine, mais la question de la définition de l'homosexualité féminine est équivalente.

Bisexuels Transsexuels) renvoie à des constructions identitaires occidentales, inadaptées pour rendre compte de la diversité des homosexualités à travers le monde (DOWSETT 2006). Cette terminologie est inadéquate, par exemple, en Afrique. Dans chaque culture, des termes identificateurs autres existent et pourraient être utilisés. Les études épidémiologiques s'intéressent aux pratiques sexuelles et, de ce fait, la terminologie utilisée par l'ONUSIDA est celle d'HSH qui porte sur les pratiques. Mais faut-il inclure uniquement des individus ayant des rapports sexuels exclusivement avec des hommes ou bien également ceux qui entretiennent également une sexualité avec des femmes ? Usuellement, la seconde approche prévaut. Reste à définir si le critère d'inclusion consiste à avoir eu au moins un rapport sexuel avec un autre homme au cours de sa vie ou si l'échantillon est restreint aux individus ayant eu récemment des pratiques homosexuelles. Cela sélectionne des populations très différentes. En effet, en Asie et en Amérique Latine, plusieurs études mettent en évidence que 6 à 20 % des hommes auraient eu au moins un partenaire masculin au cours de leur vie. Cette proportion diminue entre 2 et 8 % pour les hommes ayant eu au moins un partenaire masculin au cours des douze derniers mois (CACERES 2006)[10].

Si la taille de certaines populations peut être estimée facilement grâce à des enquêtes spécifiques (pour les camps de réfugiés par exemple) ou des données administratives (pour déterminer le nombre de militaires), il n'en est pas de même pour d'autres groupes. Il n'est pas possible de dénombrer directement le nombre d'HSH, de CDI ou de TS. Si dans certains pays, il existe des registres sanitaires de travailleuses du sexe, ceux-ci sont incomplets. Les estimations de la taille des populations resteront donc indirectes.

Le point précédent renvoie également à l'absence, le plus souvent, de base de sondage appropriée pour échantillonner des enquêtes portant sur ces sous-populations. Le recrutement des personnes enquêtées se fait donc à partir d'autres formes d'échantillonnage n'assurant plus la représentativité statistique vis-à-vis de la population visée, comme la sélection via des lieux identifiés (consultation de santé destinée aux travailleuses du sexe, bars fréquentés par une clientèle précise, arrêts routiers pour les chauffeurs…) ou la méthode dite « boules de neige » (recrutement de proche en proche par réseaux de connaissance). De fait, seule la « pointe visible de l'iceberg » est observée. Une analyse précise des biais de sélection est alors nécessaire avant toute interprétation des résultats. Les données d'enquête peuvent fournir des indications. Par exemple, l'étude, menée en 2004 au Sénégal auprès des homosexuels masculins, (WADE 2005) a concerné essentiellement des hommes jeunes (67 % ont 28 ans ou moins). Or, ces derniers

9 Par endroits remplacé par LGBTIQ (Lesbiennes, Gay, Bisexuels, Transsexuels, Intersexués, Queer).

10 Pour une synthèse, voir LARMARANGE J., « Hommes ayant des rapports sexuels avec d'autres hommes (HSH) : une épidémie toujours active », *Transcriptases*, n°129, 2006.

sont nombreux à avoir déclaré des partenaires plus âgés. L'écart entre la structure par âge des personnes enquêtées et celle de leurs partenaires sexuels permet d'estimer les biais de sélection. De plus, dans cet exemple, ne sont enquêtés que des individus acceptant plus ou moins bien leur homosexualité et se reconnaissant dans la définition d'HSH. Les personnes vivant leur sexualité dans le déni n'ont que peu de chance d'accepter de participer à ce type d'études. Bien que les enquêtes retiennent une définition comportementale pour l'inclusion des individus, la participation des personnes reste dépendante d'une perception identitaire[11].

Les enquêtes nationales en population générale pourraient permettre de contourner un certain nombre des points méthodologiques évoqués ci-dessus. D'une part, leur échantillonnage leur garantit la représentativité statistique. D'autre part, n'étant pas spécifique à une sous-population donnée, elles incluent des individus qui ne se situent pas dans une logique identitaire par rapport à certaines de leurs pratiques.

Mais ces enquêtes comportent d'autres limites. Pour les EDS, nous avons montré que les populations hors ménage ordinaire n'était pas prise en compte. Si cela a peu d'influence sur certaines sous-populations comme les TS ou les HSH qui font partie pour la majorité d'entre eux de la population en ménage ordinaire[12], cela peut affecter d'autres sous-groupes de manière importante. Par exemple, les camps de réfugiés ne sont pas pris en compte, de même que les militaires vivant en camps.

De plus, nous ne disposons pas, le plus souvent, des variables qui permettraient d'extraire de l'enquête certains sous-groupes. Aucune EDS ne pose de question concernant le nombre de partenaires de même sexe. Dans certaines, il est possible d'identifier les hommes ayant eu un rapport avec une travailleuse du sexe, voire les femmes ayant déclaré avoir eu au moins rapport sexuel tarifé. Mais, comme pour d'autres variables, il peut y avoir des erreurs de déclaration. Les personnes ayant eu une IST sont *a priori* identifiables. Cependant, il s'agit d'une variable déclarative et certaines personnes peuvent avoir contracté une IST sans le savoir.

Enfin, même lorsqu'il est possible d'identifier les individus appartenant à la sous-population visée, il faut tenir compte des faibles effectifs que cette dernière vat représenter sur l'ensemble de l'échantillon. À partir de la profession, il est théoriquement possible d'extraire de la base de données les chauffeurs routiers. Néanmoins, ils seront trop peu nombreux pour pouvoir mener une analyse fine.

[11] Le même problème peut se poser pour d'autres populations, par exemple les travailleuses du sexe. Les femmes ayant des rapports sexuels monnayés mais refusant de se définir comme prostituées ou TS seront moins susceptibles de participer à une étude.

[12] Ce point est à nuancer pour les travailleuses du sexe suivant l'organisation sociale de la prostitution dans chaque pays. Par endroits, elles peuvent vivre dans des baraquements ou des « foyers » qui ne seront pas enquêtés. Ailleurs, elles auront un logement ordinaire les incluant dans l'échantillonnage.

Une population peut néanmoins être identifiée aisément dans les EDS et présenter des effectifs suffisant pour y calculer une prévalence du VIH : il s'agit des femmes enceintes ou de celles ayant donné naissance au cours des douze derniers mois. Si elles sont usuellement suffisamment nombreuses pour calculer leur prévalence au niveau national, les effectifs se réduisent très rapidement dès lors qu'il s'agit de produire des indicateurs régionaux ou locaux.

⁂

Les enquêtes démographiques et de santé ne sont pas adaptées pour étudier les sous-populations usuellement désignées comme « groupes à risques » ou « populations passerelles ». En effet, même lorsque ces populations vivent au sein de ménage ordinaires et que l'enquête fournit les variables adéquates pour les identifier, les effectifs sont le plus souvent trop faibles pour mener une analyse statistique.

Des enquêtes spécifiques restent donc nécessaires, malgré leurs imperfections, pour suivre les tendances de l'épidémie et les changements de comportements au sein de ces sous-groupes particuliers.

5.2 La surveillance sentinelle : un indicateur possible des tendances

5.2.1 « Zones de recrutement » des cliniques prénatales

Dans le chapitre 3, nous avons mis en évidence que la prévalence observée en clinique prénatale pouvait être un indicateur de l'ordre de grandeur de la prévalence de la population adulte (hommes et femmes) à un niveau local (section 3.4.2). Cependant, nous n'avons pas précisé l'échelle correspondant à ce niveau local. Pour cela, nous pouvons avoir recours au concept anglo-saxon de *catchment area* ou « zone de recrutement ». Cette dernière peut être définie comme l'aire géographique et la population au sein desquelles un prestataire de services attire des clients[13]. La zone de recrutement d'une clinique prénatale correspondra donc au bassin de population consultant dans cette clinique donnée.

Bien que le terme de *catchment area* apparaît dans plusieurs études sur les prévalences observées en clinique prénatale (FYLKESNES 1998, KWESIGABO 2000, ZABA 2000, SAPHONN 2002), peu de travaux ont porté sur la détermination des zones de recrutement de ces cliniques. Arnaud FONTANET et ses collaborateurs ont mené une analyse à Addis-Abeba en Éthiopie où ils ont pu comparer les résultats d'une enquête en population générale avec ceux de cliniques prénatales dont ils avaient déterminé les zones de recrutement à partir de l'adresse des femmes enceintes testées (FONTANET 1998). Les prévalences observées en clinique prénatale étaient systématiquement supérieures, dans cette étude, à celles mesurées en population générale parmi les femmes. Par ailleurs, la prévalence du VIH variait d'une zone de recrutement à une autre, à la fois en population générale et parmi les femmes enceintes, montrant que l'épidémie n'est pas uniforme au sein d'une agglomération urbaine.

Nous disposons, pour le Burkina Faso et le Cameroun, de peu d'informations pour appréhender l'aire géographique couverte par les zones de recrutement des cliniques sentinelles. La méthode des cercles de même effectif nous a permis de reproduire les variations spatiales de la prévalence du VIH à des échelles infrarégionales. Nous avons précisé dans le Chapitre 4 que les niveaux produits par notre cartographie devaient être interprétés avec prudence. Ce que la carte retranscrit avant tout sont les différentiels d'une région à une autre, les niveaux

[13] Le terme de « client » doit s'entendre ici dans son sens le plus général, c'est-à-dire en tant que consommateur d'un service, monnayé ou non.

estimés pouvant être plus contrastés que les niveaux réels. Néanmoins, la prévalence estimée en un point donné reste un indicateur de l'ordre de grandeur de l'épidémie au sein de la zone définie par le cercle de lissage. En comparant les prévalences estimées par la méthode des cercles de même effectif et celles de la surveillance sentinelle des femmes enceintes, nous comparons alors la situation de l'épidémie entre l'aire géographique définie par le cercle de lissage et celle de la zone de recrutement de la clinique considérée.

Le Tableau 4.5 page 256 présente, pour les principales agglomérations urbaines du Burkina Faso et du Cameroun, les prévalences observées dans l'EDS et en cliniques prénatales. Pour les agglomérations de taille moyenne, les prévalences mesurées en clinique prénatales sont relativement proches[14] (par exemple Bertoua ou Ebolowa). Les écarts sont plus marqués pour les plus petites agglomérations. Cependant, pour ces dernières, les effectifs de personnes testées dans les EDS sont très faibles. Les prévalences qui y sont observées reflètent alors essentiellement l'erreur aléatoire de l'échantillonnage.

Dans le Tableau 5.2, nous comparons les prévalences observées en cliniques prénatales et celles estimées par la méthode des cercles (sans prise en compte du paramètre U) pour plusieurs agglomérations de taille moyenne. Pour la majorité, la prévalence des sites sentinelles est supérieure à celle estimée par la méthode des cercles. Sur la cartographie que nous avons réalisée, Bertoua, Ebolowa et Ouahigouya, lorsque le paramètre U est pris en compte, présentent un pic épidémique local. Si Tenkodogo avait été retenu[15] pour U, il aurait également présenté un pic local. Nous pouvons alors considérer que la zone de recrutement des cliniques de ces différentes agglomérations couvre un espace géographique plus restreint que celui du cercle du lissage. Les différences observées ici ne résultent pas d'une différence épidémique mais d'une différence d'échelle. Les mesures effectuées en cliniques prénatales dans une petite agglomération couvrent donc, de manière simplifiée, l'agglomération elle-même et une zone rurale plus ou moins large autour de celle-ci.

[14] Nous avons réalisé un test de Fisher pour comparer les prévalences deux à deux pour chaque agglomération. Aucune différence n'est significative à 20 %, excepté pour Douala (valeur de p : 0,01203). Cependant, les effectifs sont souvent trop faibles pour qu'un test statistique ait du sens.

[15] La prévalence observée dans l'EDS à Tenkodogo étant de 2,9, cette hypothèse pourrait être retenue, bien que ce soit avec plus de prudence que pour les autres agglomérations sélectionnées.

Tableau 5.2
Prévalences du VIH en clinique prénatale et estimée par la méthode des cercles (sans U) pour quelques agglomérations de taille moyenne

Agglomération urbaine	Prévalence du VIH en clinique prénatale (%)	Prévalence du VIH par la méthode des cercles (%)	Rayon des cercles de lissage (km)
Cameroun			
Bertoua	9,0	8,5	90
Ebolowa	11,6	7,2	74
Foumban	7,3	4,3	55
Dschang	4,3	2,6	33
Limbé	5,6	9,9	47
Burkina Faso			
Ouahigouya	3,6	0,7	89
Tenkodogo	2,6	1,5	60
Fada N'Gourma	1,3	2,0	96

Limbé présente une prévalence mesurée en cliniques prénatales largement inférieure à celle estimée par la méthode des cercles. Ville côtière, elle est située dans une région très urbanisée, à proximité de Buea et de Douala. Les zones de recrutement des cliniques sentinelles sont, de ce fait, plus complexes. Les femmes se sachant séropositives peuvent, par exemple, aller consulter dans une ville voisine pour des raisons de discrétion ou pour bénéficier d'un programme de prévention de la transmission mère-enfant. De la même manière, des femmes des localités voisines de Limbé peuvent venir y consulter. Il est fort probable, en milieu fortement urbanisé, que les zones de recrutement des différentes cliniques s'interpénètrent et se superposent en partie. L'hypothèse, que la surveillance sentinelle soit un indicateur local de l'ordre de grandeur de l'épidémie, peut alors ne pas être vérifiée.

Concernant les grandes agglomérations, les prévalences observées dans les EDS pour Yaoundé, Ouagadougou et Bobo-Dioulasso sont proches de celles mesurées par la surveillance sentinelle (Tableau 4.5 page 256). Par contre, pour Douala, la prévalence de l'EDS 2004 s'élève à 4,4 % alors qu'elle était de 8,0 % en 2002 en cliniques prénatales. Cette année-là, la surveillance sentinelle a porté sur 3 sites où 400 femmes ont été testées : Bonaberi (prévalence de 16 % sur 100 femmes), Congo II (3 % sur 100 femmes) et Deido (6,5 % sur 200 femmes), d'après (NATIONAL AIDS CONTROL COMMITTEE 2003). Les données de 2000 sont, quant à elles, assez différentes : une prévalence de 3,9 % a été mesurée à partir de 76 femmes recrutées sur deux sites, 21 sur le site de Deido et 55 sur celui de Newbell (NATIONAL AIDS CONTROL COMMITTEE 2001). La prévalence de 4,4 %, mesurée dans l'EDS, a été calculée, pour sa part, à partir de 931 individus répartis en 43 grappes.

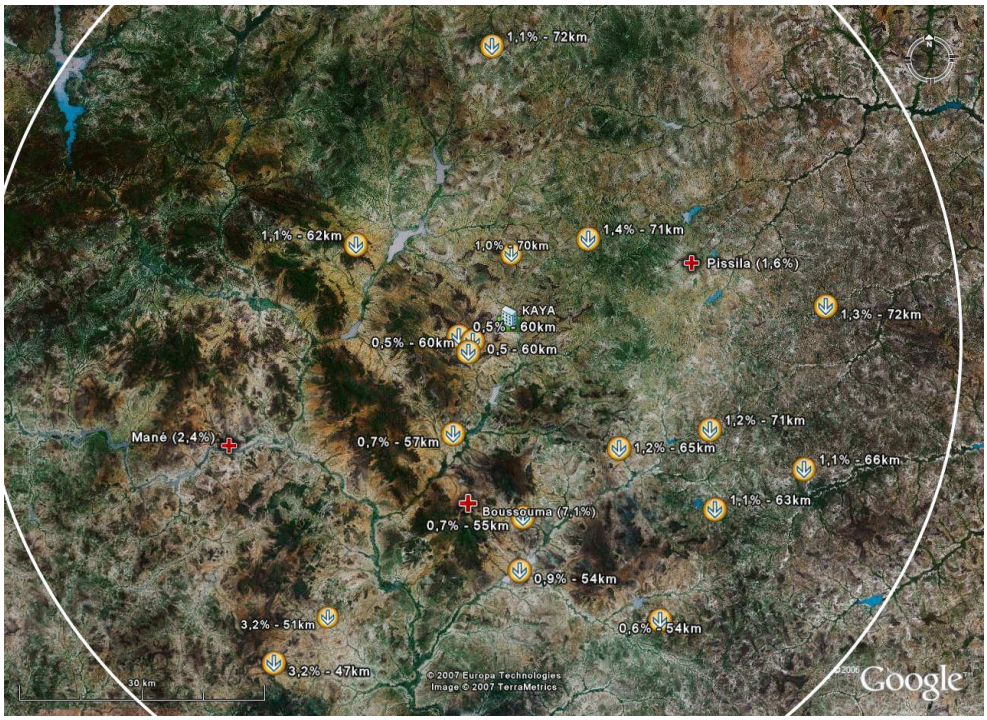
Les résultats de la surveillance sentinelle reflètent des écarts importants d'une clinique à l'autre, traduisant en partie les variations intra-urbaines de l'épidémie. D'autres études ont observé des variations comparables au sein d'une même agglomération, au Rwanda par exemple (GOTANÈGRE 1993) ou encore en Éthiopie (FONTANET 1998). Un changement dans les cliniques sélectionnées pour représenter la ville aura donc un impact sur la prévalence mesurée. Par ailleurs, dans une agglomération importante, l'offre de santé est relativement élevée. Les femmes peuvent donc avoir recours à différents services pour se faire suivre. Plusieurs facteurs peuvent influer le choix d'une clinique plutôt qu'une autre. La proximité par rapport au domicile ou au lieu de travail en est un. Le coût des soins en est un autre. Il est également connu que les cliniques proposant un programme de prévention de la transmission mère-enfant ont tendance à attirer une part plus importante de femmes séropositives. Or, la surveillance sentinelle porte sur un nombre limité de cliniques prénatales qui ne sont donc pas nécessairement représentatives de l'agglomération.

Abordons enfin la question des cliniques situées en zone rurale au travers de l'exemple de Kaya au Burkina Faso. Une prévalence de 3,6 % a été mesurée, en milieu rural autour de Kaya, lors de l'enquête sentinelle de 2003 (CNSL BURKINA FASO 2003). Plus précisément, l'enquête a porté sur trois sites de prélèvements : Boussouma, à une vingtaine de kilomètres de Kaya (prévalence de 7,1 % sur 156 femmes), Pissila, également à une vingtaine de kilomètres (1,6 % sur 191 femmes), et Mané, à environ quarante kilomètres (2,4 % sur 126 femmes). Nous avons représenté ces trois sites sur la Figure 5.2 ainsi que les différentes grappes enquêtées lors de l'EDS 2003 du Burkina Faso. Pour chacune d'elle, nous avons précisé la prévalence estimée par la méthode des cercles de même effectif et le rayon du cercle de lissage.

Les prévalences estimées dans cette zone sont de l'ordre de 0,5 à 1,4 %, excepté pour les deux grappes situées au sud-ouest (3,2 %). Cependant, les prévalences estimées ne se situent pas à la même échelle. En effet, les rayons des cercles de lissage atteignent 50 à 70 kilomètres. Pour visualiser cette distance, nous avons tracé en blanc le cercle de lissage des grappes situées dans Kaya (rayon de 60 kilomètres). Ce cercle de lissage couvre toute la zone étudiée ici. Sur les 18 grappes situées à proximité de Kaya, nous calculons une prévalence de 0,7 % (effectif de 342, intervalle de confiance à 95 % : 0,2-2,5 %), correspondant à la situation globale de cette zone géographique.

Les prévalences mesurées à Priscilla et à Mané se situent dans cet intervalle de confiance. L'épidémie y est du même niveau que dans l'ensemble de la zone. Par ailleurs, nous pouvons noter que la valeur mesurée à Priscilla est proche de celle estimée dans les grappes qui lui sont voisines.

Figure 5.2
Prévalences mesurées en cliniques prénatales et prévalences estimées aux alentours de Kaya, en 2003, au Burkina Faso



Note : cette figure a été créée avec le logiciel Google Earth (http://earth.google.com/intl/fr/). Kaya, Mané, Pissila et Boussouma sont positionnées selon leurs coordonnées dans Google Earth. Les croix rouges représentent les cliniques sentinelles. La prévalence observée en 2003 est indiquée à côté du nom. Les ronds jaunes avec une flèche représentent les grappes de l'EDS 2003. La prévalence estimée avec la méthode N et le rayon du cercle de lissage sont indiqués à côté de chaque point. Nous rappelons que les grappes sont positionnées avec une marge d'erreur de 5 kilomètres. Concernant l'échelle située en bas à gauche de la figure, 30 kilomètres correspondent à la totalité du segment. Le cercle blanc indique le cercle de lissage des trois grappes situées à Kaya. Son rayon est de 60 kilomètres.

La valeur élevée attribuée à la région de Kaya dans le cadre de la surveillance sentinelle provient de Boussouma où une prévalence de 7,1 % a été observée. Sur les 18 grappes de la Figure 5.2, aucune n'a enregistré, dans les données brutes de l'EDS, de cas positifs, à l'exception d'une grappe de Kaya (1 cas positif sur 11 personnes testées) et la grappe la plus proche de Boussouma où deux personnes ont été dépistées positives sur les 19 testées (10,5 %). La prévalence observée à Boussouma dans le cadre de la surveillance sentinelle correspond donc à un pic épidémique local. Cependant, ce dernier disparaît lors d'un changement d'échelle. Dans un rayon de 55 kilomètres, la prévalence du VIH descend à 0,7 % (calculée sur 508 individus partir de l'EDS). Les cliniques prénatales rurales mesurent donc la situation de l'épidémie au niveau local. Cependant, elles ne rendent pas forcément compte de la situation à une échelle régionale.

⁂

La prévalence du VIH mesurée auprès des femmes enceintes dans une clinique prénatale est fortement dépendante de la zone de recrutement de cette dernière. Pour les petites agglomérations isolées, le nombre limité de cliniques induit que leur zone de recrutement correspond approximativement à l'agglomération et à son voisinage plus ou moins proche. En revanche, dans les grandes villes ou dans les régions très urbanisées, la diversité des lieux de santé à disposition des populations rendent les zones de recrutement des cliniques prénatales plus complexes. Elles peuvent s'interpénétrer et/ou se superposer. De ce fait, les cliniques sélectionnées pour la surveillance sentinelle ne seront pas forcément représentatives de l'agglomération étudiée. De même, en milieu rural, les cliniques sentinelles mesurent une prévalence très localisée qui peut diverger de la tendance régionale.

5.2.2 Variations temporelles de la surveillance sentinelle

La surveillance sentinelle s'avère ne pas être adaptée pour estimer le niveau national des épidémies : la localisation des cliniques prénatales sélectionnées ne leur procure pas nécessairement une représentativité adéquate. Nous avions évoqué précédemment que la prévalence mesurée auprès des femmes enceintes pouvait être un indicateur de l'ordre de grandeur de l'épidémie à un niveau local (section 3.4.2). À la section précédente, nous avons limité la portée de ce résultat. En effet, dans certains contextes, en particulier dans les grandes agglomérations urbaines ou dans les régions fortement urbanisées, les zones de recrutement de chaque clinique peuvent s'interpénétrer et/ou se superposer, certaines cliniques attirant de manière privilégiée certaines catégories de femmes.

Il est donc préférable d'avoir recours à des enquêtes nationales en population générale telles que les EDS pour estimer le niveau la prévalence nationale. Cependant, une majorité de pays ne dispose que d'une seule enquête de ce type. Avec un seul point de mesure, il n'est pas possible d'estimer la tendance, à la hausse ou à la baisse, de l'épidémie. Certains pays devraient disposer d'ici quelques années d'une seconde EDS avec dépistage du VIH. Il sera alors possible d'estimer la tendance de l'épidémie entre les deux enquêtes. Cependant, comme il s'agit d'enquêtes lourdes et coûteuses à mettre en œuvre, elles ne sont réalisées en moyenne que tous les 4 à 8 ans. Elles ne permettront donc pas une estimation des tendances à court terme. Or, ce type d'estimation est nécessaire pour suivre les évolutions des épidémies.

La surveillance sentinelle des femmes enceintes, pour sa part, est relativement aisée à mettre en œuvre, certains pays procédant à des enquêtes annuelles. Si elle ne constitue pas un bon indicateur des niveaux, peut-elle néanmoins être un indicateur des tendances ?

À un niveau local, nous pouvons raisonnablement supposer que la prévalence mesurée dans une clinique prénatale suit les mêmes tendances que celle de la population générale, à la condition que la zone de recrutement de cette clinique reste constante. En effet, si la prévalence sentinelle surestime ou sous-estime celle de l'ensemble des adultes d'un facteur *k* et que ce facteur *k* reste constant au cours du temps, alors l'évolution mesurée en clinique prénatale sera la même que celle de la population générale.

La surveillance sentinelle peut donc constituer un indicateur local des tendances de l'épidémie. Cependant, certaines vérification sont nécessaires. En effet, dans certains pays, les cliniques retenues pour la surveillance sentinelle peuvent changer d'une année sur l'autre concernant un même site. Notre raisonnement ne tient, pour sa part, qu'à la condition de disposer de séries temporelles pour les mêmes cliniques. Par ailleurs, certains évènements contextuels peuvent modifier significativement la zone de recrutement d'une clinique prénatale. Par exemple, l'amélioration des infrastructures de transport (construction d'une route…) peut permettre aux femmes rurales de venir consulter plus facilement, diminuant ainsi la prévalence observée sans que cela ne corresponde à une évolution de l'épidémie locale (SCHWARTLANDER 1999). Autre exemple que nous avons déjà évoqué, l'implémentation d'un programme de prévention de la transmission mère-enfant (PTME) aura tendance à attirer spécifiquement les femmes séropositives (THE UNAIDS REFERENCE GROUP ON ESTIMATES MODELLING AND PROJECTIONS 2006b). Il est alors préférable de scinder les observations en deux séries temporelles lorsqu'un évènement de ce type peut être identifié.

L'hypothèse que la surveillance sentinelle soit un indicateur de tendances tient-elle au niveau national ? Bien que la représentativité ne soit pas assurée, cette hypothèse est moins forte que celle considérant que les femmes enceintes constituent un indicateur du niveau. Nous pouvons considérer, *a priori*, que les tendances temporelles devraient être, en moyenne, plus homogènes que les tendances spatiales. Au Zimbabwe, par exemple, la prévalence des femmes enceintes âgées de 15-49 ans a diminué continuellement de 32,1 % à 23,9 % entre 2000 et 2005 (MAHOMVA 2006). Le même pattern s'observe chez les femmes enceintes de 15-24 ans sur la même période. Dans la cohorte en population générale située dans la province du Manicaland, un déclin statistiquement significatif a été observé entre 1998-2000 (23,0 %) et 2001-2003 (20,5 %). Ce déclin a été observé à la fois dans les petites villes, les grandes exploitations agricoles, les zones commerciales à proximité des routes et dans des villages ruraux plus reculés.

⁂

Nous ne disposons pas pour le moment des données adéquates pour vérifier la validité de cette hypothèse. L'arrivée d'une seconde série d'EDS avec dépistage du VIH sera à ce sujet particulièrement éclairante et permettra de confirmer ou d'infirmer notre propos. Dans l'attente, nous accepterons provisoirement le fait que la surveillance sentinelle soit un indicateur des tendances de l'épidémie au niveau national.

5.3 EPP : le compromis de l'ONUSIDA

5.3.1 Un modèle épidémiologique simple pour plusieurs usages

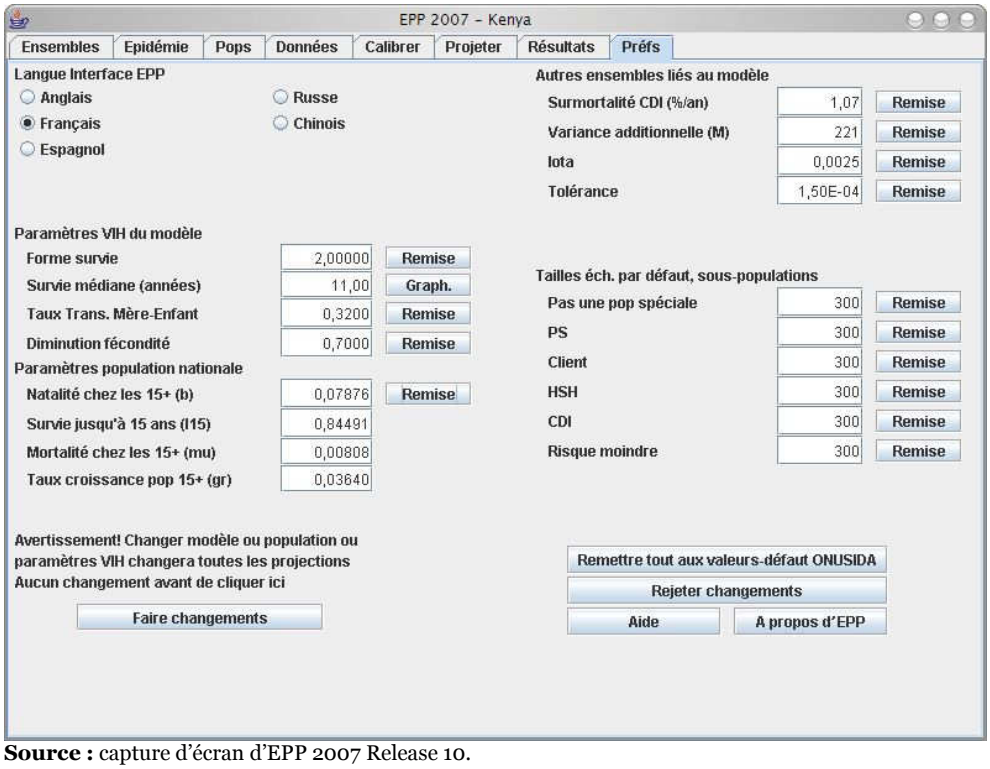
Nous avons déjà évoqué le logiciel EPP (Estimations and Projections Package) dans le Chapitre 1. Développé par l'ONUSIDA, sur les recommandations de son Groupe de Référence sur les Estimations, la Modélisation et les Projections, EPP a remplacé en 2002 EpiModel pour l'estimation des prévalences nationales pays par pays.

Il a été conçu pour pouvoir s'adapter à une grande variété de situations et répondre à plusieurs usages. Du fait d'une quantité et d'une qualité de données grandement variables d'un pays à l'autre, ils reposent sur un nombre limité de paramètres. La population des plus de 15 ans est divisée en trois groupes : les personnes sans risque d'être infectées, les personnes à risques et les personnes infectées. L'évolution de la taille de ces trois groupes est déterminée par trois équations différentielles qui définissent le modèle et le type de courbes de prévalences que ce dernier peut générer. La population n'est pas distinguée par groupe d'âges. Pour les pays à épidémie généralisée, deux sous-épidémies sont usuellement modélisées : l'une en milieu urbain et l'autre en milieu rural.

Une premier série de paramètres sont fixés dans le modèle. Ils peuvent être modifiés via l'onglet *Préfs* dans EPP (Figure 5.3) et permettent de préciser l'évolution de la population nationale, la mortalité liée au VIH et la transmission de la mère à l'enfant.

L'évolution démographique est définie par un taux de natalité chez les individus de 15 ans et plus, un taux de survie entre 0 et 15 ans pour les enfants non infectés (les enfants infectés sont supposés décéder avant 15 ans dans le modèle), un taux de mortalité (hors SIDA) des plus de 15 ans et un taux de croissance de la population adulte. Par ailleurs, sous l'onglet *Pops*, il faut définir la taille de la population nationale ainsi que celle de chaque sous-population (urbaine et rurale pour les pays à épidémie généralisée). Il est possible, au besoin, de fixer des paramètres d'évolution de la population différents entre le milieu urbain et le milieu rural.

Figure 5.3
Onglet préférences du logiciel EPP avec les paramètres par défaut



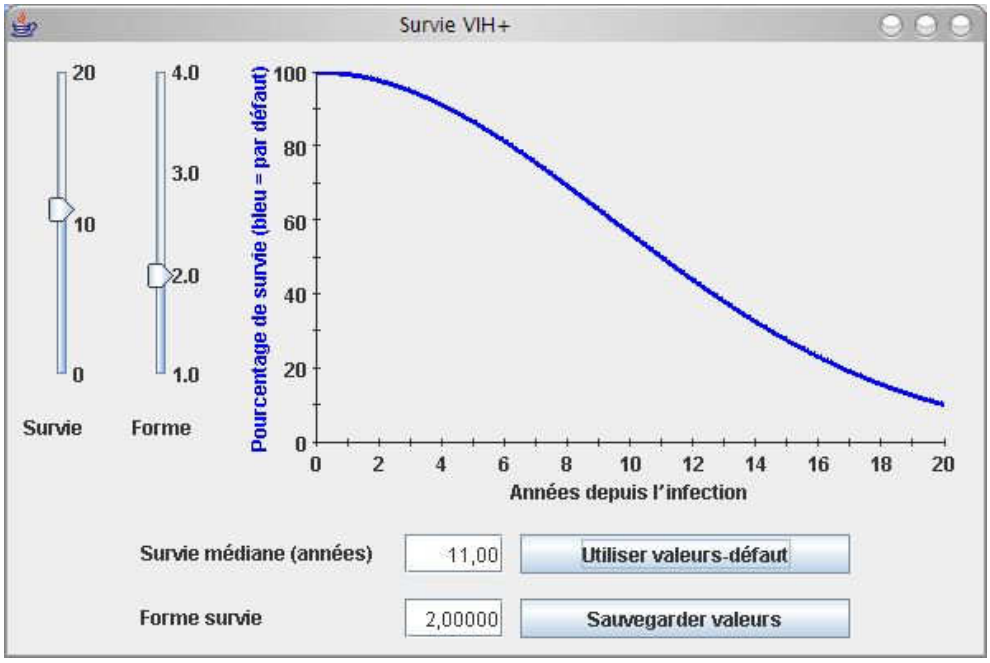
Source : capture d'écran d'EPP 2007 Release 10.

La mortalité des personnes infectées par le VIH est définie selon une courbe de survie en fonction de la durée d'infection. Il s'agit d'une courbe de Weibull (THE UNAIDS REFERENCE GROUP ON ESTIMATES MODELLING AND PROJECTIONS 2002) telle que la durée médiane de survie soit de 11 années[16]. La forme de la courbe et la durée médiane de survie peuvent être modifiées (Figure 5.4). Enfin, le modèle assume par défaut un taux de transmission du VIH de la mère à l'enfant de 32 % et une diminution de 30 % de la fécondité des femmes VIH+.

[16] À la création d'EPP, ONUSIDA considérait que la durée médiane de survie était de 9 années (THE UNAIDS REFERENCE GROUP ON ESTIMATES MODELLING AND PROJECTIONS, « Improved methods and assumptions for estimation of the HIV/AIDS epidemic and its impact: Recommendations of the UNAIDS Reference Group on Estimates, Modelling and Projections », *AIDS*, n°16(9), 2002.).

Plus récemment, suite à la publication de nouveaux résultats, le Groupe d'Experts a recommandé l'utilisation d'une durée médiane de survie de 11 ans (THE UNAIDS REFERENCE GROUP ON ESTIMATES MODELLING AND PROJECTIONS, *Improving parameter estimation, projection, methods, uncertainty estimation and epidemic classification*, Report of a meeting, Prague (CZ), 29 novembre - 1er décembre, UNAIDS, 2006a.).

Figure 5.4
Courbe de survie par défaut des personnes infectées par le VIH dans EPP



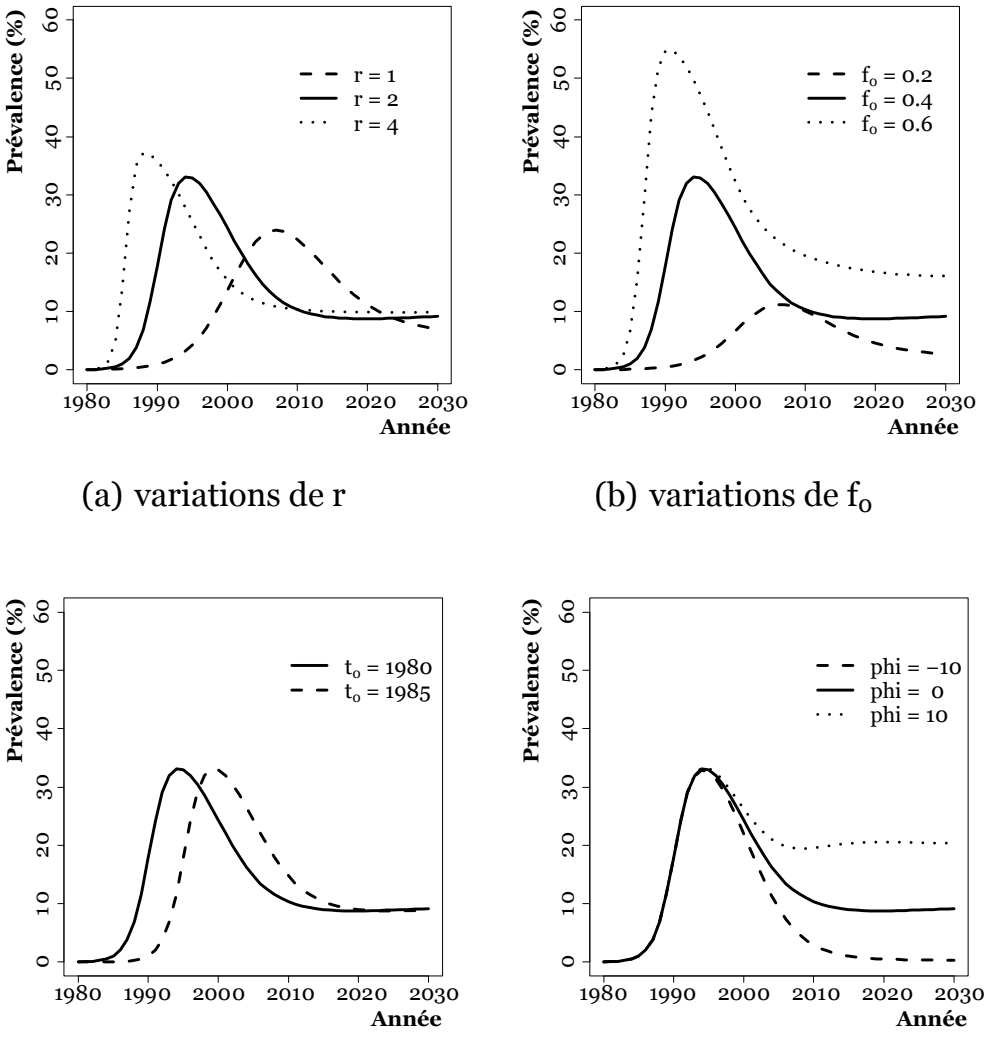
Source : capture d'écran d'EPP 2007 Release 10.

Une seconde série de quatre paramètres permettra d'ajuster la courbe des prévalences du modèle aux données observées parmi les femmes enceintes. Leur influence sur la courbe épidémique est représentée sur la Figure 5.5. Ces quatre paramètres sont :

- t_0 : l'année de démarrage de l'épidémie ;
- f_0 : la proportion de la population appartenant à la catégorie à risque au début de l'épidémie. Elle détermine en partie le maximum atteint par l'épidémie ;
- r : il s'agit du taux d'infection. Il détermine la rapidité à laquelle l'épidémie s'accroît. Si une fraction f de la population se situe dans la catégorie à risque, chaque personne infectée contaminera $r \cdot f$ individus au cours d'une année. $r \cdot f$ représente la « force » de l'épidémie ;
- φ : paramètre comportemental qui affecte la répartition des nouvelles entrées dans les catégories à risque et sans risques. S'il est nul, la proportion de personnes à risques parmi les nouveaux entrants (individus arrivant à l'âge de 15 ans) reste constante. S'il est négatif, cette proportion diminue. S'il est positif, cette proportion augmente. Ce paramètre influe sur la forme de la courbe après qu'elle ait atteint son maximum. Une valeur élevée de φ induira une stabilisation de l'épidémie à un niveau élevé.

Figure 5.5

Influence des paramètres r, f_o, t_o et φ sur la courbe épidémique produite par EPP



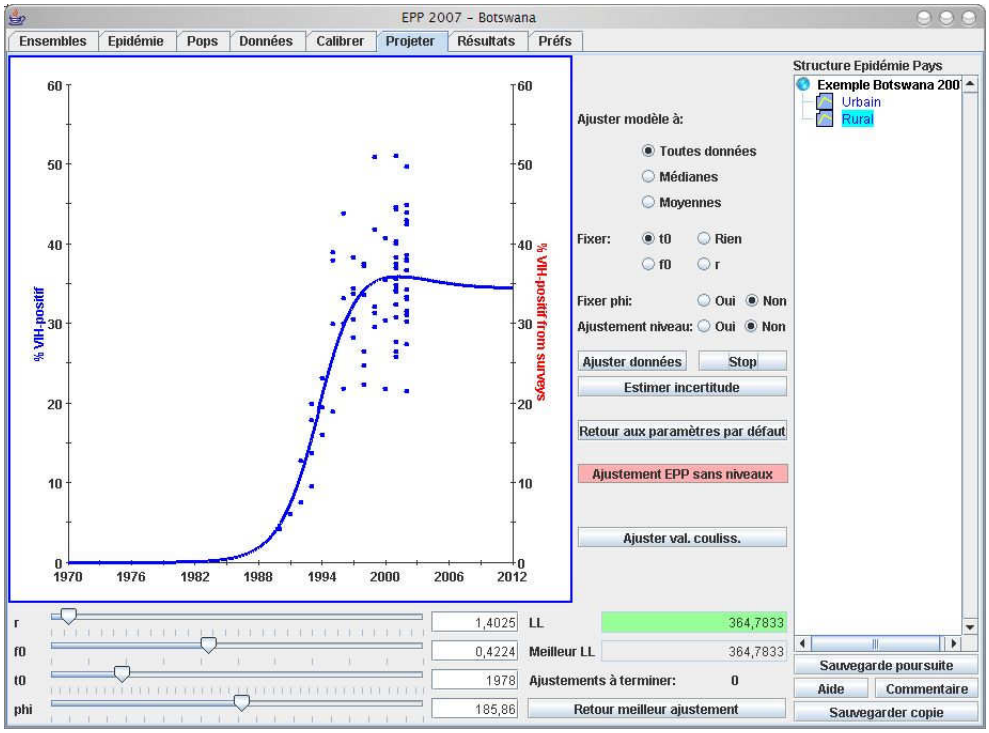
Source : d'après (ALKEMA 2007). Pour chaque graphique, la valeur des trois autres paramètres est fixe : 2 pour *r*, 0,4 pour f_o, 1980 pour t_o et 0 pour φ.

La courbe des prévalences du VIH est obtenue pour chaque sous-population en ajustant ces quatre paramètres (Figure 5.6). EPP procède par maximisation de la vraisemblance. Il est possible d'effectuer des ajustements par rapport aux valeurs médianes ou moyennes observées. Le plus usuel consiste à ajuster la courbe à l'ensemble des données observées (un point par site et par année). EPP permet de fixer éventuellement un ou plusieurs des quatre paramètres et d'effectuer l'ajustement uniquement en faisant varier les autres.

La prévalence nationale est ensuite calculée par addition des prévalences estimées pour chaque sous-population. EPP fournit au final, pour chaque année de la projection, la prévalence du VIH et le nombre d'individus infectés.

Figure 5.6

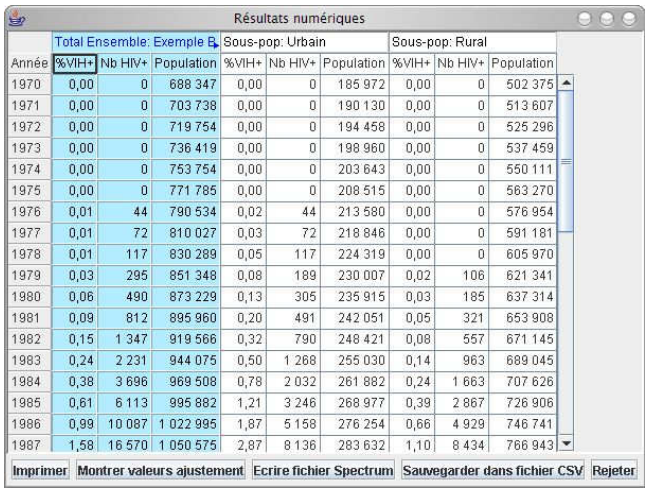
Ajustement du modèle EPP aux données de surveillance des femmes enceintes



Source : capture d'écran d'EPP 2007 Release 10.

Figure 5.7

Résultats générés par EPP



Année	Total Ensemble: Exemple B %VIH+	Nb HIV+	Population	Sous-pop: Urbain %VIH+	Nb HIV+	Population	Sous-pop: Rural %VIH+	Nb HIV+	Population
1970	0,00	0	688 347	0,00	0	185 972	0,00	0	502 375
1971	0,00	0	703 738	0,00	0	190 130	0,00	0	513 607
1972	0,00	0	719 754	0,00	0	194 458	0,00	0	525 296
1973	0,00	0	736 419	0,00	0	198 960	0,00	0	537 459
1974	0,00	0	753 754	0,00	0	203 643	0,00	0	550 111
1975	0,00	0	771 785	0,00	0	208 515	0,00	0	563 270
1976	0,01	44	790 534	0,02	44	213 580	0,00	0	576 954
1977	0,01	72	810 027	0,03	72	218 846	0,00	0	591 181
1978	0,01	117	830 289	0,05	117	224 319	0,00	0	605 970
1979	0,03	295	851 348	0,08	189	230 007	0,02	106	621 341
1980	0,06	490	873 229	0,13	305	235 915	0,03	185	637 314
1981	0,09	812	895 960	0,20	491	242 051	0,05	321	653 908
1982	0,15	1 347	919 566	0,32	790	248 421	0,08	557	671 145
1983	0,24	2 231	944 075	0,50	1 268	255 030	0,14	963	689 045
1984	0,38	3 696	969 508	0,78	2 032	261 882	0,24	1 663	707 626
1985	0,61	6 113	995 882	1,21	3 246	268 977	0,39	2 867	726 906
1986	0,99	10 087	1 022 995	1,87	5 158	276 254	0,66	4 929	746 741
1987	1,58	16 570	1 050 575	2,87	8 136	283 632	1,10	8 434	766 943

Source : capture d'écran d'EPP 2007 Release 10.

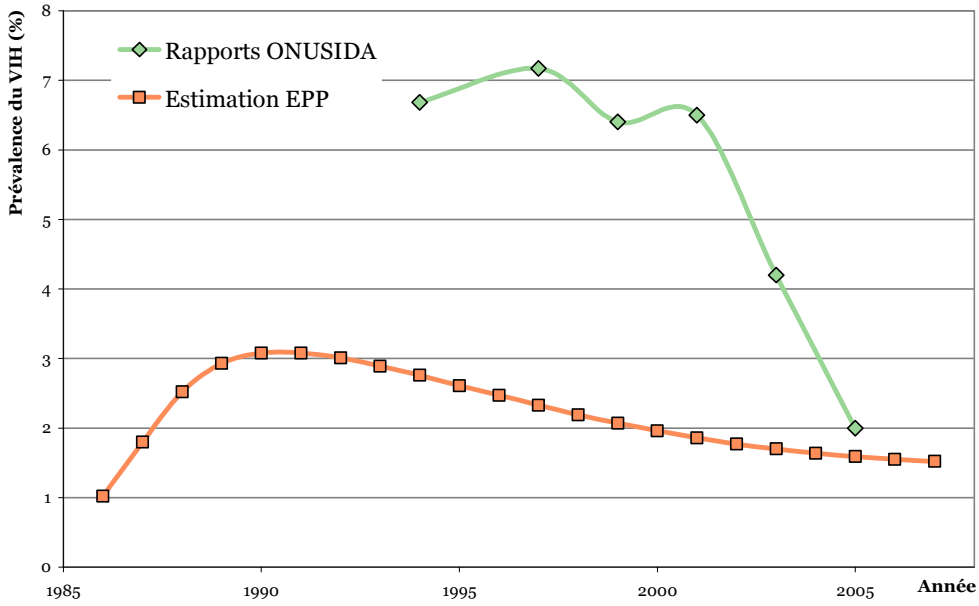
EPP répond à plusieurs usages différents. Tout d'abord, il utilisé à la fois pour des pays présentant une épidémie généralisée ou une épidémie concentrée. Pour les premiers, deux ajustements seront réalisés (urbain et rural). Pour les seconds, la prévalence sera estimée séparément dans différentes catégories de populations à

risques (travailleuses du sexe, hommes ayant des rapports sexuels avec d'autres hommes, consommateurs de drogues injectables…) et dans le reste de la population. Les projections produites par EPP sont utilisées par le logiciel Spectrum qui ventile les résultats obtenus par groupe d'âges afin d'estimer les incidences, la mortalité, le nombre d'orphelins, le nombre de personnes en besoin d'un traitement, etc. Enfin, il sert aux estimations biennales de l'ONUSIDA pays par pays.

5.3.2 Un modèle en évolution

EPP a connu plusieurs évolutions depuis 2002. Notre remarque pourra paraître triviale mais, de ce fait, les estimations réalisées par l'ONUSIDA dans ses rapports successifs ne peuvent être comparées pour déterminer les tendances de l'épidémie dans un pays. En effet, outre le fait que chaque estimation a été réalisée en prenant en compte une plus grande quantité de données, les méthodes ont évolué d'un rapport à l'autre. L'estimation réalisée avec EPP n'est pas ponctuelle : c'est une courbe qui est ajustée aux données. Afin d'éviter les confusions, il serait peut-être judicieux que l'ONUSIDA publie la courbe estimée et non simplement son estimation à la fin de l'année précédente.

Figure 5.8
Prévalences du VIH au Burkina Faso selon les différents rapports d'ONUSIDA et résultats d'une projection réalisée avec EPP



Sources : (OMS 1995, UNAIDS/WHO 1998, UNAIDS 2000, 2002, 2004a, 2006) pour les rapports ONUSIDA. La projection EPP a été réalisée avec EPP 2007 Release 10, à partir des données de surveillance sentinelle jusqu'en 2004, avec ajustement de niveaux et calibrage selon les résultats de l'EDS 2003.

À titre d'exemple, la Figure 5.8 présente les résultats d'une projection réalisée avec EPP à partir de données sur le Burkina Faso et les estimations correspondantes des rapports successifs de l'ONUSIDA.

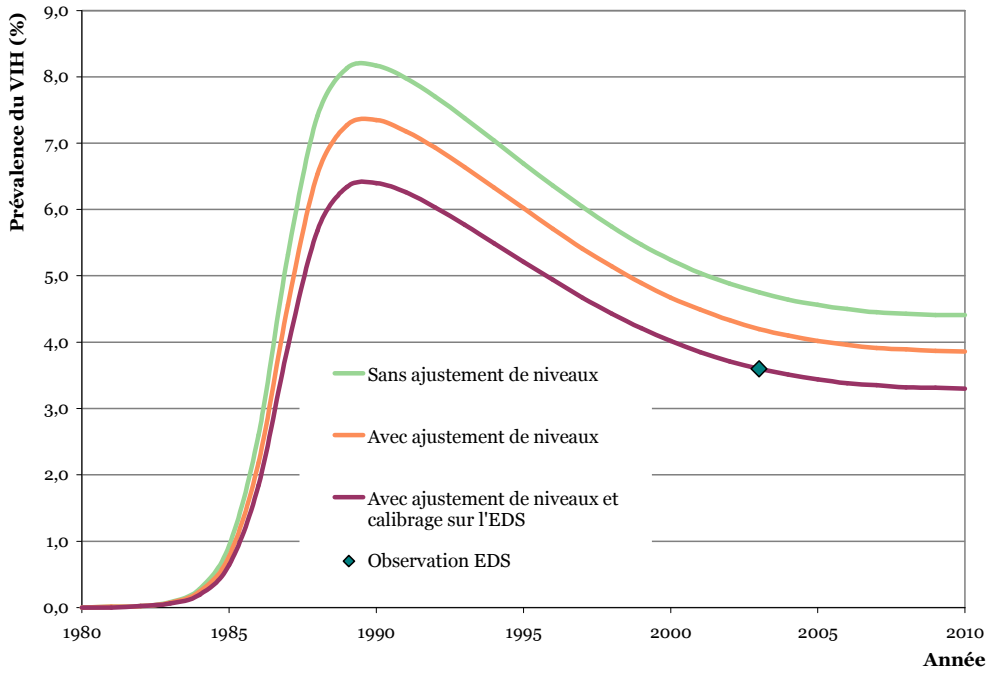
Dans les premières versions d'EPP, les courbes du modèle étaient simplement ajustées aux données de surveillance des femmes enceintes. De ce fait, les résultats souffraient de deux défauts majeurs. D'une part, nous avons montré en quoi la surveillance sentinelle ne constitue pas un bon indicateur du niveau des épidémies à l'échelle nationale. D'autre part, du fait de la variation du nombre et de la localisation des sites sentinelles au sein d'un pays donné, les tendances observées sur l'ensemble des sites sentinelles peuvent ne pas correspondre aux tendances réelles de l'épidémie. C'est typiquement le cas lorsque le système de surveillance sentinelle s'est élargi au milieu rural. Dans cette situation, pour les premières années de l'épidémie, nous ne disposons que d'observations présentant une prévalence élevée. Pour les années plus récentes, des observations présentant une prévalence moindre, provenant des nouveaux sites de surveillance, sont inclues dans le modèle. Un ajustement simple à l'ensemble des données induit alors une baisse artificielle de l'épidémie (voir la courbe en pointillé sur la Figure 1.9 B page 56).

Deux modifications apportées à EPP dans sa version 2005 permettent de contourner ces deux défauts. D'une part, la procédure *level fit* ou ajustement de niveaux qui tient compte des différentes séries temporelles. La courbe des prévalences de chaque site est supposée présenter la même forme mais à un niveau différent. EPP cherche alors à produire une courbe « moyenne » présentant les mêmes tendances que celles observées dans chaque site pris individuellement (voir Figure 1.9 page 56).

Par ailleurs, il devient possible de calibrer les courbes obtenues sur les résultats d'une enquête en population générale telle qu'une EDS. La projection EPP sera alors multipliée par un facteur d'échelle de telle sorte qu'elle soit égale à la prévalence mesurée dans l' EDS l'année de sa réalisation.

La Figure 5.9 présente, sur les données urbaines du Burkina Faso, comment la courbe des prévalences prédites par EPP est affectée par l'ajustement de niveaux et le calibrage sur les résultats de l'EDS. En procédant ainsi, EPP détermine le niveau des épidémies à partir d'une enquête nationale en population générale et les tendances à partir de la surveillance sentinelle des femmes enceintes.

Figure 5.9
Comparaison de trois projections EPP réalisées à partir des données de surveillance sentinelle en milieu urbain au Burkina Faso



Note : projections réalisées avec EPP 2007 Release 10. Les séries de données sentinelles des femmes enceintes vont jusqu'en 2004 (source : document de travail du CNLS non publiés).

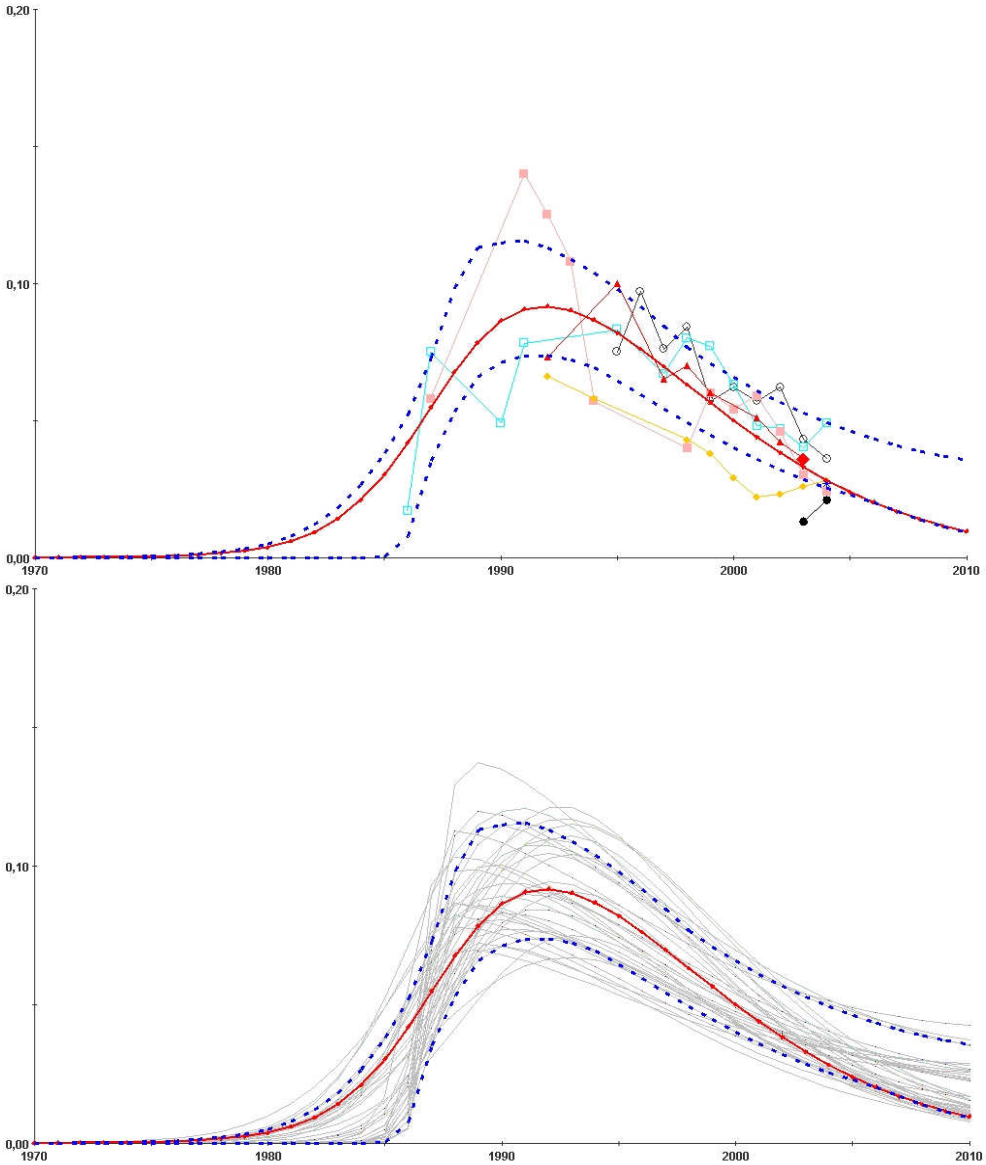
L'édition 2007 d'EPP propose deux nouveautés : l'analyse d'incertitude et la possibilité d'ajuster la courbe à partir de plusieurs enquêtes nationales en population générale (2 à 3).

L'analyse d'incertitude repose sur une approche statistique appelée *Bayesian Melding*. Son application à EPP a été discutée en 2006 par le Groupe de Référence de l'ONUSIDA (THE UNAIDS REFERENCE GROUP ON ESTIMATES MODELLING AND PROJECTIONS 2006a, 2006b). Elle a été détaillée dans un document de travail publié en janvier 2007 par Leontine ALKEMA et ses collaborateurs (ALKEMA 2007) avant d'être implémentée dans EPP en mars 2007 (Release 9).

Dans un premier temps, plusieurs dizaines de milliers de courbes (au minimum 50 000) sont générées en faisant varier les quatre paramètres r, t_0, f_0 et φ. Pour chacune d'elle est calculée sa vraisemblance par rapport aux données (points de couleurs sur la Figure 5.10). Puis, un sous-ensemble de ces courbes est ensuite tiré aléatoirement, parmi l'ensemble des courbes calculées, avec une probabilité d'échantillonnage dépendant de la vraisemblance. Ainsi, une courbe s'ajustant correctement aux données aura une probabilité importante d'être tirée au sort tandis qu'une courbe ne correspondant pas aux observations aura une probabilité presque nulle d'être sélectionnée.

Figure 5.10

Ajustement d'incertitude avec EPP sur les données urbaines du Burkina Faso



Source : captures d'écran d'EPP 2007 Release 10. Les courbes grises sur la figure du bas correspondent aux courbes tirées au sort à la seconde étape. La courbe rouge correspond au modèle suggéré par l'analyse d'incertitude. Les deux courbes en pointillées bleus correspondent aux quantiles à 2,5 % et 97,5 % des valeurs des courbes grises. Elles représentent ainsi l'intervalle de confiance de l'estimation de la courbe rouge. Les points de couleurs sur la première figure correspondent aux données observées. Les points ayant le même symbole et la même couleur proviennent du même site. Enfin, le losange rouge représente la valeur observée dans l'EDS du Burkina Faso en 2003.

À partir de cet échantillon (courbes grises), EPP peut alors calculer des paramètres moyens (produisant la courbe rouge) ainsi qu'un intervalle de confiance (courbes bleues en pointillées). Lorsque le résultat d'une EDS est entré (losange rouge), ce dernier est pris en compte du fait de l'application d'un facteur d'échelle aux

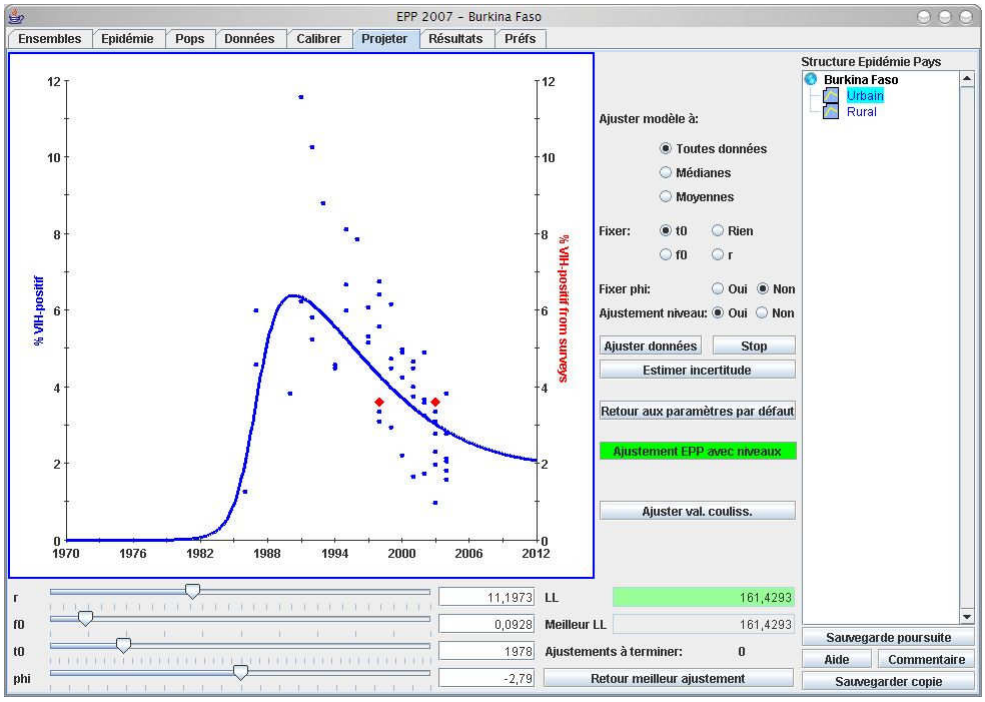
données sentinelles, bien que la courbe du modèle proposé par l'analyse d'incertitude ne passe pas exactement par le points de l'EDS.

Enfin, EPP permet, à partir de la Release 9 publiée en mars 2007, d'ajuster la courbe du modèle aux données de plusieurs enquêtes nationales en population générale (jusqu'à 3). Cette fonctionnalité a été discutée dès 2006 par le Groupe de Référence d'ONUSIDA (THE UNAIDS REFERENCE GROUP ON ESTIMATES MODELLING AND PROJECTIONS 2006b). Ce dernier recommande d'inclure les points de mesure en population générale dans l'ajustement et de calculer un paramètre ou facteur d'échelle spécifiant le biais de la surveillance sentinelle. La documentation d'EPP n'a pas encore été mise à jour et ne détaille pas comment EPP procède.

La projection produite avec un calibrage sur la seule EDS de 2003 (Figure 5.9) estime une baisse de la prévalence entre 1998 et 2003. Comment EPP se comporte-t-il si des enquêtes en population générale observaient une stagnation entre ces deux dates ? Sur la Figure 5.11, nous avons procédé à un ajustement sur les données urbaines du Burkina Faso en supposant que la valeur de 3,6 % observée dans l'EDS 2003 aurait également été observée dans une enquête nationale en population générale en 1998.

Figure 5.11

Projection EPP avec calibrage sur 2 EDS, Burkina Faso, milieu urbain



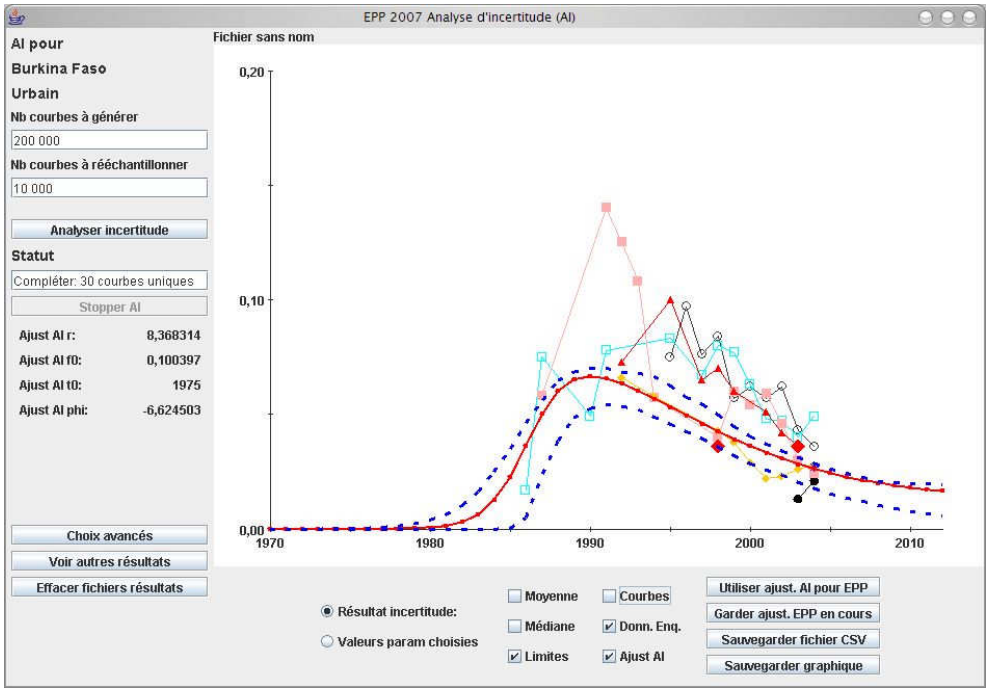
Source : capture d'écran d'EPP 2007 Release 10. Nous avons supposé qu'une prévalence de 3,6 % avait été observée en 1998 et en 2003.

La courbe produite conserve la même forme que précédemment, à un niveau situé entre les deux mesures en population générale. Il semble qu'un facteur d'échelle global soit calculé à partir des deux mesures en population générale et appliqué à la courbe. La tendance mesurée à partir des femmes enceintes est donc privilégiée à celle mesurée par les enquêtes en population générale.

Si nous procédons à une analyse d'incertitude (Figure 5.12), nous obtenons une courbe similaire. L'intervalle de confiance calculé n'inclut pas l'une des deux observations en population générale. Il semble que les résultats des EDS sont intégrés dans le calcul des vraisemblances. Cependant, nous ne savons pas exactement de quelle manière pour le moment (la documentation n'ayant pas encore été mise à jour).

Figure 5.12

Analyse d'incertitude avec calibrage sur 2 EDS, Burkina Faso, milieu urbain



Source : capture d'écran d'EPP 2007 Release 10.

Cette approche privilégie donc l'hypothèse que la surveillance sentinelle des femmes enceintes traduit correctement les tendances de l'épidémie. Pour notre part, il nous semble que cette hypothèse, que nous avons acceptée temporairement, nécessite d'être vérifiée à partir des résultats des enquêtes nationales en population générale. Ces dernières constituent de bons indicateurs de niveaux. De ce fait, la tendance observée entre deux enquêtes de ce type devrait être prise en compte. Certes, elles ne peuvent fournir la forme exacte de la courbe entre les deux observations. Mais, nous pouvons raisonnablement considérer que la courbe de l'épidémie devrait passer *a minima* au sein de leurs intervalles de confiance.

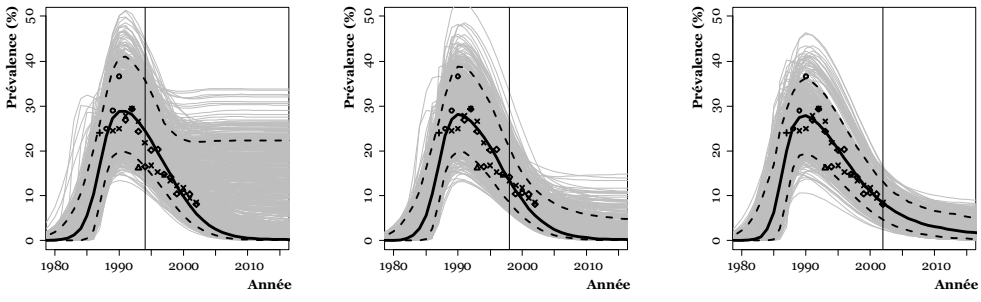
D'ici quelques années, plusieurs pays devraient disposer d'une seconde enquête nationale en population générale avec dépistage du VIH. Il devrait alors être possible de clarifier dans quelle mesure la surveillance sentinelle des femmes enceintes permet de traduire correctement ou non les tendances des épidémies au plan national. Le Groupe de Référence de l'ONUSIDA recommande néanmoins que la méthode d'échantillonnage de la surveillance sentinelle soit équivalente à travers le temps et les sites surveillés. Seuls des sites pour lesquels une série temporelle de données est disponibles (au moins trois points) devraient être sélectionnés pour une analyse de tendances (THE UNAIDS REFERENCE GROUP ON ESTIMATES MODELLING AND PROJECTIONS 2006b).

5.3.3 Un modèle qui montre ses limites

Outre les limites provenant des données de surveillance sentinelle des femmes enceintes, EPP repose sur l'hypothèse que son modèle épidémiologique simple est en capacité de rendre compte de la plupart des épidémies. Ce modèle implique que les courbes des prévalences du VIH estimées auront une certaine forme. L'un des atouts d'EPP a été de pouvoir s'ajuster de manière correcte aux données de certains pays bien documentés. Par ailleurs, pour ces derniers, les prédictions réalisées avec EPP ont été confirmées lorsque de nouvelles données étaient introduites dans le modèle. C'est le cas, par exemple, sur les données urbaines de surveillance sentinelle de l'Ouganda (Figure 5.13).

Figure 5.13

Projections sur les données urbaines de l'Ouganda avec ajouts successifs de données dans le modèle



(a) données jusqu'en 1994 (b) données jusqu'en 1998 (c) données jusqu'en 2002

Source : (ALKEMA 2007). Projections obtenues par analyse d'incertitude avec ajout progressif de données. La courbe noire représente la projection obtenue par l'analyse d'incertitude. Les courbes en pointillés représentent les intervalles de confiance. Les courbes grises correspondent aux courbes échantillonnées par l'analyse d'incertitude. Les points représentent les données observées.

L'intérêt d'un modèle épidémiologique par rapport à un simple ajustement géométrique des données est de calculer, pour des pays ayant peu de données, une projection dont les valeurs rétrospectives et prospectives[17] seraient plausibles. En effet, si un simple ajustement géométrique[18] permet de déterminer une courbe représentant correctement les données observées, les valeurs rétrospectives ou prospectives de cette courbe peuvent être totalement aberrantes vis-à-vis du phénomène étudié. Or, pour des estimations plus complexes comme celles de Spectrum, il importe de disposer d'une courbe des prévalences capable de rendre compte du démarrage d'une épidémie et de fournir une projection future raisonnable.

Cependant, le modèle implémenté dans EPP n'arrive pas à rendre compte de certaines épidémies observées aujourd'hui, en particulier lorsque la prévalence du VIH se stabilise après un déclin rapide. Des difficultés d'ajustement ont été observées sur des données du Kenya, du Zimbabwe, du Rwanda ou encore d'Éthiopie (THE UNAIDS REFERENCE GROUP ON ESTIMATES MODELLING AND PROJECTIONS 2007). La Figure 5.14 présente les données de surveillance sentinelle chez les femmes enceintes en milieu urbain au Kenya, selon l'*Epidemiological Fact Sheet 2006* (WHO 2006b).

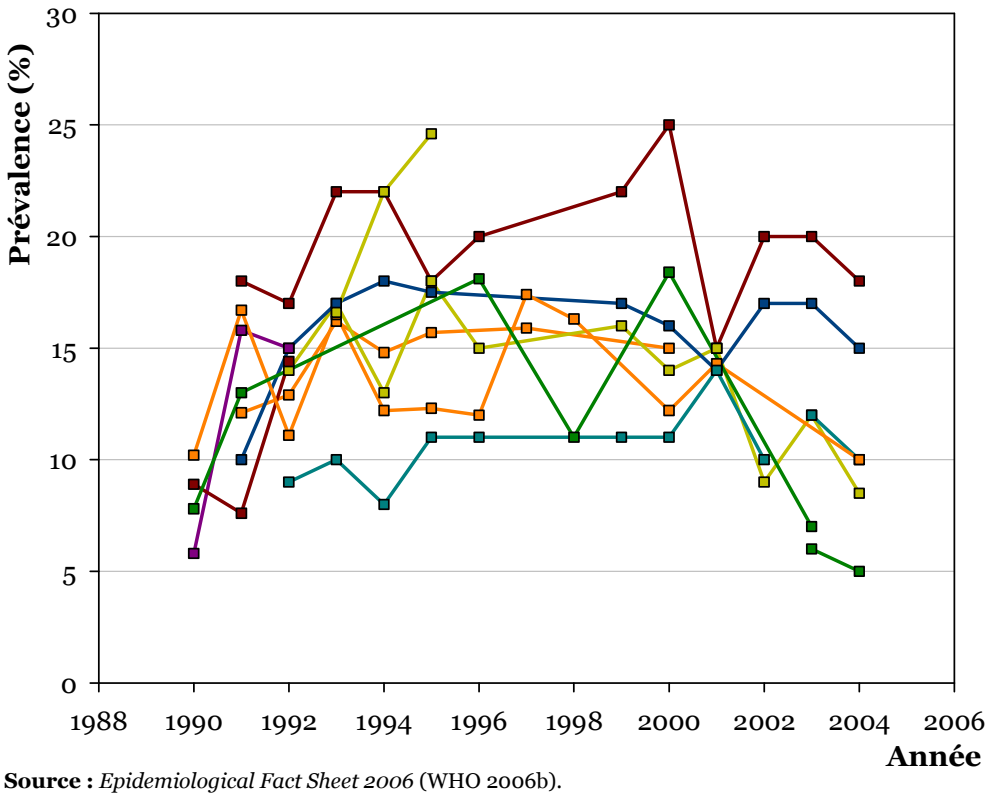
Nous avons ensuite réalisé une projection (Figure 5.15) sur ces données avec ajustement de niveaux et calibrage à partir des résultats de la dernière EDS (prévalence de 10,0 % en 2003). Les valeurs moyennes et médianes fournissent une première approximation, grossière, des tendances observées chez les femmes enceintes. EPP n'arrive pas à rendre compte du déclin rapide observé depuis 1998. La courbe estimée stagne à partir de 2002 tandis que les prévalences mesurées parmi les femmes enceintes continuent de décroître jusqu'en 2004 (année des dernières observations). EPP estime donc une prévalence de 9,9 % en 2007 alors que, si nous suivons les tendances observées chez les femmes enceintes, cette dernière devrait être plus faible.

[17] C'est-à-dire les valeurs estimées à des dates antérieures ou postérieures à celles des données disponibles.

[18] Nous entendons par ajustement géométrique l'ajustement d'une fonction mathématique usuelle aux données observées, que ce soit une droite (régression linéaire), une courbe polynomiale ou une autre famille de courbes. Dans ce type d'ajustement, aucune hypothèse n'est faite quant à la dynamique de l'épidémie, ses modes de transmission, etc. Il s'agit simplement d'identifier une fonction mathématique en capacité de s'ajuster le mieux possible aux observations.

Figure 5.14

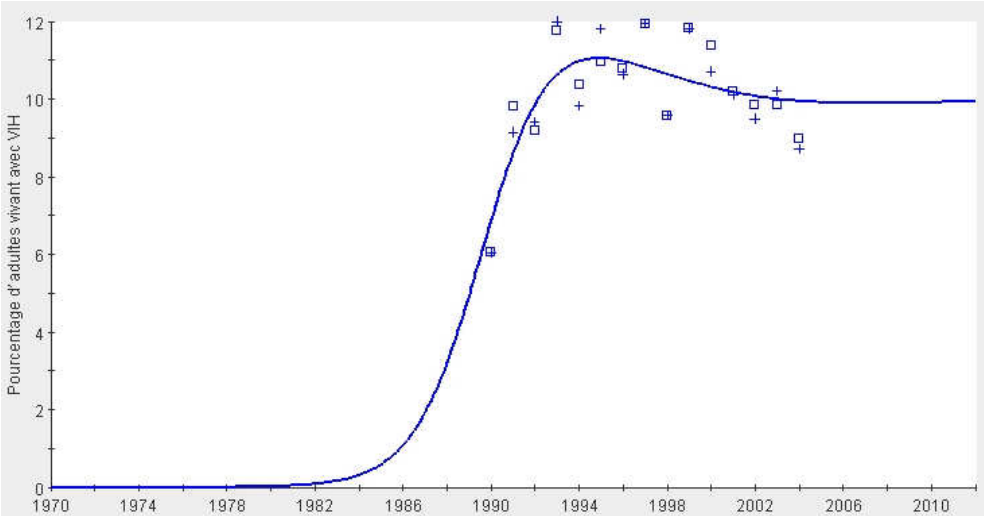
Données de surveillance sentinelle des femmes enceintes, milieu urbain, Kenya



Source : *Epidemiological Fact Sheet 2006* (WHO 2006b).

Figure 5.15

Projection EPP réalisée avec ajustement de niveaux et calibrage EDS sur les données urbaines du Kenya



Source : capture d'écran d'EPP 2007 Release 10. La courbe représente la projection estimée, les croix les valeurs médianes des prévalences sentinelles et les carrés leurs valeurs moyennes.

Afin de rendre le modèle implémenté dans EPP plus souple, le Groupe de Référence de l'ONUSIDA envisage de permettre à certains paramètres d'évoluer au cours du temps (THE UNAIDS REFERENCE GROUP ON ESTIMATES MODELLING AND PROJECTIONS 2007). Plusieurs pistes doivent être explorées :

1. *Modèle contraint :* correspond au modèle actuel, les paramètres conservant une valeur constante au cours du temps.
2. *Saut de r :* autorise un changement du paramètre r à une date donnée. Ce modèle requiert deux nouveaux paramètres : l'année du changement et la valeur du changement, r étant constant sur les deux périodes considérées.
3. *Changement progressif de r :* le paramètre r évoluerait progressivement sur une période de temps donnée. Trois nouveaux paramètres sont nécessaires : année de début et année de fin de la période de changement et un taux annuel de changement.
4. *Changement progressif de r et saut de φ :* scénario précédant incluant une modification de φ à une date donnée (5 nouveaux paramètres).
5. *Modèle non contraint :* les paramètres r et φ peuvent évoluer au cours du temps, indépendamment l'un de l'autre.

Il faudra attendre les prochaines versions d'EPP pour vérifier si ces changements permettront ou non un meilleur ajustement de la courbe aux données. Cette possibilité de faire varier les différents paramètres au cours du temps est évoquée depuis plusieurs années. Si, sur une période courte, il est raisonnable de considérer que ces paramètres restent relativement constants, ce n'est plus le cas dès lors que la période de temps étudiée devient plus importante. Le modèle implémenté dans EPP correspond à la situation d'une épidémie naturelle. Or, de nombreuses interventions ont eu lieu depuis 25 ans. Les comportements sexuels ont évolué, différemment selon les pays et les générations. La mise en place à grande échelle de programmes d'accès aux antirétroviraux vient progressivement modifier à la fois la mortalité des personnes infectées et les probabilités de transmission par acte. La possibilité de faire varier les paramètres d'EPP au cours du temps s'avère ainsi de plus en plus nécessaire.

Reste la problématique d'estimer les prévalences du VIH dans les pays où EPP s'avère, pour le moment, inadapté pour rendre compte des tendances de l'épidémie à court terme. Si un ajustement géométrique est inadapté pour réaliser une projection rétrospective et prospective de l'épidémie, ce type d'approche peut-il nous permettre d'estimer le niveau actuel d'une épidémie ?

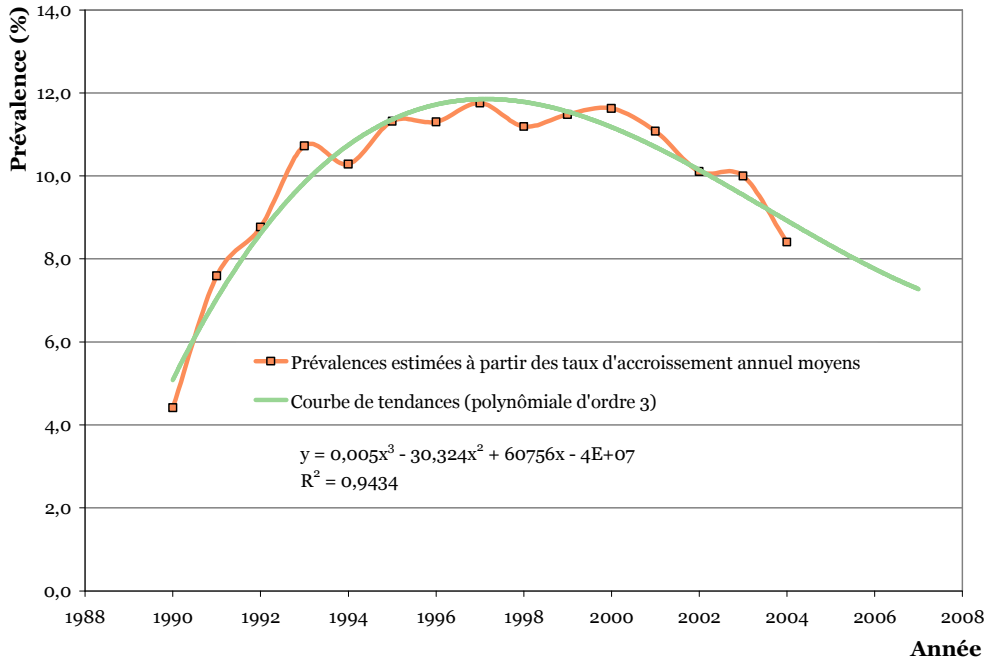
Nous avons montré dans ce chapitre que la prévalence mesurée dans une EDS était un bon indicateur du niveau (section 5.1) tandis que la surveillance sentinelle pouvait être considérée, au moins provisoirement, comme un indicateur de tendances (section 5.2). Nous pouvons alors envisager une approche purement

descriptive pour estimer la prévalence du VIH du milieu urbain kenyan en 2007, en posant un minimum d'hypothèses.

Considérant la surveillance sentinelle des femmes enceintes comme un indicateur de tendances, nous calculons, pour chaque site et chaque année, un taux d'accroissement annuel. Ce dernier traduit ainsi la tendance observée sur ce site. Puis, nous estimons les taux d'accroissement annuel de l'épidémie pour l'ensemble du milieu urbain en faisant la moyenne, pour chaque année, des taux d'accroissement annuel de chaque site. Connaissant le niveau de l'épidémie en 2003 à partir de l'EDS (10,0 %), nous calculons la courbe de l'épidémie à partir des taux d'accroissement annuel moyens et obtenons la courbe orange sur la Figure 5.16. Bien que cette courbe présente des irrégularités, elle fournit une image assez précise des tendances de l'épidémie. Nous ajustons ensuite une courbe de tendances par une simple régression.

Figure 5.16

Estimation géométrique de la prévalence urbain du milieu urbain kenyan



Dans notre exemple, nous avons opté pour une courbe polynômiale d'ordre 3. La prévalence en 2007 serait alors de 7,3 %. Plusieurs types de courbes peuvent être essayés avant d'en choisir une. Il est également possible d'effectuer un ajustement visuel en traçant une courbe de tendances à la main. Des résultats différents seront obtenus. Seule une connaissance de la situation locale peut permettre de déterminer quelle estimation sera la plus « probable » *a priori*. En effet, nous ne pouvons garantir notre résultat de 7,3 %. Cependant, il est, en l'absence d'autres informations, plus raisonnable de considérer que la prévalence du VIH en milieu

urbain au Kenya est de l'ordre de 7 % en 2007 plutôt que de 10 %. Il ne peut s'agir que d'un résultat provisoire qui pourra être confirmé ou infirmé ultérieurement lorsque de nouvelles données seront disponibles et, en particulier, lorsqu'une nouvelle enquête nationale en population générale sera réalisée.

Si, lorsque le logiciel EPP ne s'ajuste pas correctement aux données observées, ce type d'ajustement graphique peut se révéler plus adéquat pour estimer la prévalence à court terme, il importe de disposer de données récentes. Dans notre exemple, nous avons utilisé des données allant jusqu'en 2004 pour estimer la prévalence de 2007. Une durée prospective de trois ans s'avère ici élevée. En effet, si différents types de courbes produisent des résultats assez semblables dans le cadre d'une prospection à un ou deux ans, les résultats divergent très rapidement ensuite, surtout si les variations sont importantes parmi les données observées sur la période récente.

*
**

Les EDS s'avèrent être un bon indicateur du niveau des épidémies à l'échelle nationale. Elles permettent également d'estimer les prévalences selon le milieu de résidence, les régions et les variables sociodémographiques usuelles, après vérification des taux de participation. Par contre, elles ne sont pas adaptées pour l'étude de populations spécifiques et l'analyse des déterminants comportementaux de la prévalence doit rester prudente. Elles ne peuvent pour le moment être utilisées pour déterminer les tendances des épidémies puisque nous ne disposons que d'un seul point de mesure.

La surveillance sentinelle des femmes enceintes est en capacité de traduire les tendances locales des épidémies, à la condition que les zones de recrutement des cliniques étudiées restent stables dans le temps. Nous pouvons accepter provisoirement l'hypothèse selon laquelle ces tendances observées chez les femmes enceintes traduisent les tendances nationales de l'épidémie. Il importera néanmoins de confirmer ou d'infirmer ce fait lorsque nous aurons à notre disposition une seconde mesure de la prévalence nationale en population générale.

EPP, le logiciel de projection développé par l'ONUSIDA, a connu de nombreuses évolutions ces dernières années. En permettant un ajustement des niveaux et le calibrage du modèle sur les résultats des enquêtes en population générale, les tendances sont estimées à partir de la surveillance sentinelle tandis que le niveau est déterminé par une enquête telle qu'une EDS. En intégrant une analyse d'incertitude, à l'aide d'une technique nommée *Bayesian Melding*, il devient possible d'estimer un intervalle de confiance des projections.

Depuis peu, EPP autorise le calibrage du modèle à partir de plusieurs enquêtes en population générale. Cependant, il importera d'analyser plus en détails la manière dont EPP effectue ce calibrage, lorsque les enquêtes seront à disposition. Enfin,

EPP est limité par son modèle épidémiologique simple. En effet, ce modèle impose un certain type de courbes épidémiques qui ne peuvent rendre compte de ce qui est observé dans certains pays. Une des solutions en cours de développement consiste à rendre le modèle plus flexible en autorisant certains paramètres à évoluer dans le temps. Il faudra attendre quelques mois avant de pouvoir vérifier si les modifications apportées à EPP lui permettront ou non de s'ajuster correctement aux données des pays pour lesquels il échoue actuellement. En attendant, EPP s'avère inadapté pour estimer, à court terme, la prévalence nationale du VIH de ces pays. Il est alors possible d'avoir recours à un ajustement géométrique moins contraignant. Si les résultats de ce type d'ajustements doivent être interprétés avec prudence, ils restent *a priori* plus proches de la réalité qu'une estimation effectuée avec EPP. Par contre, pour les pays où la courbe produite par EPP s'ajuste raisonnablement aux données sentinelles, l'estimation réalisée avec ce dernier reste valable. Quelque soit l'approche utilisée, les estimations à court terme de la prévalence reposent toutes sur des hypothèses plus ou moins fortes concernant la capacité de la surveillance sentinelle à rendre compte des évolutions des épidémies. Pour chaque pays, il importera de vérifier la pertinence d'une telle hypothèse lorsque nous disposerons de plusieurs mesures de la prévalence en population générale.

Conclusion

Historiquement, la surveillance sentinelle des femmes enceintes a été mise en place en Afrique subsaharienne afin de suivre l'apparition de l'épidémie de VIH sur ce continent (voir Chapitre 1). Les femmes enceintes présentaient l'avantage de pouvoir être recrutées facilement, un dépistage anonyme du VIH étant réalisé à partir d'échantillons sanguins prélevés pour d'autres usages. Quelques enquêtes en population générale ont été menées à la fin des années 1980. Outre le fait qu'elles étaient lourdes et coûteuses à mettre en place, elles se heurtaient à des taux de refus de se faire dépister non négligeables. Elles ont été abandonnées tandis que la surveillance sentinelle s'est généralisée sur le continent.

En raison de leur disponibilité, les données de surveillance sentinelle des femmes enceintes sont devenues la source privilégiée pour l'estimation des prévalences nationales du VIH à partir du milieu des années 1990. L'OMS va alors développer EpiModel qui sera repris par l'ONUSIDA avant d'être remplacé par EPP en 2002.

À la fin des années 1990, les cliniques prénatales sélectionnées pour la surveillance sentinelle surreprésentaient, dans la majorité des pays, le milieu urbain. L'ONUSIDA va alors promouvoir, à partir de 2000, la surveillance sentinelle de seconde génération qui recommandait, outre un suivi comportemental des populations, l'extension de la surveillance au milieu rural.

La même année, le groupe de travail ONUSIDA/OMS sur la surveillance globale du VIH et des IST suggère la possibilité de tester des échantillons sanguins prélevés dans le cadre d'enquêtes nationales en population générale. En 2001 au Mali, est conduite la première Enquête Démographique et de Santé (EDS) avec dépistage du VIH. Comme les EDS sont réalisées régulièrement dans une majorité de pays

d'Afrique subsaharienne, le nombre d'EDS avec dépistage du VIH va se développer dans les années qui vont suivre. Plus d'une quinzaine d'enquêtes de ce type ont déjà été conduites et plusieurs autres sont en cours de réalisation.

Dans nombre de pays, les prévalences du VIH mesurées dans les EDS ont divergé sensiblement de celles estimées jusqu'alors par la surveillance sentinelle des femmes enceintes (voir Figure 1.8 page 52). Si ces résultats ont questionné la communauté scientifique internationale, EPP calibre depuis 2005 ses projections sur ce type d'enquêtes. Cependant, des taux de non réponse élevés dans les enquêtes en population générale pourraient constituer un biais important dans leurs estimations.

Reste que la question de savoir s'il faut privilégier les estimations réalisées à partir des femmes enceintes ou à partir d'enquêtes en population générale n'est pas tranchée. Il s'agit alors de déterminer la validité, la signification et la portée qui peuvent être attribuées à ces différentes mesures.

Dans le Chapitre 2, nous rappelons que tout énoncé d'observation n'est pas donné par lui-même mais naît de l'application, à des données d'observation brutes, de la relation répétable définissant un concept opératoire. Nos énoncés d'observation ont une portée limitée puisque leur ensemble population-espace-temps est restreint aux individus effectivement enquêtés, sur les zones d'enquête, le temps de l'enquête. Or, nos assertions portent sur des populations plus larges et, en l'occurrence, nos estimations de la prévalence nationale du VIH concernent l'ensemble de la population adulte d'un pays, répartie sur tout son territoire.

Quelle que soit la source de données utilisées, nous avons donc recours à des hypothèses anticipatrices afin de transformer nos énoncés d'observation en assertions plus générales. Dès lors, nous prenons un risque d'erreur radicale, plus ou moins important selon nos hypothèses. D'après notre Postulat 2.8 (page 102), le risque d'erreur radicale pourra être considéré comme minimisé si l'hypothèse anticipatrice posée a été vérifiée expérimentalement pour des cas analogues et/ou si plusieurs hypothèses anticipatrices différentes, appliquées à un même énoncé d'observation, conduisent aux mêmes énoncés déduits. Nous pouvons alors étudier la validité des différentes sources de données en analysant les hypothèses requises pour transformer leurs énoncés d'observations en assertion sur l'ensemble de la population.

Dans un premier temps, nous nous sommes consacré à la question de la représentativité de chaque source (Chapitre 3). Nous avons défini cette notion comme la possibilité d'extrapoler les résultats obtenus sur un échantillon à l'ensemble de la population dont il est extrait, un échantillon pouvant n'être représentatif que pour certaines variables et/ou que pour certaines caractéristiques (moyenne, variance, distribution…) de ces dernières.

Pour les enquêtes nationales en population générale de type EDS, nous avons analysé cinq sources de biais : populations hors ménages, ancienneté de la base de sondage, ménages non enquêtés, individus non testés et fenêtre sérologique des algorithmes de dépistage utilisés. L'ampleur de ces bais a été estimée et, bien qu'en maximisant ces biais, les prévalences ajustées pour le Burkina Faso, le Cameroun et le Kenya se situaient au sein des intervalles de confiance à 95 % des prévalences observées dans les EDS. Ces enquêtes sont donc adaptées pour estimer le niveau de la prévalence nationale. Par ailleurs, elles sont échantillonnées pour être représentatives par région et milieu de résidence. Par contre, en raison de leurs effectifs limités, les EDS sont inadéquates pour l'estimation des prévalences du VIH à un niveau local.

La surveillance sentinelle en cliniques prénatales, quant à elle, est soumise à de nombreuses sources de biais difficilement identifiables et quantifiables. Localement, la prévalence mesurée parmi les femmes enceintes peut être considérée comme une estimation *a minima* de la prévalence de l'ensemble des femmes et comme un indicateur de l'ordre de grandeur de la prévalence de la population générale adulte (hommes et femmes). Par contre, les patterns observés selon certaines caractéristiques sociodémographiques peuvent différer entre la surveillance sentinelle et la population générale.

Au niveau national, les estimations du niveau de la prévalence du VIH à partir des femmes enceintes seront plus ou moins proches de la prévalence réelle en fonction de la configuration particulière de la localisation des sites sentinelles retenus pour la surveillance et des variations spatiales de la prévalence au sein du pays considéré. Les biais peuvent ainsi se compenser ou, au contraire, se compléter. Dans certains pays, la prévalence estimée parmi les femmes enceintes s'est ainsi avérée jusqu'à quatre ou six fois plus élevée que celle observée en population générale.

Au Chapitre 4, nous avons développé une méthodologie afin d'estimer les tendances infrarégionales de la prévalence du VIH à partir des EDS. Pour cela, nous inspirant de techniques d'analyse spatiale en composantes d'échelle, nous avons eu recours à des cercles de même effectif pour estimer la tendance régionale de chaque zone d'enquête avant de procéder à une interpolation spatiale par krigeage ordinaire.

La simulation d'enquêtes sur un pays modèle nous a permis d'asseoir notre méthodologie et de déterminer des valeurs optimales pour nos paramètres de lissage. Notre approche présente l'avantage de poser un minimum d'hypothèses *a priori* sur la distribution spatiale de nos données et de maximiser l'information que nous pouvons en déduire. Par ailleurs, des cartes complémentaires permettent de faciliter l'interprétation des résultats en indiquant les zones d'incertitude.

Les résultats que nous avons obtenu sur les données des EDS du Burkina Faso et du Cameroun s'avèrent cohérents avec d'autres sources d'information. Ils suggèrent, par ailleurs, une diffusion de l'épidémie selon les principaux axes routiers et dans des zones présentant une activité économique particulière, comme ont pu le mettre en évidence d'autres travaux. Cependant, notre approche reste avant tout descriptive plus qu'explicative, les données à notre disposition n'étant pas suffisamment fines.

Au Kenya, nos résultats sont relativement proches de ceux obtenus par une autre équipe à partir d'hypothèses différentes. Notre approche est par contre plus aisée à mettre en œuvre, notamment grâce à la librairie logicielle prevR que nous avons écrite. Distribuée gratuitement sous licence libre, cette dernière constitue un outil à la disposition des programmes nationaux pour les aider à planifier leurs programmes d'action en identifiant les régions les plus touchées.

Le Chapitre 5 a été consacré à deux dimensions importantes de la mesure : les niveaux et les tendances. À l'échelle nationale, les EDS constituent un bon indicateur du niveau des épidémies. Cela reste vrai selon les régions, le milieu de résidence et les variables sociodémographiques usuelles, sous réserve d'une vérification des taux de participation.

Si la surveillance sentinelle est inadaptée pour estimer avec précision le niveau national des épidémies, elle fournit, pour sa part, des séries temporelles de données. Si les zones de recrutement des cliniques prénatales restent stables dans le temps, ces données peuvent être considérées comme un indicateur de tendances au niveau local. Nous pouvons, au moins temporairement, admettre qu'elles constituent également un indicateur des tendances au niveau national. Cependant, cette hypothèse nécessitera d'être vérifiée, dans différents contextes, lorsque nous disposerons, pour un même pays, d'au moins deux mesures de la prévalence en population générale.

En proposant un ajustement par niveaux et un calibrage sur les résultats des enquêtes nationales en population générale, le logiciel EPP de projection des prévalences du VIH, développé par l'ONUSIDA, utilise de fait la surveillance sentinelle pour l'estimation des tendances et les EDS pour l'estimation du niveau. Par ailleurs, la nouvelle procédure d'analyse d'incertitude implémentée en 2007 permet d'estimer les marges d'erreur d'une projection réalisée avec EPP.

L'inconvénient majeur d'EPP reste le modèle épidémiologique simple qu'il présuppose. Si ce dernier permet de réduire de manière importante le nombre de paramètres d'ajustement, il ne peut rendre compte de la diversité des épidémies observées. Le Groupe de Référence de l'ONUSIDA sur les Estimations, la Modélisation et les Projections travaille actuellement sur un assouplissement du modèle afin de permettre à certains de ses paramètres d'évoluer dans le temps. En attendant, il peut être préférable, pour une projection à court terme (un à trois ans)

de la prévalence du VIH, de procéder à un ajustement géométrique, c'est-à-dire par prolongation de la tendance observée sur la période récente, parmi les femmes enceintes.

L'arrivée d'ici quelques années d'une deuxième EDS avec mesure de la prévalence du VIH dans plusieurs pays d'Afrique subsaharienne devrait permettre de vérifier certaines des hypothèses du modèle implémenté dans EPP et notamment sa capacité à rendre compte des tendances de l'épidémie dans certains pays.

Du fait des hypothèses importantes qu'il nous est nécessaire de poser dès lors que nous travaillons, à une échelle nationale, sur des données sentinelles auprès de femmes enceintes, toute estimation réalisée à partir de celles-ci doit être considérée comme provisoire, dans l'attente d'une confirmation ou d'une infirmation éventuelle lorsque des données nouvelles seront disponibles.

Bien que nous ne soyons jamais à l'abri d'un risque d'erreur radicale, les résultats des enquêtes en population générale sont, pour leur part, plus solides, à condition de ne pas leur conférer une précision supérieure à celle qui est la leur. En raison de leur échantillonnage, ces enquêtes disposent d'une représentativité statistique que n'a pas la surveillance sentinelle. Malgré les discussions sur le sens épistémologique qui peut être accordé à la notion de probabilité, les méthodes d'échantillonnage probabilistes ont prouvé maintes fois, dans la pratique, leur efficacité. Certes, les EDS ne sont pas exempts de biais. Cependant, en ce qui concerne le dépistage du VIH, ces derniers restent relativement faibles. Il peut y avoir des exceptions comme la région de Lilongwe, au Malawi, pour l'EDS 2004. Mais, même dans ces situations, nous avons à notre disposition des méthodes d'ajustement qui permettent de redresser les résultats de manière raisonnable.

Les enquêtes nationales en population générale ont permis de clarifier et de préciser notre connaissance des niveaux épidémiques. Il faudra encore attendre quelques années pour savoir si elles modifieront également notre vision des tendances.

*
**

Si aujourd'hui nous disposons d'une meilleure image du niveau des épidémies de VIH sur le continent africain grâce, notamment, à l'apport des enquêtes en population générale, la connaissance des dynamiques épidémiques reste, quant à elle, fragmentaire.

Peu de mesures directes des incidences et de la mortalité sont disponibles sur le continent et les données existantes portent sur un nombre limité de cohortes. L'ONUSIDA utilise le logiciel Spectrum pour estimer, à partir des données de prévalences générées par EPP, des incidences par âge et le nombre de décès imputables au VIH/SIDA. Cependant, il s'agit, en l'occurrence, d'estimations indirectes. Outre les limites propres aux estimations réalisées avec EPP,

notamment concernant les projections rétrospectives et prospectives sur des périodes où nous ne disposons pas de donnée, les prévalences du VIH sont ventilées par groupe d'âges selon un pattern mesuré en population générale. Or, il est fort probable que ce dernier ait évolué au cours du temps. Celui que nous observons aujourd'hui peut différer de celui qui prévalait au démarrage des épidémies. La mortalité liée au VIH est modélisée en fonction de la durée d'infection. Pourtant, elle peut évoluer également en fonction de l'âge. Quant au pattern, il peut être amené à se modifier au fur et à mesure que l'épidémie évolue.

Spectrum demeure un outil précieux dans des contextes où les données sont peu nombreuses. Il permet de réaliser des projections raisonnables sur ce qu'a pu être l'épidémie par le passé et sur ses possibles tendances futures. Cependant, il reste dépendant des nombreuses hypothèses anticipatrices sur lesquelles il repose. Dans quelle mesure les patterns observés sur certaines cohortes peuvent être transposés sur d'autres populations et à d'autres échelles ? La simplification nécessaire du réel qu'opère Spectrum pour transcrire les relations entre différentes grandeurs mathématiques permet-elle de décrire avec suffisamment de précision la dynamique des épidémies ? De telles questions ne peuvent être résolues simplement. Cependant, une première piste pourrait consister à réaliser des projections avec Spectrum sur des cohortes de populations pour lesquelles nous disposons de mesures directes des nouvelles infections et de la mortalité : dans un premier temps, en calibrant le modèle sur les différents patterns mesurés dans la cohorte elle-même puis, dans un second temps, en utilisant les patterns par défaut de Spectrum. Ce type d'analyses permet de vérifier empiriquement, au moins localement, la validité et la portée d'un tel modèle.

Par ailleurs, d'autres pistes peuvent être explorées, s'appuyant sur notre Postulat 2.8. En effet, le risque d'erreur radical peut être considéré minimisé si plusieurs approches méthodologiques, reposant sur des hypothèses anticipatrices différentes, induisent des résultats similaires. Bien que d'un point de vue formel cela ne constitue pas une vérification empirique, une telle comparaison constitue un argument permettant d'affermir certains de nos résultats.

Or, si plusieurs modèles épidémiologiques ont été élaborés à la fin des années 1980, force est de constater qu'il y a peu de travaux sur ces questions aujourd'hui. Concernant l'estimation des incidences, nous pouvons citer la thèse de Charlotte SAKAROVITCH[1] sur les données de surveillance sentinelle des femmes enceintes à Abidjan, en Côte d'Ivoire. Elle détermine les incidences en modélisant la probabilité qu'une femme infectée soit observée en clinique prénatale, par

[1] SAKAROVITCH C., *Estimation de l'incidence de l'infection par le VIH en Afrique : application à la Côte d'Ivoire*, Thèse de Doctorat option Épidémiologie et Intervention en Santé Publique, Université Victor Segalen Bordeaux 2, 2006.

maximisation de la vraisemblance. De nouvelles approches biologiques permettent de tester, sur un échantillon sanguin, s'il s'agit d'une nouvelle infection (JANSSEN 1998). Certaines EDS, comme par exemple celle de 2003 au Burkina Faso, ont intégré des tests de ce genre à leur algorithme de dépistage. Ce type de résultats est encore peu exploité mais constitue une source potentielle d'estimation des incidences. L'Actuarial Society of South Africa a développé un modèle épidémiologique comportemental, ASSA 2003, pour décrire l'épidémie de l'Afrique du Sud (DORRINGTON 2006). L'adaptation de ce type de modèles à d'autres pays permettrait d'affiner les hypothèses applicables à une population donnée.

Le nombre relativement faible de travaux sur la modélisation des épidémies renvoie à la complexité d'une telle tâche. Un modèle tel que Spectrum est conçu pour réaliser des estimations à partir de données parcellaires. Cependant, le fonctionnement mathématique interne d'un tel modèle constitue une forme très simplifiée et parfois non vérifiée de la dynamique interne des épidémies. Il ne s'agit donc pas d'un modèle explicatif permettant de décrire le mécanisme de propagation du VIH dans le temps à travers les populations et les territoires.

Les sciences sociales sont ici confrontées aux difficultés liées à la complexité des phénomènes qu'elles étudient. Expliquer une épidémie dans ses différentes composantes nécessite de prendre en compte une multitude d'interactions complexes entre individus mobiles.

Le premier élément fondamental d'une telle modélisation consiste à estimer correctement les probabilités de transmission du VIH par acte. Si ces dernières sont fonction des pratiques des individus (sexuelles, consommation de drogues par voie injectable, etc.), elles dépendent très fortement de cofacteurs tels que la circoncision masculine, la coïnfection avec d'autres IST ou encore l'histoire individuelle de l'infection chez un individu donné.

Des travaux récents ont montré que les hommes circoncis auraient un risque moitié moindre d'être infectés que des hommes non circoncis (AUVERT 2005, BAILEY 2007, GRAY 2007). Une modélisation effectuée par des chercheurs de la London School of Hygiene and Tropical Medicine a montré l'importance de ce facteur sur le niveau actuel des épidémies (ORROTH 2006). Si la circoncision masculine avait été de 0 % au lieu de 100 % à Yaoundé, au Cameroun, la prévalence actuelle aurait été le triple de celle observée. À Ndola, en Zambie, la prévalence actuelle aurait été le tiers de celle observée si la circoncision masculine avait atteint 100 % au lieu de 10 %.

L'importance de la coïnfection avec le virus de l'Herpès (HSV-2) est également documentée depuis plusieurs années (GREENBLATT 1988, SIMONSEN 1988, PEREZ 1998, DEL MAR PUJADES RODRIGUEZ 2002). Les personnes infectées par HSV-2 seraient plus susceptibles d'être infectées par le VIH. Les résultats de l'essai ANRS 1285, au Burkina Faso, mettent en évidence qu'un traitement suppressif de HSV-2

chez des personnes coïnfectées par le VIH permet de réduire de 50 à 70% la charge virale de ce dernier (OUEDRAOGO 2006, NAGOT 2007).

Les taux de transmission par acte évoluent fortement au cours de l'infection et sont particulièrement élevés pendant la primo-infection (LEYNAERT 1998, WAWER 2005). L'infectiosité des personnes séropositives au VIH se maintiendrait ensuite à un niveau bas pendant la période asymptomatique, à l'exception des épisodes d'IST pendant lesquels la concentration virale remonterait. Ainsi, près de la moitié des nouvelles infections pourrait être le fait de personnes elles-mêmes nouvellement infectées.

Par ailleurs, les probabilités de transmission par acte sont dépendantes des différentes souches virales. Ces dernières peuvent évoluer dans le temps chez un même individu. Certaines variantes résistantes, apparaissant chez des personnes sous traitement antirétroviral, pourraient être moins contaminantes. Enfin, tous les individus ne présentent pas la même susceptibilité d'être infectés face à une exposition au virus, du fait d'une résistance à la fois en partie innée (génétique) et en partie acquise (immunitaire).

Il semble que les cofacteurs biologiques de la transmission affecte les probabilités de transmission au point que leur impact sur les niveaux nationaux des prévalences du VIH puisse être supérieur à celui des différences comportementales d'un pays à un autre.

Les probabilités de transmission par acte ne constitue qu'un des éléments explicatifs de la dynamique épidémique. Elles s'appliquent à des comportements individuels, en particulier sexuels, dont la mesure reste imparfaite. Par exemple, les travaux récents de DE WALQUE suggèrent que la sexualité extraconjugale des femmes est largement sous-estimée (DE WALQUE 2007).

Les comportements sexuels des individus ne sont pas uniformes. Ils évoluent, que ce soit les pratiques sexuelles, leur fréquence ou le multi-partenariat, au cours des trajectoires de vie des individus et en fonction des évolutions du contexte social (normes, cultures, politiques de santé publique, perceptions de l'épidémie…).

L'analyse des réseaux sexuels est déterminante pour comprendre comment l'épidémie se propage au sein de certaines catégories, d'une part, et entre ces catégories, d'autre part. Les différences d'âges entre partenaires sexuels et le fait qu'une même personne puisse avoir simultanément des partenaires plus jeunes et plus âgés induit, par exemple, une propagation du VIH entre les générations. Certaines populations, les travailleurs du sexe en particulier, ont pu constituer des foyers de propagation des épidémies à leur démarrage. Cependant, il est évident que cela ne plus être le cas lorsqu'une épidémie atteint des niveaux élevés en population générale.

Une étude menée au Malawi montre que les réseaux sexuels des adolescents et des jeunes adultes sont extrêmement interconnectés et décentralisés (KOHLER 2006). La majorité des individus enquêtés est ainsi liée par une chaîne de partenaires sexuels sans que cela ne soit le fait d'un petit nombre d'individus extrêmement actifs. Aucun groupe « passerelle » n'émerge en particulier, contrairement à ce que prédisent des modèles théoriques plus « centralisés » (noyau/périphérie par exemple).

Ces réseaux s'inscrivent dans l'espace et sont interdépendants des réseaux de transport et de mobilité. Si la mobilité des êtres humains est un élément essentiel pour permettre aux épidémies de se diffuser, comme tendent à le montrer nos résultats cartographiques par exemple, les liens entre mobilités et infections à VIH restent complexes. Nathalie LYDIÉ montre ainsi que les catégories de populations qui se déplacent sont remarquables par leur diversité. Si certains facteurs de prises de risques peuvent être associés aux diverses formes de mobilités, la situation reste plus complexe qu'elle n'y parait (LYDIÉ 2001). Autre exemple, les déplacements de population dans le cadre de conflits armés ne s'accompagnent pas systématiquement d'une augmentation des épidémies (SPIEGEL 2007) bien que cela ait pu être le cas pour certains conflits.

Les dynamiques épidémiques relèvent de phénomènes complexes dont les connaissances actuelles se contentent d'en esquisser les contours. Une mesure relativement fine du niveau et des tendances des prévalences constitue l'un des premiers indicateurs sur lequel fonder une connaissance. Mais, de la description des épidémies (modèles monoscopiques) à l'explication de leurs mécanismes (modèles panscopiques), il reste un pas immense à franchir qui requerra, très probablement, le développement de théories multidimensionnelles de la complexité applicables aux sciences sociales.

Références bibliographiques

Remarques sur la présente bibliographie :

Lorsque le nombre d'auteurs est supérieur à six, seuls les trois premiers auteurs sont listés, suivi de l'abréviation *et al.* Si nous avions connaissance de l'existence d'une version électronique accessible gratuitement sur le web, l'adresse internet a été mentionnée entre parenthèses à la suite du document concerné. Les noms de villes sont suivis d'un code à deux lettres spécifiant le pays concerné. Il s'agit du code international *ISO 3166-1-alpha-2*. Un tableau de correspondance avec les noms de pays peut être obtenu sur *http://www.iso.org/iso/fr/prods-services/iso3166ma/02iso-3166-code-lists/list-fr1.html* ou bien encore *http://fr.wikipedia.org/wiki/ISO_3166-1*.

Dans le corps du texte, les références sont mentionnées, entre parenthèses, par le nom du premier auteur et l'année de publication. Lorsque plusieurs documents ont le même premier auteur et la même année de publication, une lettre minuscule a été ajoutée à l'année de publication afin de les distinguer (par exemple : 1983a, 1983b). Lorsque la référence suit une citation d'un passage précis d'un ouvrage, la ou les pages dont est extraite la citation sont mentionnées sous la forme *p. 210* pour la page 210 ou *p. 185-190* pour les pages 185 à 190.

Concernant les ouvrages ayant été plusieurs fois réédités et/ou traduits, l'année d'édition utilisée dans le corps de la thèse correspond à l'année de la première édition originale. S'il s'agit d'une réédition corrigée et augmentée, l'année de la première réédition corrigée sera utilisée. Les pages indiquées suite à une citation correspondent néanmoins à la réédition consultée. Les détails concernant cette réédition sont mentionnés dans la bibliographie présentée ci-après.

Dans les notes de bas de page, la première occurrence d'une référence bibliographique est détaillée mais a été allégée de certaines informations (telles que les particularités de l'édition ou le nombre de pages) afin d'éviter d'allonger inutilement les notes de bas de page. À partir de la seconde occurrence d'une même référence, seuls sont mentionnés l'auteur et le titre court de l'ouvrage.

Les références propres aux annexes sont publiées à la fin de celles-ci.

Nombre de références : 347

AKWARA P. A., FOSU G. B., GOVINDASAMY P., ALAYÓN S. et HYSLOP A., *An In-Depth Analysis of HIV Prevalence in Ghana: Further Analysis of Demographic and Health Surveys Data*, Calverton, Maryland (US), ORC Macro, 2005, 44 pages. (http://www.measuredhs.com/pubs/pdf/FA46/FA46.pdf)

ALKEMA L., RAFTERY A. E. et CLARK S. J., *Probabilistic Projections of HIV Prevalence Using Bayesian Melding*, Washington, DC (US), Center for Statistics and the Social Sciences, University of Washington, 5 janvier, 2007, Working Paper n°69, 16 pages. (http://www.csss.washington.edu/Papers/wp69.pdf)

ALLEN T., « AIDS and evidence: interrogating [corrected] some Ugandan myths », *Journal of Biosocial Science*, n°38(1), 2006, pages 7-28.

ALVAREZ M., OYONARTE S., RODRIGUEZ P. M. et HERNANDEZ J. M., « Estimated risk of transfusion-transmitted viral infections in Spain », *Transfusion*, n°42(8), 2002, pages 994-998.

ARISTOTE, *Les Seconds Analytiques - Organon IV*, Wikisource, nouvelle traduction pour Internet par sœur Pascale NAU, à partir de la version grecque, de la traduction Vrin et de celle de G. R. G. Mure. (http://fr.wikisource.org/wiki/Seconds_Analytiques)

ASSEFA T., DAVEY G., DUKERS N. *et al.*, « Overall HIV-1 prevalence in pregnant women over-estimates HIV-1 in the predominantly rural population of Afar Region », *Ethiopian Medical Journal*, n°41 Suppl 1, 2003, pages 43-49.

AUBRY P. et PIÉGAY H., « Pratique de l'analyse de l'autocorrélation spatiale en géomorphologie : définitions opératoires et tests », *Géographie physique et Quaternaire*, n°55(2), 2001, pages 111-129. (http://www.erudit.org/revue/gpq/2001/v55/n2/008297ar.pdf)

AUVERT B., TALJAARD D., LAGARDE E. *et al.*, « Randomized, controlled intervention trial of male circumcision for reduction of HIV infection risk: the ANRS 1265 Trial », *PLoS Med*, n°2(11), 2005, pages e298.

AYAD M., BARRÈRE B. et OTTO J., *Demographic and Socioeconomic Characteristics of Households*, Calverton, Mayland (US), Macro International, coll. *DHS Comapartive Studies*, septembre, 1997, n°26, 75 pages.

BACHELARD G., *La Formation de l'esprit scientifique : contribution à une psychanalyse de la connaissance*, J. Vrin, réédition en format poche de 2004, collection Bibliothèque des textes philosophiques, Paris (FR), 1938, 306 pages.

BAILEY R. C., MOSES S., PARKER C. B. *et al.*, « Male circumcision for HIV prevention in young men in Kisumu, Kenya: a randomised controlled trial », *Lancet*, n°369(9562), 2007, pages 643-656.

BAILLARGEON S., *Le krigeage : revue de la théorie et application à l'interpolation spatiale de données de précipitations*, mémoire présenté pour l'obtention du grade de Maître ès Sciences (M.Sc.), sous la direction de RIVEST L.-P. et POULIOT J., Université de Laval, Faculté des Sciences et de Génie, Québec (CA), 2005, 137 pages. (http://www.theses.ulaval.ca/2005/22636/22636.pdf)

BAROUILLET T., MAURIN T. et OLLIER L., « Dépistage du VIH : rappel sur la technique », *Bulletin du Réseau Ville-Hôpital des Alpes maritimes*, n°25, 2005, pages 5. (http://www.revihop06.org/Archives/revihop25.pdf)

BARRÉ-SINOUSSI F., CHERMANN J. C., REY F. *et al.*, « Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS) », *Science*, n°220(4599), 1983, pages 868-871.

BARRETT J. E., DAWSON G., HELLER J. *et al.*, « Performance evaluation of the Abbott HTLV III EIA, a test for antibody to HTLV III in donor blood », *American Journal of Clinical Pathology*, n°86(2), 1986, pages 180-185.

BARRY A., KABA D. et DIOP I., *Rapport final ESSIDAGUI/2001*, Conakry (GN), Stat-View International, 2002, rapport non publié.

BECKER K. M., GLASS G. E., BRATHWAITE W. et ZENILMAN J. M., « Geographic epidemiology of gonorrhea in Baltimore, Maryland, using a geographic information system », *American Journal of Epidemiology*, n°147(7), 1998, pages 709-716.

BENOIT S. N., GERSHY-DAMET G. M., COULIBALY A. *et al.*, « Seroprevalence of HIV infection in the general population of the Côte d'Ivoire, West Africa », *Journal of Acquired Immune Deficiency Syndromes*, n°3(12), 1990, pages 1193-1196.

BERKLEY S., NAAMARA W., OKWARE S. *et al.*, « AIDS and HIV infection in Uganda: are more women infected than men? », *AIDS*, n°4(12), 1990, pages 1237-1242.

BERKLEY S. W., NAAMARA W. et OKWARE S., *The epidemiology of AIDS and HIV infection in women in Uganda*, IV International Conference on AIDS and Associatied Cancers in AFrica, Marseille (FR), Octobre, 1989.

BIGNAMI-VAN ASSCHE S., SALOMON J. A. et MURRAY C. J., *Evidence from National Population-Based Surveys on Bias in Antenatal Clinic-Based Estimates of HIV Prevalence*, Population Association of America Annual Meeting, Philadelphia, Pennsylvania (US), March 31 - April 2, 2005, 29 pages. (http://paa2005.princeton.edu/)

BIZIMUNGU C., NTILIVAMUNDA A. et TAHIMANA M., « Nationwide community-based serological survey of HIV-1 and other human retrovirus infections », *Lancet*, n°1(8644), 1989, pages 941-943.

BLANC A., *The Relationship between sexual behaviour and level of education in developing countries*, Genève (CH), UNAIDS, 2000.

BLANC A. K. et RUTENBERG N., « Assesment of the quality of data on age at first sexual intercourse, age at first marriage and age at first birth in the Demographic and Health Surveys », dans INSTITUT FOR RESOURCE DEVELOPMENT, *An Assessment of DHS-I Data Quality*, Institut for Resource Development, Macro Systems Inc., Columbia, Maryland 'US), 1990, pages 41-82.

BOERMA J. T., URASSA M., SENKORO K., KLOKKE A. et NGWESHEMI J. Z., « Spread of HIV infection in a rural area of Tanzania », *AIDS*, n°13(10), 1999, pages 1233-1240.

BOERMA J. T., GHYS P. D. et WALKER N., « Estimates of HIV-1 prevalence from national population-based surveys as a new gold standard », *Lancet*, n°362(9399), 2003, pages 1929-1931.

BOISSON E., NICOLL A., ZABA B. et RODRIGUES L. C., « Interpreting HIV seroprevalence data from pregnant women », *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, n°13(5), 1996, pages 434-439.

BONGAARTS J., *Modeling the Spread of HIV and the Demographic Impact of AIDS in Africa*, New York City, New York (US), The Population Council, coll. *Working Papers of the Center for Policy Studies*, 1988, n°140, 42 pages.

BONITZER J., « Réflexions sur les modèles statistiques de décision », *Revue de statistique appliquée*, n°32(1), 1984, pages 9-37. (http://www.numdam.org/numdam-bin/fitem?id=RSA_1984__32_1_9_0)

BOURDIEU P., CHAMBOREDON J.-C. et PASSERON J.-C., *Le Métier de sociologue : préalables épistémologiques*, Mouton, 2e édition révisée, première édition en 1968, Paris-La Haye (FR), 1973, 357 pages.

BOUZITAT C., BOUZITAT P. et PAGÈS G., *Statistique, Probabilités, Estimation ponctuelle : cours et exercice d'application*, Cujas, Paris (FR), 1990, 224 pages.

BROCKLEHURST P. et FRENCH R., « The association between maternal HIV infection and perinatal outcome: a systematic review of the literature and meta-analysis », *British Journal of Obstetrics and Gynaecology*, n°105(8), 1998, pages 836-848.

BROWN T. et PEERAPATANAPOKIN W., « The Asian Epidemic Model: a process model for exploring HIV policy and programme alternatives in Asia », *Sexually Transmitted Infections*, n°80 Suppl 1, 2004, pages I19-I24.

BROWN T., GRASSLY N. C., GARNETT G. et STANECKI K., « Improving projections at the country level: the UNAIDS Estimation and Projection Package 2005 », *Sexually Transmitted Infections*, n°82(suppl_3), 2006, pages iii34-40.

BUSCH M. P., LEE L. L., SATTEN G. A. *et al.*, « Time course of detection of viral and serologic markers preceding human immunodeficiency virus type 1 seroconversion: implications for screening of blood and tissue donors », *Transfusion*, n°35(2), 1995, pages 91-97.

BUSCH M. P., KLEINMAN S. H., JACKSON B. *et al.*, « Committee report. Nucleic acid amplification testing of blood donors for transfusion-transmitted infectious diseases: Report of the Interorganizational Task Force on Nucleic Acid Amplification Testing of Blood Donors », *Transfusion*, n°40(2), 2000, pages 143-159.

BUSCH M. P., GLYNN S. A., STRAMER S. L. *et al.*, « A new strategy for estimating risks of transfusion-transmitted viral infections based on rates of detection of recently infected donors », *Transfusion*, n°45(2), 2005, pages 254-264.

BUVE A., CARAEL M., HAYES R. J. *et al.*, « The multicentre study on factors determining the differential spread of HIV in four African cities: summary and conclusions », *AIDS*, n°15 Suppl 4, 2001a, pages S127-131.

BUVE A., LAGARDE E., CARAEL M. *et al.*, « Interpreting sexual behaviour data: validity issues in the multicentre study on factors determining the differential spread of HIV in four African cities », *AIDS*, n°15 Suppl 4, 2001b, pages S117-126.

CACERES C., KONDA K., PECHENY M., CHATTERJEE A. et LYERLA R., « Estimating the number of men who have sex with men in low and middle income countries », *Sexually Transmitted Infections*, n°82 Suppl 3, 2006, pages iii3-9.

CARAEL M., *La Surveillance de l'infection à VIH en Afrique*, Chaire Quételet, Louvain-la-Neuve (BE), 19 novembre, Université Catholique de Louvain, 2004. (http://www.demo.ucl.ac.be/cq04/powerpoint/Carael.pdf)

CARBALLO M. et SOLBY S., *HIV/AIDS, conflict and reconstruction in sub-saharan Africa*, Preventing and Coping with HIV/AIDS in Post-Conflict Societies: gender-based lessons from sub-saharan Africa, Durban (ZA), 26-28 mars, African Centre for the Constructive Resolution of Disputes (ACCORD), Tulane University Payson Center for International Development and Technology Transfer, 2001. (http://www.certi.org/publications/AIDS_symp/pub/carb.PDF)

CARPENTER L. M., NAKIYINGI J. S., RUBERANTWARI A. *et al.*, « Estimates of the impact of HIV-1 infection on fertility in a rural Ugandan population cohort », *Health Transition Review*, n°7(Supplement 2), 1997, pages 113-126.

CARPENTER L. M., KAMALI A., RUBERANTWARI A., MALAMBA S. S. et WHITWORTH J. A., « Rates of HIV-1 transmission within marriage in rural Uganda in relation to the HIV sero-status of the partners », *AIDS*, n°13(9), 1999, pages 1083-1089.

CARPENTER L. M., KAMALI A., PAYNE M. *et al.*, « Independent effects of reported sexually transmitted infections and sexual behavior on HIV-1 prevalence among adult women, men, and teenagers in rural Uganda », *Journal of Acquired Immune Deficiency Syndromes*, n°29(2), 2002, pages 174-180.

CARRAT F. et VALLERON A. J., « Epidemiologic mapping using the "kriging" method: application to an influenza-like illness epidemic in France », *American Journal of Epidemiology*, n°135(11), 1992, pages 1293-1300.

CDC, « Kaposi's sarcoma and Pneumocystis pneumonia among homosexuel men - New York City and California », *Morbidity and Mortality Weekly Report*, n°30, 1981a, pages 305-308.

CDC, « Pneumocystis pneumonia - Los Angeles », *Morbidity and Mortality Weekly Report*, n°30, 1981b, pages 250-252.

CDC, « Update on Acquired Immune Deficiency Syndrome (AIDS) - United States », *Morbidity and Mortality Weekly Report*, n°31(37), 1982, pages 507-508. (http://www.cdc.gov/mmwr/preview/mmwrhtml/00001163.htm)

CDC, « Update: Acquired Immunodeficiency Syndrome (AIDS) - United States », *Morbidity and Mortality Weekly Report*, n°32(52), 1984, pages 688-691. (http://www.cdc.gov/mmwr/preview/mmwrhtml/00000254.htm)

CDC, « Revision of the Case Definition of Acquired Immunodeficiency Syndrome for National Reporting - United States », *Morbidity and Mortality Weekly Report*, n°34(25), 1985, pages 373-375. (http://www.cdc.gov/mmwr/preview/mmwrhtml/00000567.htm)

CELLULE DE PLANIFICATION ET DE STATISTIQUE, MINISTÈRE DE LA SANTÉ, DIRECTION NATIONALE DE LA STATISTIQUE ET DE L'INFORMATIQUE (DNSI) et ORC MACRO, *HIV testing in Mali: findings from the 2001 Mali Demographic and Health Survey*, Calverton, Maryland (US), Cellule de Planification et de Statistique, ministère de la Santé, Direction Nationale de la Statistique et de l'Informatique (DNSI), ORC Macro, 2002.

CENTER FOR INTERNATIONAL EARTH SCIENCE INFORMATION NETWORK (CIESIN) OF COLUMBIA UNIVERSITY, INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE (IFPRI), THE WORLD BANK et CENTRO INTERNACIONAL DE AGRICULTURA TROPICAL (CIAT), *Global Rural-Urban Mapping Project (GRUMP), Alpha Version: Settlement Points*, Palisades, New York (US), Socioeconomic Data and Applications Center (SEDAC) of Columbia University, 2004a. (http://sedac.ciesin.columbia.edu/gpw)

CENTER FOR INTERNATIONAL EARTH SCIENCE INFORMATION NETWORK (CIESIN) OF COLUMBIA UNIVERSITY, INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE (IFPRI), THE WORLD BANK et CENTRO INTERNACIONAL DE AGRICULTURA TROPICAL (CIAT), *Global Rural-Urban Mapping Project (GRUMP), Alpha Version: Urban Extents*, Palisades, New York (US), Socioeconomic Data and Applications Center (SEDAC) of Columbia University, 2004b. (http://sedac.ciesin.columbia.edu/gpw)

CENTER FOR INTERNATIONAL EARTH SCIENCE INFORMATION NETWORK (CIESIN) OF COLUMBIA UNIVERSITY, INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE (IFPRI), THE WORLD BANK et CENTRO INTERNACIONAL DE AGRICULTURA TROPICAL (CIAT), *Global Rural-Urban Mapping Project (GRUMP), Alpha Version: Population Density Grids*, Palisades, New York (US), Socioeconomic Data and Applications Center (SEDAC) of Columbia University, 2004c. (http://sedac.ciesin.columbia.edu/gpw)

CENTRAL BUREAU OF STATISTICS, *Kenya 1999 Population and Housing Census - Volume I Population distribution by administrative areas and urban centers*, Nairobi (KE), Ministry of planning and national development, Central bureau of statistics, janvier, 2001, 443 pages.

CENTRAL BUREAU OF STATISTICS, *Kenya 1999 Population and Housing Census - Volume VII Analytical Report on Population Projections*, Nairobi (KE), Ministry of planning and national development, Central bureau of statistics, août, 2002a, 155 pages.

CENTRAL BUREAU OF STATISTICS, *Kenya 1999 Population and Housing Census - Volume X Analytical Report on Housing Conditions and Household amenities*, Nairobi (KE), Ministry of planning and national development, Central bureau of statistics, août, 2002b, 109 pages.

CENTRAL BUREAU OF STATISTICS, MINISTRY OF HEALTH et ORC MACRO, *Kenya Demographic and Health Survey 2003*, Calverton, Maryland (US), CBS, MOH, ORC Macro, Juillet, 2004, 389 pages. (http://www.measuredhs.com/pubs/pdf/FR151/FR151-KE03.pdf)

CENTRAL STATISTICAL OFFICE, CENTRAL BOARD OF HEALTH et ORC MACRO, *Zambia Demographic and Health Survey 2001–2002: preliminary report*, Calverton, Maryland (US), Central Statistical Office, Central Board of Health, ORC Macro, 2002.

CHALMERS A. F., *Qu'est-ce que la science ? Récents développements en philosophie des sciences : Popper, Kuhn, Lakatos, Feyerabend*, La Découverte, réédition de 2006 en format poche dans la collection Livre de poche, 1987 pour la première édition française, 1976 pour la première édition anglaise, Paris (FR), 1976, 287 pages.

CHANGALUCHA J., GROSSKURTH H., MWITA W. *et al.*, « Comparison of HIV prevalences in community-based and antenatal clinic surveys in rural Mwanza, Tanzania », *AIDS*, n°16(4), 2002, pages 661-665.

CHIN J. et MANN J. M., « The global patterns and prevalence of AIDS and HIV infection », *AIDS*, n°2(suppl. 1), 1988, pages S247-S252.

CHIN J. et MANN J. M., « Global surveillance and forecasting of AIDS », *Bulletin of the World Health Organization*, n°67(1), 1989, pages 1-7. (http://libdoc.who.int/bulletin/1989/Vol67-No1/bulletin_1989_67(1)_1-7.pdf)

CHIN J., « Public health surveillance of AIDS and HIV infections », *Bulletin of the World Health Organization*, n°68(5), 1990, pages 529-536. (http://whqlibdoc.who.int/bulletin/1990/Vol68-No5/bulletin_1990_68(5)_529-536.pdf)

CHIN J. et LWANGA S. K., « Estimation and projection of adult AIDS cases: a simple epidemiological model », *Bulletin of the World Health Organization*, n°69(4), 1991a, pages 399-406. (http://whqlibdoc.who.int/bulletin/1991/Vol69-No4/bulletin_1991_69(4)_399-406.pdf)

CHIN J. et LWANGA S. K., « The World Health Organization approach: projections of non-paediatric HIV infection and AIDS in pattern II areas », dans UN/WHO, *The AIDS epidemic and its demographic consequences*, UN/WHO, New York City, New York (US), 1991b, pages 137-140.

CHORLEY R. J. et HAGGETT P., « Trend-Surface Mapping in Geographical Research », *Transactions of the Institute of British Geographers*(37), 1965, pages 47-67.

CITY POPULATION, *Burkina Faso - City Population - Cities, Towns & Provinces - Statistics and Maps*, 2006, page web consultée le September, 25th 2006. (http://www.citypopulation.de/BurkinaFaso.html)

CNLS-IST, *Sida Retro-Infos : Bulletin de rétro-information sur le SIDA au Burkina Faso*, Ougadougou (BF), CNLS-ISt, juin, 2004, n°7, 9 pages.

CNSL BURKINA FASO, *Récapitulatif des taux de prévalence du VIH chez les femmes enceintes dans les sites sentinnelles de 1997 à 2003*, CNLS, 2003.

COTE J. A. et BUCKLEY M. R., « Estimating Trait, Method, and Error Variance: Generalizing across 70 Construct Validation Studies », *Journal of Marketing Research*, n°24(3), 1987, pages 315-318.

COURGEAU D., « Probabilités, démographie et sciences sociales », *Mathématiques & sciences humaines*, n°167, 2004, pages 27-50. (http://www.ehess.fr/revue-msh/pdf/N167R933.pdf)

CRAIG M. H., SHARP B. L., MABASO M. L. et KLEINSCHMIDT I., « Developing a spatial-statistical model and map of historical malaria prevalence in Botswana using a staged variable selection procedure », *Int J Health Geogr*, n°6(1), 2007, pages 44.

CRAMPIN A. C., GLYNN J. R., NGWIRA B. M. *et al.*, « Trends and measurement of HIV prevalence in northern Malawi », *AIDS*, n°17(12), 2003, pages 1817-1825.

CRITON C. et FENER P., *Dépistage du VIH/SIDA chez la femme à risque*, Vandoeuvre-lès-Nancy (FR), INIST - CNRS, 2007, 36 pages. (http://www.inist.fr/article188.html)

CURTIS S. L. et SUTHERLAND E. G., « Measuring sexual behaviour in the era of HIV/AIDS: the experience of Demographic and Health Surveys and similar enquiries », *Sexually Transmitted Infections*, n°80 Suppl 2, 2004, pages ii22-27.

DANDONA L., LAKSHMI V., SUDHA T., KUMAR G. A. et DANDONA R., « A population-based study of human immunodeficiency virus in south India reveals major differences from sentinel surveillance-based estimates », *BMC Med*, n°4, 2006, pages 31.

DARE O. O. et CLELAND J. G., « Reliability and validity of survey data on sexual behaviour », *Health Transition Review*, n°4 Suppl, 1994, pages 93-110.

DE COCK K. M., COLEBUNDERS R., FRANCIS H. *et al.*, « Evaluation of the WHO clinical case definition for AIDS in rural Zaire », *AIDS*, n°2(3), 1988, pages 219-221.

DE WALQUE D., *Who Gets AIDS and How? The determinants of HIV infection and sexual behaviors in Burkina Faso, Cameroon, Ghana, Kenya and Tanzania*, Washington, DC (US), World Bank, coll. *Policy Research Working Papers*, 2006, WPS3844, 51 pages. (http://www-wds.worldbank.org/external/default/WDSContentServer/IW3P/IB/2006/02/03/000016406_20060203104911/Rendered/PDF/wps3844.pdf)

DE WALQUE D., « Sero-Discordant Couples in Five African Countries: Implications for Prevention Strategies », *Population and Development Review*, n°33(3), 2007, pages 501-523.

DECOSAS J., « Special report: West Africa. Migration factor makes regional approach essential », *AIDS Anal Afr*, n°5(3), 1995, pages 8-9.

DEL MAR PUJADES RODRIGUEZ M., OBASI A., MOSHA F. *et al.*, « Herpes simplex virus type 2 infection increases HIV incidence: a prospective study in rural Tanzania », *AIDS*, n°16(3), 2002, pages 451–462.

DESGRÉES DU LOÛ A., MSELLATI P., YAO A. *et al.*, « Impaired fertility in HIV-1-infected pregnant women: a clinic-based survey in Abidjan, Cote d'Ivoire, 1997 », *AIDS*, n°13(4), 1999, pages 517-521.

DGS, « Le Point sur le SIDA », *Bulletin Épidémiologique Hebdomadaire*, n°51/1983, 1983, pages 2-3. (http://www.invs.sante.fr/beh/1983/51/beh_51_1983.pdf)

DGS, « Définition du SIDA avéré (révision 1987) », *Bulletin Épidémiologique Hebdomadaire*, n°51/1987, 1987, pages 201-203. (http://www.invs.sante.fr/beh/1987/51/beh_51_1987.pdf)

DGS, « Révision de la définition du SIDA en France », *Relevé Épidémiologique Hebdomadaire*, n°11/1993, 1993, pages 47-48. (http://www.invs.sante.fr/beh/1993/11/beh_11_1993.pdf)

DI SALVO M., GADAIS M. et ROCHE-WOILLEZ M., *L'estimation de la densité par la méthode du noyau : méthodes et outils*, Lyon (FR), CERTU, coll. *Rapports d'études du CERTU*, 2005, 26 pages. (http://www1.certu.fr/catalpres/0949T1.zip)

DIAZ T., DE COCK K., BROWN T., GHYS P. D. et BOERMA J. T., « New strategies for HIV surveillance in resource-constrained settings: an overview », *AIDS*, n°19 Suppl 2, 2005, pages S1-S8.

DIRECTION NATIONALE DU DEUXIÈME RECENSEMENT DE LA POPULATION ET DE L'HABITAT, *Deuxième Recensement Général de la Population et de l'Habitat du Cameroun 1987 - Volume I Résultats bruts - Tome 1 République du Cameroun*, Yaoundé (CM), 1992, 838 pages.

DORRINGTON R. E., BRADSHAW D., JOHNSON L. et DANIEL T., *The Demographic Impact of HIV/AIDS in South Africa: National and Provincial Indicators for 2006*, Cape Town (ZA), Centre for Actuarial Research, the Burden of Disease Research Unit (Medical Research Council) and the Actuarial Society of South Africa, 2006, 116 pages. (http://www.assa.org.za/applications/cms/documents/file_build.asp?id=100000148)

DOWSETT G., *The Problematic Category of MSM: Masculinity, sexuallity and HIV/AIDS*, XVIe Conférence Internationale sur le SIDA, Toronto (CA), 13-18 août, 2006, THSA09.

DURAND J. P., MUSI S., JOSSE R. *et al.*, « Prévalence des porteurs d'anticorps contre les virus de l'immunodéficience humaine (VIH1 et VIH2) dans le Sud-Cameroun. Résultats des tentatives d'isolement de rétrovirus. », *Medecine Tropicale*, n°48(4), 1988, pages 391-395.

DURKHEIM É., *Les règles de la méthode sociologique*, Félix Alcan, première édition reproduite en ligne sur Gallica (BNF), Paris (FR), 1895, 186 pages. (http://gallica.bnf.fr/ark:/12148/bpt6k1055050)

DZEKEDZEKE K. et FYLKESNES K., « Reducing uncertainties in global HIV prevalence estimates: the case of Zambia », *BMC Public Health*, n°6, 2006, pages 83.

EINSTEIN A., *La Théorie de la Relativité restreinte et générale*, Dunod, réédition de 1999, traduction de Maurice Solovine à partir de la 14e édition allemande, Paris (FR), 1917, 192 pages.

ELLIOTT L. J., BLANCHARD J. F., BEAUDOIN C. M. *et al.*, « Geographical variations in the epidemiology of bacterial sexually transmitted infections in Manitoba, Canada », *Sexually Transmitted Infections*, n°78 Suppl 1, 2002, pages i139-144.

FABIANI M., FYLKESNES K., NATTABI B., AYELLA E. O. et DECLICH S., « Evaluating two adjustment methods to extrapolate HIV prevalence from pregnant women to the general female population in sub-Saharan Africa », *AIDS*, n°17(3), 2003, pages 399-405.

FABIANI M., NATTABI B., AYELLA E. O., OGWANG M. et DECLICH S., « Differences in fertility by HIV serostatus and adjusted HIV prevalence data from an antenatal clinic in northern Uganda », *Tropical Medicine and International Health*, n°11(2), 2006, pages 182-187.

FANG C. T., FIELD S. P., BUSCH M. P. et HEYNS ADU P., « Human immunodeficiency virus-1 and hepatitis C virus RNA among South African blood donors: estimation of residual transfusion risk and yield of nucleic acid testing », *Vox Sanguinis*, n°85(1), 2003, pages 9-19.

FERRY B., DEHENEFFE J.-C., MAMDANI M. et INGHAM R., « Characteristics of Surveys and Data Quality », dans CLELAND J. et FERRY B., *Sexual Behaviour and AIDS in the Developing World*, WHO, Taylor & Francis, coll. *Social Aspects of AIDS*, Londres (UK), 1995, pages 10-42.

FONTANET A. L., MESSELE T., DEJENE A. *et al.*, « Age- and sex-specific HIV-1 prevalence in the urban community setting of Addis Ababa, Ethiopia », *AIDS*, n°12(3), 1998, pages 315-322.

FREEDBERG K. et YAZDANPANAH Y., « Cost-effectiveness of HIV Therapies in Resource-Poor Countries », dans MOATTI J.-P., CORIAT B., SOUTEYRAND Y., BARNETT T., DUMOULIN J. et FLORI Y.-A., *Economics of AIDS and Access to HIV/AIDS Care in Developing Countries: Issues and Challenges*, ANRS, coll. *Sciences Sociales et Sida*, Paris (FR), 2003, pages 267-291.

FYLKESNES K., NDHLOVU Z., KASUMBA K., MUBANGA MUSONDA R. et SICHONE M., « Studying dynamics of the HIV epidemic: population-based data compared with sentinel surveillance in Zambia », *AIDS*, n°12(10), 1998, pages 1227-1234.

FYLKESNES K., MUSONDA R. M., SICHONE M. *et al.*, « Declining HIV prevalence and risk behaviours in Zambia: evidence from surveillance and population-based surveys », *AIDS*, n°15(7), 2001, pages 907-916.

GALVANI L., « Révision critique de certains points de la méthode représentative », *Revue de l'Institut International de Statistique*, n°19(1), 1951, pages 1-12.

GARCIA-CALLEJA J. M., ZANIEWSKI E., GHYS P. D., STANECKI K. et WALKER N., « A global analysis of trends in the quality of HIV sero-surveillance », *Sexually Transmitted Infections*, n°80 Suppl 1, 2004, pages I25-I30.

GARCIA-CALLEJA J. M., GOUWS E. et GHYS P. D., « National population based HIV prevalence surveys in sub-Saharan Africa: results and implications for HIV and AIDS estimates », *Sexually Transmitted Infections*, n°82(suppl_3), 2006, pages iii64-70.

GEARY R. C., « The Contiguity Ratio and Statistical Mapping », *The Incorporated Statistician*, n°5(3), 1954, pages 115-145.

GEMPERLI A., VOUNATSOU P., SOGOBA N. et SMITH T., « Malaria mapping using transmission models: application to survey data from Mali », *American Journal of Epidemiology*, n°163(3), 2006, pages 289-297.

GENDREAU F., GUBRY F., LOHLE-TART L., VAN DE WALLE É. et WALTISPERGER D., *Manuel de Yaoundé : Estimations indirectes en démographie africaine*, UIESP, IFORD, Groupe de Démographique Africaine, Ordina éditions, Liège (BE), 1985, 247 pages.

GERSHY-DAMET G. M., KOFFI K., SORO B. *et al.*, « Seroepidemiological survey of HIV-1 and HIV-2 infections in the five regions of Ivory Coast », *AIDS*, n°5(4), 1991, pages 462-463.

GERSOVITZ M., « The HIV Epidemic in Four African Countries Seen through the Demographic and Health Surveys », *Journal of African Economies*, n°14(2), 2005, pages 191-246.

GHYS P. D., BROWN T., GRASSLY N. C. *et al.*, « The UNAIDS Estimation and Projection Package: a software package to estimate and project national HIV epidemics », *Sexually Transmitted Infections*, n°80 Suppl 1, 2004, pages I5-I9.

GLYNN J. R., BUVE A., CARAEL M. *et al.*, « Factors influencing the difference in HIV prevalence between antenatal clinic and general population in sub-Saharan Africa », *AIDS*, n°15(13), 2001a, pages 1717-1725.

GLYNN J. R., CARAEL M., AUVERT B. *et al.*, « Why do young women have a much higher prevalence of HIV than young men? A study in Kisumu, Kenya and Ndola, Zambia », *AIDS*, n°15 Suppl 4, 2001b, pages S51-60.

GLYNN J. R., SONNENBERG P., NELSON G. *et al.*, « Survival from HIV-1 seroconversion in Southern Africa: a retrospective cohort study in nearly 2000 gold-miners over 10 years of follow-up », *AIDS*, n°21(5), 2007, pages 625-632.

GLYNN S. A., KLEINMAN S. H., WRIGHT D. J. et BUSCH M. P., « International application of the incidence rate/window period model », *Transfusion*, n°42(8), 2002, pages 966-972.

GODIFROID B., AUGUSTIN N. et DIDACE N., « Étude sur la séropositivité liée à l'infection au Virus de l'Immunodéficience Humaine au Rwanda », *Revue Médicale Rwandaise*, n°20(54), 1988, pages 37-42.

GOTANÈGRE J. F., « Analyse géographique de l'incidence du VIH et du SIDA au Rwanda en 1990 », *Les cahiers d'Outre-Mer*, n°183, 1993, pages 233-252.

GOUWS E., WHITE P. J., STOVER J. et BROWN T., « Short term estimates of adult HIV incidence by mode of transmission: Kenya and Thailand as examples », *Sexually Transmitted Infections*, n°82 Suppl 3, 2006, pages iii51-55.

GPA, *Unlinked anonymous screening for the public health surveillance of HIV infection*, Genève (CH), WHO, Juin, 1989, GPA/SFI/89.3, 8 pages. (http://whqlibdoc.who.int/hq/1989/GPA_SFI_89.3.pdf)

GRASSLY N. C., MORGAN M., WALKER N. *et al.*, « Uncertainty in estimates of HIV/AIDS: the estimation and application of plausibility bounds », *Sexually Transmitted Infections*, n°80 Suppl 1, 2004, pages I31-I38.

GRATTON Y., « Le Krigeage : la méthode optimale d'interpolation spatiale », *les articles de l'Institut d'Analyse Géographique*, 2002. (http://www.iag.asso.fr/articles/krigeage_juillet2002.htm)

GRAY R. H., WAWER M. J., SERWADDA D. *et al.*, « Population-based study of fertility in women with HIV-1 infection in Uganda », *Lancet*, n°351(9096), 1998, pages 98-103.

GRAY R. H., KIGOZI G., SERWADDA D. *et al.*, « Male circumcision for HIV prevention in men in Rakai, Uganda: a randomised trial », *Lancet*, n°369(9562), 2007, pages 657-666.

GREENBLATT R. M., LUKEHART S. A., PLUMMER F. A. *et al.*, « Genital ulceration as a risk factor for human immunodeficiency virus infection », *AIDS*, n°2(1), 1988, pages 47-50.

GREGSON S., ZHUWAU T., ANDERSON R. M., CHIMBADZWA T. et CHIWANDIWA S. K., « Age and religion selection biases in HIV-1 prevalence data from antenatal clinics in Manicaland, Zimbabwe », *Central African Journal of Medicine*, n°41(11), 1995, pages 339-346.

GREGSON S., ZHUWAU T., ANDERSON R. M. et CHANDIWANA S. K., « Is there evidence for behaviour change in response to AIDS in rural Zimbabwe? », *Social Science and Medicine*, n°46(3), 1998, pages 321-330.

GREGSON S. et GARNETT G. P., « Contrasting gender differentials in HIV-1 prevalence and associated mortality increase in eastern and southern Africa: artefact of data or natural course of epidemics? », *AIDS*, n°14 Suppl 3, 2000, pages S85-99.

GREGSON S., WADDELL H. et CHANDIWANA S. K., « School education and HIV control in sub-saharan Africa : from discord to harmony ? », *Journal of international development*, n°13, 2001, pages 467-485.

GREGSON S., NYAMUKAPA C. A., GARNETT G. P. *et al.*, « Sexual mixing patterns and sex-differentials in teenage exposure to HIV infection in rural Zimbabwe », *Lancet*, n°359(9321), 2002a, pages 1896-1903.

GREGSON S., TERCEIRA N., KAKOWA M. *et al.*, « Study of bias in antenatal clinic HIV-1 surveillance data in a high contraceptive prevalence population in sub-Saharan Africa », *AIDS*, n°16(4), 2002b, pages 643-652.

GRIFFIN W. R., « Residual Gravity in Theory and Practice », *Geophysics*, n°14(1), 1949, pages 39-56.

GROUPE DE RÉFÉRENCE DE L'ONUSIDA SUR LES ESTIMATIONS MODÈLES ET PROJECTIONS, *Estimation et projection des Épidémies nationales de VIH/SIDA - Modèles et méthodologie permettant à l'ONUSIDA/OMS de procéder à l'estimation et à la projection des épidémies nationales de VIH/SIDA*, Genève (CH), ONUSIDA/OMS, 2003, ONUSIDA/01.83, 65 pages.

GROUPE DE SURVEILLANCE SÉRO-ÉPIDÉMIOLOGIQUE, *Bulletin séro-épidémiologique n°7 de surveillance du SIDA*, Dakar (SN), CNLS du Sénégal, juin, 1999, 36, 53 pages.

GROUPE ÉPIDÉMIOLOGIE PNLS, *Bulletin séro-épidémiologique n°11 de surveillance du SIDA*, Dakar (SN), CNLS du Sénégal, Laboratoire de Bactériologie et de Virologie CHUA Le Dantec,, septembre, 2004, 53 pages.

GUBRY P., NEGADI G. et TAYO J., « La Population du Cameroun au recensement de 1976 », *Revue science et technique. Série Sciences humaines / Science and technology review. Social sciences series*(1-2), 1983, pages 7-38.

GURTLER L., MUHLBACHER A., MICHL U. *et al.*, « Reduction of the diagnostic window with a new combined p24 antigen and human immunodeficiency virus antibody screening assay », *Journal of Virological Methods*, n°75(1), 1998, pages 27-38.

HAGGETT P., *L'Analyse spatiale en géographie humaine*, A. Colin, coll. *Collection U*, traduction de 1973 à partir de la quatrième édition britannique de 1968, Paris (FR), 1968, 390 pages.

HANKINS C. A., FRIEDMAN S. R., ZAFAR T. et STRATHDEE S. A., « Transmission and prevention of HIV and sexually transmitted infections in war settings: implications for current and future armed conflicts », *AIDS*, n°16(17), 2002, pages 2245-2252.

HARROWER M. A. et BREWER C. A., « ColorBrewer.org: An Online Tool for Selecting Color Schemes for Maps », *The Cartographic Journal*, n°40(1), 2003, pages 27-37.

HIRA S. K., NKOWANE B. M., KAMANGA J. *et al.*, « Epidemiology of human immunodeficiency virus in families in Lusaka, Zambia », *Journal of Acquired Immune Deficiency Syndromes*, n°3(1), 1990, pages 83-86.

HIV-SIDA.COM, *L'Épidémie de 1980 à 2000*, 2001, page web consultée le 30 mai 2007. (http://www.hiv-sida.com/historique2.shtml)

HOGARTH J., *Vocabulaire de la santé publique*, OMS Bureau régional de l'Europe, coll. *La Santé publique en Europe*, Copenhague (DK), 1977. (http://whqlibdoc.who.int/euro/phie/WHO_PHIE_4_fre.pdf)

HULL H. F., « Comparison of HIV-antibody prevalence in patients consenting to and declining HIV-antibody testing in an STD clinic », *Journal of the American Medical Association*, n°260(7), 1988, pages 935-938.

INSTITUT FOR RESOURCE DEVELOPMENT, *An Assesment of DHS-I Data Quality*, Columbia, Maryland (US), Institute for Resource Development, Macro Systems Inc., coll. *DHS Methodlogical report*, décembre, 1990, n°1, 143 pages. (http://www.measuredhs.com/pubs/pdf/MR1/MR1.pdf)

INSTITUT NATIONAL DE LA STATISTIQUE et ORC MACRO, *Enquête Démographique et de Santé 2004 du Cameroun*, Calverton, Maryland (US), INS, ORC Macro, juin, 2005, 479 pages. (http://www.measuredhs.com/pubs/pdf/FR163/FR163-CM04.pdf)

INSTITUT NATIONAL DE LA STATISTIQUE, *Annuaire Statistique du Cameroun 2006*, Yaoundé (CM), INS, 2006, 593 pages. (http://www.statistics-cameroon.org/ins/annuaire.htm)

INSTITUT NATIONAL DE LA STATISTIQUE ET DE LA DÉMOGRAPHIE, *Analyse des résultats du Recensement Général de la Population et de l'Habitation de 1996 du Burkina Faso - Volume I*, Ouagadougou (BF), INSD, 2000a, 374 pages. (http://www.insd.bf/Publications/Enq_Recens/RGP96/RGP96_Vol1_document_entier.pdf)

INSTITUT NATIONAL DE LA STATISTIQUE ET DE LA DÉMOGRAPHIE, *Annuaire Statistique du Burkina faso 1999*, Ouagadougou (BF), Ministère de l'Économie et des Finances, December, 2000b, 219 pages.

INSTITUT NATIONAL DE LA STATISTIQUE ET DE LA DÉMOGRAPHIE, *Projections de population du Burkina Faso*, Ouagadougou (BF), INSD, février, 2004, 85 pages. (http://www.insd.bf/Publications/Autres_publi/f_Projections_de_population.pdf)

INSTITUT NATIONAL DE LA STATISTIQUE ET DE LA DÉMOGRAPHIE, *Statistiques Structurelles*, Ougadougou (BF), INSD, 2006. (http://www.insd.bf/donnees/donnees_structurelles/Accueil_Données_structurelles_2.htm)

JACKSON D. J., NGUGI E. N., PLUMMER F. A. *et al.*, « Stable antenatal HIV-1 seroprevalence with high population mobility and marked seroprevalence variation among sentinel sites within Nairobi, Kenya », *AIDS*, n°13(5), 1999, pages 583-589.

JAFFE H. W., BREGMAN D. J. et SELIK R. M., « Acquired immune deficiency syndrome in the United States: the first 1,000 cases », *Journal of Infectious Diseases*, n°148(2), 1983, pages 339-345.

JANSSEN R. S., SATTEN G. A., STRAMER S. L. *et al.*, « New testing strategy to detect early HIV-1 infection for use in incidence estimates and for clinical and prevention purposes », *JAMA*, n°280(1), 1998, pages 42-48.

JENUM P., « Anti-HIV screening of pregnant women in south-east Norway », *NIPH Annals*, n°11, 1988, pages 53-58.

KARON M. J., *Methods for estimating HIV prevalence in the United States*, Atlanta, Georgia (US), Centers for Disease Control and Prevention, coll. *Technical Report n°D821*, 1997.

KEITA M. L. et TOURE H., *Estimation et analyse de la variation spatiale du risque de mortalité maternelle en Guinée*, document de travail non publié, 2006, 33 pages.

KENGEYA-KAYONDO J. F., AMAANA A. et NAAMARA W., *Anti-HIV seroprevalence in adult rural populations of Uganda and its implications for preventive strategies*, V International Conference on AIDS, Montreal (CA), Juin, 1989.

KIAËR A. N., « Observations et expériences concernant des dénombrements représentatifs », *Bulletin de l'Institut International de Statistique*, n°IX(2), 1895, pages 176.

KILIAN A. H., GREGSON S., NDYANABANGI B. *et al.*, « Reductions in risk behaviour provide the most consistent explanation for declining HIV-1 prevalence in Uganda », *AIDS*, n°13(3), 1999, pages 391-398.

KILLEWO J., NYAMUREKUNGE A. et SANDSTROM A., « Prevalence of HIV-1 infection in the Kagera region of Tanzania: A population-based study », *AIDS*, n°4, 1990, pages 1081-1085.

KLEINSCHMIDT I., BAGAYOKO M., CLARKE G. P., CRAIG M. et LE SUEUR D., « A spatial statistical approach to malaria mapping », *International Journal of Epidemiology*, n°29(2), 2000, pages 355-361.

KLEINSCHMIDT I., SHARP B. L., CLARKE G. P., CURTIS B. et FRASER C., « Use of generalized linear mixed models in the spatial analysis of small-area malaria incidence rates in Kwazulu Natal, South Africa », *American Journal of Epidemiology*, n°153(12), 2001, pages 1213-1221.

KOHLER H.-P. et HELLERINGER S., *The Structure of Sexual Networks and the Spread of HIV in Sub-Saharan Africa: Evidence from Likoma Island (Malawi)*, Philadelphie, Pennsylvanie (US), Population Agin Research Center - University of Pennsylvania, 2006, WPS 06-02, 22 pages. (http://www.pop.upenn.edu/rc/parc/aging_center/2006/PARCwps06-02.pdf)

KORBER B., GASCHEN B., YUSIM K. *et al.*, « Evolutionary and immunological implications of contemporary HIV-1 variation », *British Medical Bulletin*, n°58, 2001, pages 19-42.

KREMER-MARIETTI A., *Comment Popper comprit Einstein… et comment Einstein pensait réellement*, La Science einsteinienne : ses origines, son contenu et sa portée, Tunis (TN), 12-14 décembre, 2005. (http://dogma.free.fr/txt/AKM-PopperEinstein.htm)

KRIEGER J. N., COOMBS R. W., COLLIER A. C. *et al.*, « Fertility parameters in men infected with human immunodeficiency virus », *Journal of Infectious Diseases*, n°164(3), 1991, pages 464-469.

KRIGE D., « A statiscal approach to some basic mine valuation problemens on the witwatersrand », *Journal of the Chemical, Metallurgical and Mining Society*, n°52, 1951, pages 119-139.

KRUMBEIN W. C., « Regional and local components in facies maps », *AAPG Bulletin*, n°40(9), 1956, pages 2163-2194.

KWESIGABO G., KILLEWO J. Z. et SANDSTROM A., « Sentinel surveillance and cross sectional survey on HIV infection prevalence: a comparative study », *East African Medical Journal*, n°73(5), 1996, pages 298-302.

KWESIGABO G., KILLEWO J. Z., URASSA W. *et al.*, « Monitoring of HIV-1 infection prevalence and trends in the general population using pregnant women as a sentinel population: 9 years experience from the Kagera region of Tanzania », *Journal of Acquired Immune Deficiency Syndromes*, n°23(5), 2000, pages 410-417.

LACHAUD J. P., « HIV prevalence and poverty in Africa: micro- and macro-econometric evidences applied to Burkina Faso », *Journal of Health Economics*, n°26(3), 2007, pages 483-504.

LALOU R. et PICHE V., « Migration et sida en Afrique de l'Ouest: un état des connaissances », *Les Dossiers du Ceped*, n°28, 1994, pages 1-48.

LAPERCHE S. et LY T. D., « Sensitivity of HIV infection screening assays in 2001 », *Annales de Biologie Clinique*, n°60(3), 2002, pages 307-315.

LARMARANGE J. et FERRY B., *Estimation des niveaux de prévalence du VIH dans les pays d'Afrique sub-saharienne et ajustement possible à partir des femmes enceintes*, Chaire Quételet, Louvain-la-Neuve (BE), 19 novembre, Université Catholique de Louvain, 2004. (http://www.demo.ucl.ac.be/cq04/textes/Larmarange-Ferry.pdf)

LARMARANGE J., « Hommes ayant des rapports sexuels avec d'autres hommes (HSH) : une épidémie toujours active », *Transcriptases*, n°129, 2006, pages 72-74. (http://www.pistes.fr/transcriptases/129_549.htm)

LAW D. C., SERRE M. L., CHRISTAKOS G., LEONE P. A. et MILLER W. C., « Spatial analysis and mapping of sexually transmitted diseases to optimise intervention and prevention strategies », *Sexually Transmitted Infections*, n°80(4), 2004, pages 294-299.

LEE L., « Fertility reduction and duration of HIV infection : Findings from the United States », dans USAID, MEASURE EVALUATION et UNAIDS, *HIV, STI and Infertility: Past Trends and Current Monitoring Problems. Conference abstracts*, Measure Evaluation, Arlington, Virginia (US), 1998.

LEVINE N., *CrimeStat III: A Spatial Statistics Program for the Analysis of Crime Incident Locations*, Ned Levine & Associates, National Institute of Justice, Houston, Texas et Washington, District of Columbia (US), 2004. (http://www.icpsr.umich.edu/CRIMESTAT/)

LEYNAERT B., DOWNS A. M. et DE VINCENZI I., « Heterosexual transmission of human immunodeficiency virus: variability of infectivity throughout the course of infection. European Study Group on Heterosexual Transmission of HIV », *American Journal of Epidemiology*, n°148(1), 1998, pages 88-96.

LI L. et REVESZ P., *A Comparison of Spatio-temporal Interpolation Methods*, Geographic Information Science: Second International Conference, Boulder, Colorado (US), September 25-28, 2002.

LOUA A., SOW E. M., MAGASSOUBA F. B., CAMARA M. et BALDE M. A., « Evaluation du risque infectieux residuel chez les donneurs de sang au Centre national de transfusion sanguine de Conakry », *Transfusion Clinique et Biologique*, n°11(2), 2004, pages 98-100.

LU F., WANG N., WU Z. *et al.*, « Estimating the number of people at risk for and living with HIV in China in 2005: methods and results », *Sexually Transmitted Infections*, n°82(suppl_3), 2006, pages iii87-91.

LURIE M. N., WILLIAMS B. G., ZUMA K. *et al.*, « Who infects whom? HIV-1 concordance and discordance among migrant and non-migrant couples in South Africa », *AIDS*, n°17(15), 2003, pages 2245-2252.

LWANGA S. K. et CHIN J., *Projections of non-paediatric infection and AIDS in pattern II areas*, III International Conference on AIDS and Associated Cancers in Africa, Arusha (TZ), 14-16 septembre, 1988.

LYDIE N., ROBINSON N. J., FERRY B. *et al.*, « Mobility, sexual behavior, and HIV infection in an urban population in Cameroon », *Journal of Acquired Immune Deficiency Syndromes*, n°35(1), 2004, pages 67-74.

LYDIÉ N., *Les Chemins du Sida : migrations, mouvements de population et infection à VIH au Cameroun*, thèse de doctorat en géographie, sous la direction de POURTIER R., Université Paris 1 Panthéon-Sorbonne, UFR de Géographie, Paris (FR), 2001, 346 pages.

LYERLA R., GOUWS E., GARCIA-CALLEJA J. M. et ZANIEWSKI E., « The 2005 Workbook: an improved tool for estimating HIV prevalence in countries with low level and concentrated epidemics », *Sexually Transmitted Infections*, n°82(suppl_3), 2006, pages iii41-44.

MACH E., *La Connaissance et l'erreur*, Flammarion, coll. *Bibliothèque de philosophie scientifique*, traduit en 1908 sur la dernière édition allemande, par le Dr Marcel Dufour, disponible en ligne sur Gallica, Paris (FR), 1905, 392 pages. (http://catalogue.bnf.fr/ark:/12148/bpt6k655583)

MAHOMVA A., GREBY S., DUBE S. *et al.*, « HIV prevalence and trends from data in Zimbabwe, 1997-2004 », *Sexually Transmitted Infections*, n°82 Suppl 1, 2006, pages i42-47.

MANN J. M., TARANTOLA D. J. M. et NETTER T. W., *AIDS in the World*, Harvard University Press, Cambridge (US) et Londres (GB), 1992, 1037 pages.

MANN J. M. et TARANTOLA D. J. M., *AIDS in the World II*, Oxford University Press, New York City, New York (US), 1996, 616 pages.

MARCH L., « Observations sur la méthode représentative et sur le projet de rapport relatif à cette méthode », *Bulletin de l'Institut International de Statistique*, n°XXII(1), 1925, pages 444.

MARTIN P. M., GRESENGUET G., HERVE V. M. *et al.*, « Decreased number of spermatozoa in HIV-1-infected individuals », *AIDS*, n°6(1), 1992, pages 130.

MATHERON G., *Traité de géostatistique appliquée, Tome I*, Fontainebleau (FR), Éditions Technip, coll. *Mémoires du Bureau de Recherches Géologiques et Minières*, 1962, n°14, 334 pages.

MATHERON G., *Traité de géostatistique appliquée, Tome II : le krigeage*, Fontainebleau (FR), Éditions Technip, coll. *Mémoires du Bureau de Recherches Géologiques et Minières*, 1963a, n°24, 172 pages.

MATHERON G., « Principles of geostatistics », *Economic Geology*, n°58, 1963b, pages 1246-1266.

MATHERON G., *Estimer et Choisir : essai sur la pratique des probabilités*, Centre de Géostatistique, École Nationale Supérieure des Mines de Paris, coll. *Les cahiers du CMM de Fontainebleau*, Fontainebleau (FR), 1978, 175 pages. (http://cg.ensmp.fr/bibliotheque/1978/MATHERON/Ouvrage/DOC_00208/MATHERON_Ouvrage_00208.pdf)

MAUNY F., VIEL J. F., HANDSCHUMACHER P. et SELLIN B., « Multilevel modelling and malaria: a new method for an old disease », *International Journal of Epidemiology*, n°33(6), 2004, pages 1337-1344.

MCCUTCHAN F. E., « Understanding the genetic diversity of HIV-1 », *AIDS*, n°14 Suppl 3, 2000, pages S31-44.

MEASURE DHS, « Online HIV/AIDS Survey Indicator Database », *DHS+ Dimensions*, n°4(2), 2002, pages 9. (http://www.measuredhs.com/pubs/pub_details.cfm?Filename=Vol4no2.pdf&id=383)

MEASURE DHS, *Methodology - Collecting Geographic Data*, 2006, page web consultée le 10 juillet 2006. (http://www.measuredhs.com/topics/gis/methodology.cfm)

MÉDA N., GAUTIER-CHARPENTIER L., SOUDRE R. B. *et al.*, « Serological diagnosis of human immunodeficiency virus in Burkina Faso: reliable, practical strategies using less expensive commercial test kits », *Bulletin of the World Health Organization*, n°77(9), 1999, pages 731-739.

MEEKERS D., « Immaculate conceptions in sub-saharan Africa: exploratory analysis of inconsistencies in the timing of first sexual intercourse and first birth », *Social Biology*, n°42(3-4), 1995, pages 151-161.

MISHRA V., VAESSEN M., BOERMA J. T. *et al.*, « HIV testing in national population-based surveys: experience from the Demographic and Health Surveys », *Bulletin of the World Health Organization*, n°84(7), 2006, pages 537-545.

MOCK N. B., DUALE S., BROWN L. F. *et al.*, « Conflict and HIV: A framework for risk assessment to prevent HIV in conflict-affected settings in Africa », *Emerg Themes Epidemiol*, n°1(1), 2004, pages 6.

MONTAGNIER L., CHERMANN J. C., BARRE-SINOUSSI F. *et al.*, « Lymphadenopathy associated virus and its etiological role in AIDS », *Princess Takamatsu Symposia*, n°15, 1984, pages 319-331.

MONTANA L., NEUMAN M. et MISHRA V., *Spatial Modeling of HIV Prevalence in kenya*, coll. *DHS Worling Papers*, janvier, 2007, n°27, 31 pages. (http://www.measuredhs.com/pubs/pub_details.cfm?Filename=WP27.pdf&id=637)

NAAMARA W., *Official release of the National Serosurvey for human immunodeficiency virus (HIV) in Uganda*, Kampala (UG), AIDS Control Program, Ministry of Health, 1990.

NAGOT N., OUEDRAOGO A., FOULONGNE V. *et al.*, « Reduction of HIV-1 RNA levels with therapy to suppress herpes simplex virus », *New England Journal of Medicine*, n°356(8), 2007, pages 790-799.

NATIONAL AIDS CONTROL COMMITTEE, *Technical Report National Serosurvey on VIH/Syphilis*, Yaoundé (CM), Ministry of Public Health, April, 2001, 37 pages.

NATIONAL AIDS CONTROL COMMITTEE, *National HIV Sentinel surveillance Report (2002)*, Yaoundé (CM), Ministry of Public Health, June, 2003, 42 pages.

NATIONAL INSTITUTE OF STATISTICS, *Cameroon Statistical Yearbook 2004*, Yaoundé (CM), NIS, 2004, 635 pages. (http://www.statistics-cameroon.org/ins/annuaire.htm)

NATIONAL STATISTICAL OFFICE et ORC MACRO, *Malawi Demographic and Health Survey 2004*, Calverton, Maryland (US), NSO, ORC Macro, décembre, 2005, 482 pages. (http://www.measuredhs.com/pubs/pdf/FR175/FR-175-MW04.pdf)

NDINYA-ACHOLA J. O., WAMOLA I. A., NAGELKERKE N. et AL. E., « Impact of post-partum counselling of HIV infected women of their subsequent reproduction behaviour », *Kenya AIDS Technical Bulletin*, n°1, 1990.

NETTLETON L. L., « Regionals, Residuals and Structures », *Geophysics*, n°19(1), 1954, pages 1-22.

NICOLL A., STEPHENSON J., GRIFFIOEN A. *et al.*, « The relationship of HIV prevalence in pregnant women to that in women of reproductive age: a validated method for adjustment », *AIDS*, n°12(14), 1998, pages 1861-1867.

NNKO S., BOERMA J. T., URASSA M., MWALUKO G. et BASIA Z., *Secretive females or swaggering males? An assessment of the quality of sexual partnership reporting in rural Tanzania*, Chape Hill, North carolina (US), Carolina Population Center, coll. *Measure Evaluation Working Paper*, septembre, 2002, WP-02-57, 34 pages. (http://www.cpc.unc.edu/measure/publications/pdf/wp-02-57.pdf)

NTOZI J. P., « Widowhood, remarriage and migration during the HIV/AIDS epidemic in Uganda », *Health Transition Review*, n°7(Supplement), 1997, pages 125-144.

OLIVER M. A., MUIR K. R., WEBSTER R. *et al.*, « A geostatistical approach to the analysis of pattern in rare disease », *Journal of Public Health Medicine*, n°14(3), 1992, pages 280-289.

OLIVER M. A., WEBSTER R., LAJAUNIE C. *et al.*, « Binomial cokriging for estimating and mapping the risk of childhood cancer », *IMA Journal of Mathematics Applied in Medicine and Biology*, n°15(3), 1998, pages 279-297.

OMS, « Syndrome d'Immunodéficience Acquise (SIDA) Réunion de l'OMS », *Relevé Épidémiologique Hebdomadaire*, n°58(48), 1983a, pages 369-370. (http://whqlibdoc.who.int/wer/WHO_WER_1983/WER1983_58_369-376%20(N%C2%B048).pdf)

OMS, « Syndrome Immunodéficitaire Acquis (SIDA) », *Relevé Épidémiologique Hebdomadaire*, n°58(14), 1983b, pages 101-102. (http://whqlibdoc.who.int/wer/WHO_WER_1983/WER1983_58_101-108%20(N%C2%B014).pdf)

OMS, « Syndrome Immunodéficitaire Acquis (SIDA) », *Relevé Épidémiologique Hebdomadaire*, n°58(29), 1983c, pages 227-228. (http://whqlibdoc.who.int/wer/WHO_WER_1983/WER1983_58_221-228%20(N%C2%B029).pdf)

OMS, « Atelier sur le SIDA en Afrique Centrale, Bangui, 22-25 octobre 1985 », *Relevé Épidémiologique Hebdomadaire*, n°60(44), 1985, pages 342. (http://whqlibdoc.who.int/wer/WHO_WER_1985/WER1985_60_337-344%20(N%C2%B044).pdf)

OMS, « SIDA - Plan d'action pour la lutte dans la Région Africaine », *Relevé Épidémiologique Hebdomadaire*, n°61(13), 1986a, pages 93. (http://whqlibdoc.who.int/wer/WHO_WER_1986/WER1986_61_93-100%20(N%C2%B013).pdf)

OMS, « SIDA - Données Mondiales », *Relevé Épidémiologique Hebdomadaire*, n°61(47), 1986b, pages 361. (http://whqlibdoc.who.int/wer/WHO_WER_1986/WER1986_61_361-368%20(N%C2%B047).pdf)

OMS, « Estimations de travail provisoires de la prévalence du VIH chez les adultes, à la fin 1994, par pays », *Relevé Épidémiologique Hebdomadaire*, n°70(50), 1995, pages 355-357. (http://whqlibdoc.who.int/wer/WHO_WER_1995/WER1995_70_353-360%20(N%C2%B050).pdf)

OMS, « Surveillance du SIDA - Partie 1 », *Relevé Épidémiologique Hebdomadaire*, n°72(48), 1997, pages 357-360. (http://www.who.int/docstore/wer/pdf/1997/wer7248.pdf)

OMS, *Le Virus de l'immunodéficience humaine et son diagnostic : manuel de référence à l'usage des personnels de laboratoire*, Brazzaville (CG), OMS, Bureau régional de l'Afrique, Division des maladies transmissibles, Programme régional SIDA, 2004, 60 pages. (http://www.afro.who.int/aids/laboratory_services/hiv_french.pdf)

OMS, *Constitution de l'Organisation Mondiale de la Santé*, suppléments à la 45e édition, Genève (CH), 2006. (http://www.who.int/entity/governance/eb/who_constitution_fr.pdf)

OMS EURO, *L'Enseignement de l'épidémiologie en médecine et santé publique : rapport sur un symposium*, OMS, document EURO 0393, Copenhague (DK), 1968.

ONU, *Manuel X. Techniques indirectes d'estimation démographique*, ONU, New York City, New York (US), 1984, 324 pages.

ONUSIDA/OMS, *Les maladies sexuellement transmissibles : politiques et principes de prévention et de soins*, coll. *Meilleures Pratiques de l'ONUSIDA*, ONUSIDA/OMS/97.6, Genève (CH), 1997a, 50 pages. (http://www.who.int/hiv/pub/sti/en/prev_care_fr.pdf)

ONUSIDA/OMS, « Recommandations concernant le choix et l'utilisation des tests de mise en évidence des anti-corps anti-VIH - Version révisée », *Relevé Épidémiologique Hebdomadaire*, n°72(12), 1997b, pages 81-87. (http://www.who.int/docstore/wer/pdf/1997/wer7212.pdf)

ORROTH K. K., WHITE R., FREEMAN E. E. *et al.*, *Four cities modelling: #2 the dynamic impact of male circumcision and curable STIs on the heterogeneity of HIV epidemics in sub-Saharan Africa - simulation results*, XVI International AIDS Conference, Toronto (CA), 2006. (http://www.aids2006.org/PAG/Abstracts.aspx?AID=12017)

OUATTARA H., SIRANSY-BOGUI L., FRETZ C. *et al.*, « Residual risk of HIV, HVB and HCV transmission by blood transfusion between 2002 and 2004 at the Abidjan National Blood Transfusion Center », *Transfusion Clinique et Biologique*, n°13(4), 2006, pages 242-245.

OUEDRAOGO A., NAGOT N., VERGNE L. *et al.*, « Impact of suppressive herpes therapy on genital HIV-1 RNA among women taking antiretroviral therapy: a randomized controlled trial », *AIDS*, n°20(18), 2006, pages 2305-2313.

PALLONI A. et GLICKLICH M., « Review of approaches to modelling the demograpic impact of the AIDS epidemic », dans UN/WHO, *The AIDS epidemic and its demographic consequences*, UN/WHO, New York City, New York (US), 1991, pages 20-50.

PARENT C., « La "prostitution" ou le commerce des services sexuels », dans DUMONT F., LANGLOIS S. et MARTIN Y., *Traité des problèmes sociaux*, Institut québécois de recherche sur la culture, Québec (CA), 1994, pages 393-410. (http://classiques.uqac.ca/contemporains/parent_colette/prostitution_commerce_sexe/prostitution_commerce_sexe.pdf)

PARK C., GACHET D., UDRESSY O. et VARONE M.-P., *Fiche technique dépistage du VIH*, Groupe SIDA Genève, 2005, page web consultée le 27 août 2007. (http://www.groupesida.ch/04/elements/documents/article_depistage.pdf)

PASSERON J.-C., *Le Raisonnement sociologique, un espace non poppérien de l'argumentation*, Albin Michel, nouvelle édition revue et augmentée (première édition 1991), bibliothèque de l'Évolution de l'Humanité, Paris (FR), 2006, 666 pages.

PEBESMA E. J., « Multivariable geostatistics in S: the gstat package », *Computers & Geosciences*, n°30, 2004, pages 683-691.

PEREZ G., SKURNICK J. H., DENNY T. N. *et al.*, « Herpes simplex type II and Mycoplasma genitalium as risk factors for heterosexual HIV transmission: report from the heterosexual HIV transmission study », *International Journal of Infectious Diseases*, n°3(1), 1998, pages 5–11.

PETERS P., MEINZEN-DERR J., KAYITENKORE K. *et al.*, *HIV-1 positive Rwandan women have a high frequency of long-terme survival: 20-year follow-up from a prospective cohort study*, XVIth International AIDS Conference, Toronto (CA), août, 2006, abstract n°WEAX0304.

PETERSEN L. R., SATTEN G. A., DODD R. *et al.*, « Duration of time from onset of human immunodeficiency virus type 1 infectiousness to development of detectable antibody. The HIV Seroconversion Study Group », *Transfusion*, n°34(4), 1994, pages 283-289.

PILLONEL J., LAPERCHE S., SAURA C., DESENCLOS J. C. et COUROUCE A. M., « Trends in residual risk of transfusion-transmitted viral infections in France between 1992 and 2000 », *Transfusion*, n°42(8), 2002, pages 980-988.

PLANTIER J.-C. et SIMON F., « Diagnistic sérologique des infections à VIH », *Développement et Santé*, n°162, 2002. (http://documentation.ledamed.org/IMG/html/doc-10797.html)

POPPER K. R., *Conjectures et Réfutations : la croissance du savoir scientifique*, Payot, coll. *Bibliothèque scientifique*, traduction de Michelle-Irène et Marc-B de Launey de 1985, Paris (FR), 1963, 610 pages.

POPPER K. R., *La Logique de la découverte scientifique*, éditions Payot, 5e édition de 2002 de l'édition française de 1989, basée sur l'édition anglaise de 1968, elle-même une édition augmentée et corrigée de la première édition allemande de 1935, Paris (FR), 1968, 480 pages.

QUINN T. C., « Population migration and the spread of types 1 and 2 human immunodeficiency viruses », *Proceedings of the National Academy of Sciences of the United States of America*, n°91(7), 1994, pages 2407-2414.

R DEVELOPMENT CORE TEAM, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienne (AT), 2006. (http://www.R-project.org)

REMY G., « L'Infection à VIH1 en Afrique Sud-Saharienne : la priorité urbaine reconsidérée », *Médecine d'Afrique Noire*, n°46(8-9), 1999, pages 388-393.

RÉMY G., « Image géographique de l'infection à VIH1 en Afrique Centrale: des discontinuités remarquables », *Annales de la Societe Belge de Medecine Tropicale*, n°73(2), 1993, pages 127-142.

RÉSEAU MIGRATIONS ET URBANISATION EN AFRIQUE DE L'OUEST (REMUAO), *Enquête sur les Migrations et l'Urbanisation au Burkina Faso (EMUBF) 1992-1993 Rapport national descriptif*, Bamako (ML), CERPOD, december, 1997, 140 pages.

RICE B. D., BATZING-FEIGENBAUM J., HOSEGOOD V. *et al.*, « Population and antenatal-based HIV prevalence estimates in a high contracepting female population in rural South Africa », *BMC Public Health*, n°7, 2007, pages 160.

ROSS A., MORGAN D., LUBEGA R. *et al.*, « Reduced fertility associated with HIV: the contribution of pre-existing subfertility », *AIDS*, n°13(15), 1999, pages 2133-2141.

RUMEAU-ROUQUETTE C., BRÉART G. et PADIEU R., *Méthodes en épidémiologie*, Flammarion Médecine-Sciences, coll. *Statistique en Biologie et en Médecine*, seconde édition, Paris (FR), 1981, 306 pages.

RYDER R. W., BATTER V. L., NSUAMI M. *et al.*, « Fertility rates in 238 HIV-1-seropositive women in Zaire followed for 3 years post-partum », *AIDS*, n°5(12), 1991, pages 1521-1527.

SAKAROVITCH C., *Estimation de l'incidence de l'infection par le VIH en Afrique : application à la Côte d'Ivoire*, Thèse de Doctorat option Épidémiologie et Intervention en Santé Publique, sous la direction de ALIOUM A., Université Victor Segalen Bordeaux 2, Bordeaux (FR), 2006, 154 pages.

SALAMA P. et DONDERO T. J., « HIV surveillance in complex emergencies », *AIDS*, n°15 Suppl 3, 2001, pages S4-12.

SANGARE A., LEONARD G. et GERSHY-DAMET G., *Epidemiology of HIV-1 and HIV-2 Virus in Ivory Coast during the period 1986-1989*, IV International Conference on AIDS and Associated Cancers in Africa, Marseille (FR), Octobre, 1989.

SAPHONN V., HOR L. B., LY S. P. *et al.*, « How well do antenatal clinic (ANC) attendees represent the general population? A comparison of HIV prevalence from ANC sentinel surveillance sites with a population-based survey of women aged 15-49 in Cambodia », *International Journal of Epidemiology*, n°31(2), 2002, pages 449-455.

SCHREIBER G. B., BUSCH M. P., KLEINMAN S. H. et KORELITZ J. J., « The risk of transfusion-transmitted viral infections. The Retrovirus Epidemiology Donor Study », *New England Journal of Medicine*, n°334(26), 1996, pages 1685-1690.

SCHWARTLANDER B., STANECKI K. A., BROWN T. *et al.*, « Country-specific estimates and models of HIV and AIDS: methods and limitations », *AIDS*, n°13(17), 1999, pages 2445-2458.

SCHWARTZ D., *L'explication en épidémiologie*, Chaire Quetelet 1987, Louvain-la-Neuve (BE), 13-16 octobre 1987, Ciaco, Institut de Démographie - Université Catholique de Louvain, 1989, 127-140 pages.

SELIK R. M., HAVERKOS H. W. et CURRAN J. W., « Acquired immune deficiency syndrome (AIDS) trends in the United States, 1978-1982 », *American Journal of Medicine*, n°76(3), 1984, pages 493-500.

SÉNÈQUE, *Questions Naturelles*, traduction de Charpentier et Lemaistre publiée originalement en 1861 chez Garnier, numérisée et mise en ligne dans le cadre du projet Itinera Electronica de l'Université Catholique de Louvain, Paris (FR), vers 62. (http://bcs.fltr.ucl.ac.be/sen/qnII.html)

SERWADDA D., WAWER M. J., MUSGRAVE S. D. *et al.*, « HIV risk factors in three geographic strata of rural Rakai Distric, Uganda », *AIDS*, n°6, 1992, pages 983-989.

SERWADDA D., GRAY R. H., WAWER M. J. *et al.*, « The social dynamics of HIV transmission as reflected through discordant couples in rural Uganda », *AIDS*, n°9(7), 1995, pages 745-750.

SHAHMANESH M., GAYED S., ASHCROFT M. *et al.*, « Geomapping of chlamydia and gonorrhoea in Birmingham », *Sexually Transmitted Infections*, n°76(4), 2000, pages 268-272.

SHAIKH N., ABDULLAH F., LOMBARD C. J. *et al.*, « Masking through averages--intraprovincial heterogeneity in HIV prevalence within the Western Cape », *South African Medical Journal*, n°96(6), 2006, pages 538-543.

SHANG G., SEED C. R., WANG F., NIE D. et FARRUGIA A., « Residual risk of transfusion-transmitted viral infections in Shenzhen, China, 2001 through 2004 », *Transfusion*, n°47(3), 2007, pages 529-539.

SHEPARD D., *A two-dimensional interpolation function for irregularly-spaced data*, Proceedings of the 1968 23rd ACM national conference, ACM Press, 1968, 517-524 pages. (http://doi.acm.org/10.1145/800186.810616)

SHISANA O. et SIMBAYI L., *Nelson Mandela/HSRC study of HIV/AIDS: South African national HIV prevalence, behavioural risks and mass media-household survey 2002*, Cape Town (ZA), Human Sciences Research Council, 2002.

SIMONSEN J. N., CAMERON D. W., GAKINYA M. N. *et al.*, *Human immunodeficiency virus infection among men with sexually transmitted diseases. Experience from a center in Africa*, 1988, 274-278 pages.

SLUTKIN G., CHIN J., TARANTOLA D. et MANN J. M., *Sentinel surveillance for HIV infection: a method to monitor HIV infection trends in population groups*, IV International Conference on AIDS, Stockholm (SE), Juin, WHO, 1988, WHO/GPA/DIR/88.8. (http://whqlibdoc.who.int/hq/1988/WHO_GPA_DIR_88.8.pdf)

SLUTKIN G., CHIN J., TARANTOLA D. et MANN J., *Use of HIV surveillance data in national AIDS control programmes: a review of current data use with recommendations for strengthening future use*, Genève (CH), GPA/WHO, 1990, WHO/GPA/SFI/90.1, 12 pages. (http://whqlibdoc.who.int/hq/1990/WHO_GPA_SFI_90.1.pdf)

SP/CONASUR, UNICEF et PAM, *Analyse des données sur les rapatriés de Côte d'Ivoire*, Ouagadougou (BF), Comité National de Secours d'Urgence et de Réhabilitation, septembre, 2004, 65 pages.

SPIEGEL P. B., « HIV/AIDS among conflict-affected and displaced populations: dispelling myths and taking action », *Disasters*, n°28(3), 2004, pages 322-339.

SPIEGEL P. B., BENNEDSEN A. R., CLAASS J. *et al.*, « Prevalence of HIV infection in conflict-affected and displaced people in seven sub-Saharan African countries: a systematic review », *The Lancet*, n°369(9580), 2007, pages 2187-2195.

STOVER J., « Projecting the demographic consequences of adult HIV prevalence trends: the Spectrum Projection Package », *Sexually Transmitted Infections*, n°80 Suppl 1, 2004, pages I14-I18.

STOVER J., *AIM version 4 : Programme informatique pour réaliser des projectionssur le VIH/SIDA et examiner ses impacts socio-économiques*, Washington, DC (US), USAID, Health Policy Initiative, mars, 2007, 87 pages. (http://data.unaids.org/pub/Manual/2007/aim_manual_2007_fr.pdf)

TAFFÉ P., *Cours de Régression Logistique Appliquée*, Lausanne (CH), Institut Universitaire de Médecine Sociale et Préventive et Centre d'Épidémiologie Clinique, août, 2004, 64 pages. (http://www.tesser-pro.org/stat/Cours_regression_logistique.pdf)

TATT I. D., BARLOW K. L., NICOLL A. et CLEWLEY J. P., « The public health significance of HIV-1 subtypes », *AIDS*, n°15 Suppl 5, 2001, pages S59-71.

THE UNAIDS EPIDEMIOLOGY REFERENCE GROUP, *Recommendations*, UNAIDS Epidemiology Reference Group Meeting, Rome (IT), 8-10 octobre, UNAIDS, 2000, 18 pages. (http://www.epidem.org/Publications/Rome2000rec.pdf)

THE UNAIDS EPIDEMIOLOGY REFERENCE GROUP, *Recommended methodology for the estimation and projection of HIV prevalence and AIDS mortality in the short-term*, Meeting of the UNAIDS Epidemiology Reference Group, Gex (FR), Janvier, 2001, 29 pages. (http://www.epidem.org/Publications/Meeting%20summary%20and%20recommendations%20final.pdf)

THE UNAIDS REFERENCE GROUP ON ESTIMATES MODELLING AND PROJECTIONS, « Improved methods and assumptions for estimation of the HIV/AIDS epidemic and its impact: Recommendations of the UNAIDS Reference Group on Estimates, Modelling and Projections », *AIDS*, n°16(9), 2002, pages W1-14.

THE UNAIDS REFERENCE GROUP ON ESTIMATES MODELLING AND PROJECTIONS, *Responding to surveillance: Methods and software to produce HIV/AIDS estimates in the era of population-based prevalence surveys*, Technical Report and Recommendations - Report of a meeting on the UNAIDS Reference Group, Glion (CH), May 10-11th, UNAIDS, 2004a, 18 pages. (http://www.epidem.org/Publications/Glion%20report%20final.pdf)

THE UNAIDS REFERENCE GROUP ON ESTIMATES MODELLING AND PROJECTIONS, *Development of EPPv2 and Spectrum, and Measuring and tracking concentrated HIV epidemics*, Technical Report and Recommendations - Report of a meeting on the UNAIDS Reference Group, Sintra (PT), December 8-10th 2004, 2004b, 22 pages. (http://www.epidem.org/Publications/Sintra%20Dec%202004%20report.pdf)

THE UNAIDS REFERENCE GROUP ON ESTIMATES MODELLING AND PROJECTIONS, *Improving parameter estimation, projection, methods, uncertainty estimation and epidemic classification*, Report of a meeting, Prague (CZ), 29 novembre - 1er décembre, UNAIDS, 2006a, 25 pages. (http://www.epidem.org/Publications/Prague2006report.pdf)

THE UNAIDS REFERENCE GROUP ON ESTIMATES MODELLING AND PROJECTIONS, *Improvements to estimation packages*, Technical Report and Recommendations, Glion (CH), 19-20 juillet, UNAIDS, 2006b, 11 pages. (http://www.epidem.org/Publications/Glion2006Report.pdf)

THE UNAIDS REFERENCE GROUP ON ESTIMATES MODELLING AND PROJECTIONS, *Methods for estimation of ART's impact on death/delayed; and Developmentin EPP 2007*, Technical Report and Recommendations, Baltimore, Maryland (US), 13 juillet, UNAIDS, 2007, 10 pages. (http://www.epidem.org/Publications/Baltimore2007_13July.pdf)

THUAN T. X., « Préface », dans GREENE B., *L'Univers élégant*, Robert Laffont, Paris (FR), 2000, pages 9-13.

TOULEMON L., *Régression logistique et régression sur les risques : deux supports de cours*, Paris (FR), INED, coll. *Dossiers et Recherches*, 1995, n°46, 56 pages.

ULLMO J., « Les Concepts physiques », dans PIAGET J., *Logique et connaissance scientifique*, Gallimard, coll. *Encyclopédie de la Pléiade*, Paris (FR), 1967.

ULLMO J., *La Pensée scientifique moderne*, Flammarion, réédition de 2000 en format poche dans la collection Champs, Paris (FR), 1969, 315 pages.

UN/WHO, *The AIDS epidemics and its demographic consequences*, New York City, New York (US), United Nations, World Health Organization, Mai, 1991, ST/ESA/SER.A/119, 140 pages.

UNAIDS, *Report of the global HIV/AIDS epidemic June 2000*, Genève (CH), UNAIDS, 2000, UNAIDS/00.13E. (http://whqlibdoc.who.int/unaids/2000/global_report_2000.pdf)

UNAIDS, *Report on the global HIV/AIDS epidemic 2002*, Genève (CH), UNAIDS, July, 2002, UNAIDS/02.26E. (http://whqlibdoc.who.int/unaids/2002/global_report_2002.pdf)

UNAIDS, *2004 Report on the global AIDS epidemic 4th global report*, Genève (CH), UNAIDS, June, 2004a, 236 pages. (http://www.unaids.org/bangkok2004/GAR2004_pdf/UNAIDSGlobalReport2004_en.pdf)

UNAIDS, *UNAIDS response to Kenyan HIV prevalence survey*, Press statement, Genève (CH), 13 janvier, 2004b. (http://data.unaids.org/Media/PressStatements01/ps_kenyan_report_13jan04_en.pdf)

UNAIDS, *EPP manual for generalized epidemics*, 2005, page web consultée le march 10th 2007. (http://data.unaids.org/Topics/Epidemiology/Manuals/EPP_GeneralizedEpidemic_05_en.pdf)

UNAIDS, *Report on the global AIDS epidemic*, Genève (CH), May, 2006, UNAIDS/06.13E, 629 pages. (http://www.unaids.org/en/HIV_data/2006GlobalReport/default.asp)

UNAIDS, *UNAIDS/WHO working group on global HIV/AIDS and STI surveillance*, 2007a, page web consultée le 30 juillet 2007. (http://www.unaids.org/en/HIV_data/Epidemiology/epiworkinggrp.asp)

UNAIDS, *Practical Guidelines for Intensifying HIV Prevention: towards universal access*, Genève (CH), UNAIDS, 6 mars, 2007b, UNAIDS 07/07.E, 70 pages. (http://data.unaids.org/pub/Agenda/2007/20070306_prevention_guidelines_towards_universal_access%5D_en.pdf?preview=true)

UNAIDS/WHO, *Report on the global HIV/AIDS epidemic, June 1998*, Genève (CH), UNAIDS, WHO, Juin, 1998, 75 pages. (http://whqlibdoc.who.int/unaids/1998/global_report_1998.pdf)

UNAIDS/WHO, *Reconciling antenatal clinic-based surveillance and population-based survey estimates of HIV prevalence in sub-Saharan Africa*, UNAIDS/WHO, Genève (CH), 2003, 30 pages.

UNAIDS/WHO WORKING GROUP ON GLOBAL HIV/AIDS AND STI SURVEILLANCE, *Guidelines for Second Generation HIV Surveillance*, Genève (CH), UNAIDS/WHO, 2000, WHO/CDS/CSR/EDC/2000.5, UNAIDS/00.03E, iv+42 pages. (http://data.unaids.org/Publications/IRC-pub01/jc370-2ndgeneration_en.pdf?preview=true)

UNAIDS/WHO WORKING GROUP ON GLOBAL HIV/AIDS AND STI SURVEILLANCE, *Initiating second generation HIV surveillance systems: practical guidelines*, Genève (CH), UNAIDS/WHO, 2002, UNAIDS/02.45E, WHO/HIV/2002.17, 27 pages. (http://data.unaids.org/Publications/IRC-pub02/jc742-initiatingsgs_en.pdf?preview=true)

UNAIDS/WHO WORKING GROUP ON GLOBAL HIV/AIDS AND STI SURVEILLANCE, *Guidelines for measuring national HIV prevalence in population-based surveys*, Genève (CH), UNAIDS/WHO, 2005, 70 pages. (http://data.unaids.org/pub/Manual/2005/20050101_GS_GuideMeasuringPopulation_en.pdf)

UNHCR, *Refugees and others of concern to UNHCR - 1999 statiscal overview*, Genève (CH), UNHCR, juillet, 2000, 132 pages. (http://www.unhcr.org/cgi-bin/texis/vtx/statistics/opendoc.pdf?tbl=STATISTICS&id=3ae6bc834#zoom=100)

UNHCR, *2003 Global Refugee Trends: Overview of refugee populations, new arrivals, durable solutions, ayslum-seekers and other persons of concern to UNHCR*, Genève (CH), UNHCR, 15 juin, 2004a, 94 pages. (http://www.unhcr.org/cgi-bin/texis/vtx/statistics/opendoc.pdf?tbl=STATISTICS&id=40d015fb4)

UNHCR, *Global Report 2003 - UNHCR in Kenya (Map)*, UNHCR, Genève (CH), 2004b. (http://www.unhcr.org/publ/PUBL/40c573b50.pdf)

UNITED NATIONS CARTOGRAPHIC SECTION, *Map of Kenya*, United Nations, carte 4187 rev 1, 2004a. (http://www.un.org/Depts/Cartographic/map/profile/kenya.pdf)

UNITED NATIONS CARTOGRAPHIC SECTION, *Map of Cameroon*, United Nations, carte 4227, 2004b. (http://www.un.org/Depts/Cartographic/map/profile/cameroon.pdf)

UNITED NATIONS CARTOGRAPHIC SECTION, *Map of Burkina Faso*, United Nations, carte 4230, 2004c. (http://www.un.org/Depts/Cartographic/map/profile/burkina.pdf)

UNITED NATIONS POPULATION DIVISION, *World Population Prospects: the 2006 Revision and World urbanization Prospects: the 2005 Revision*, Population Division of the Department of Economic and Social Affairs of the United Nations Secretariat, 2007. (http://esa.un.org/unpp)

US CENSUS BUREAU, *HIV/AIDS Surveillance Data Base - June 2001*, Washington, DC (US), Health Studies Branch, International Programs Center, Population Division, US Census Bureau, Juin, 2001, CD-Rom. (http://www.census.gov/ipc/www/hivaidsn.html)

VELATI C., ROMANO L., BARUFFI L. *et al.*, « Residual risk of transfusion-transmitted HCV and HIV infections by antibody-screened blood in Italy », *Transfusion*, n°42(8), 2002, pages 989-993.

WADE A. S., KANE C. T., DIALLO P. A. *et al.*, « HIV infection and sexually transmitted infections among men who have sex with men in Senegal », *AIDS*, n°19(18), 2005, pages 2133-2140.

WALKER N., GARCIA-CALLEJA J. M., HEATON L. *et al.*, « Epidemiological analysis of the quality of HIV sero-surveillance in the world: how well do we track the epidemic? », *AIDS*, n°15(12), 2001, pages 1545-1554.

WALKER N., STOVER J., STANECKI K. *et al.*, « The workbook approach to making estimates and projecting future scenarios of HIV/AIDS in countries with low level and concentrated epidemics », *Sexually Transmitted Infections*, n°80 Suppl 1, 2004, pages I10-I13.

WANE H. R., TRAORÉ S., PICHÉ V., DIÈYE A. M. et DICKO A. B., *Atlas Migration et Environnement au Sahel*, Bamako (ML), CERPOD, INSAH, CILSS, novembre, 2000, 24 + 45 planches pages.

WAWER M., SERWADDA D., MUSGRAVE S. *et al.*, « Dynamics of HIV-1 infection in a rural district of Uganda », *British Medical Journal*, n°303, 1991, pages 1303-1306.

WAWER M. J., GRAY R. H., SEWANKAMBO N. K. *et al.*, « Rates of HIV-1 Transmission per Coital Act, by Stage of HIV-1 Infection, in Rakai, Uganda », *Journal of Infectious Diseases*, n°191(9), 2005, pages 1403-1409.

WAY P. O., *HIV/AIDS in Sub-Saharan Africa*, présentation à la National Academy of Sciences Committee on Population and Demography, 5-6 mars, Center for International Research US Bureau of the Census, 1992.

WEBER B., FALL E. H., BERGER A. et DOERR H. W., « Reduction of diagnostic window by new fourth-generation human immunodeficiency virus screening assays », *Journal of Clinical Microbiology*, n°36(8), 1998, pages 2235-2239.

WHO, *Workshop on AIDS in Central Africa 1985*, Genève (CH), World Health Organization, 1986, WHO/CDS/AIDS 85.1.

WHO, UNICEF et UNAIDS, *Burkina Faso - Epidemiological Fact Sheets on HIV/AIDS and Sexually Transmitted Infections 2006 Update*, Genève (CH), WHO, UNICEF, UNAIDS, 2006a, 20 pages. (http://www.who.int/GlobalAtlas/predefinedReports/EFS2006/index.asp)

WHO, UNICEF et UNAIDS, *Kenya - Epidemiological Fact Sheets on HIV/AIDS and Sexually Transmitted Infections 2006 Update*, Genève (CH), WHO, UNICEF, UNAIDS, 2006b, 26 pages. (http://www.who.int/GlobalAtlas/predefinedReports/EFS2006/index.asp)

WHO, UNICEF et UNAIDS, *Cameroon - Epidemiological Fact Sheets on HIV/AIDS and Sexually Transmitted Infections 2006 Update*, Genève (CH), WHO, UNICEF, UNAIDS, 2006c, 24 pages. (http://www.who.int/GlobalAtlas/predefinedReports/EFS2006/index.asp)

WIDY-WIRSKI R., BERKLEY S., DOWNING R. *et al.*, « Evaluation of the WHO clinical case definition for AIDS in Uganda », *JAMA*, n°260(22), 1988, pages 3286-3289.

WIKIPEDIA, *Enzyme-linked immunosorbent assay*, page web consultée le 30 mai 2007. (http://fr.wikipedia.org/wiki/Enzyme-linked_immunosorbent_assay)

WUNSCH G., *Techniques d'analyse des données démographiques déficientes*, Ordina, Liège (BE), 1984, 202 pages.

YAHYA-MALIMA K. I., OLSEN B. E., MATEE M. I. et FYLKESNES K., « The silent HIV epidemic among pregnant women within rural Northern Tanzania », *BMC Public Health*, n°6, 2006, pages 109.

YAHYA-MALIMA K. I., MATEE M. I., EVJEN-OLSEN B. et FYLKESNES K., « High potential of escalating HIV transmission in a low prevalence setting in rural Tanzania », *BMC Public Health*, n°7, 2007, pages 103.

YENI P. et GROUPE DES EXPERTS « PRISE EN CHARGE MÉDICALE DES PERSONNES INFECTÉES PAR LE VIH », *Prise en charge médicale des personnes infectées par le VIH - Recommendations du groupe d'experts*, Paris (FR), Ministère de la Santé et des Solidarités, Médecine-Sciences Flammarion, 2006, 368 pages. (http://www.sante.gouv.fr/htm/actu/yeni_sida/rapport_experts_2006.pdf)

ZAAIJER H. L., V EXEL-OEHLERS P., KRAAIJEVELD T., ALTENA E. et LELIE P. N., « Early detection of antibodies to HIV-1 by third-generation assays », *Lancet*, n°340(8822), 1992, pages 770-772.

ZABA B., BOERMA T., PISANI E. et BAPTISTE N., *Estimation of levels and trends in age at first sex from surveys using survival analysis*, Chape Hill, North carolina (US), Carolina Population Center, coll. *Measure Evaluation Working Paper*, mars, 2002, WP-02-51, 29 pages. (http://www.cpc.unc.edu/measure/publications/pdf/wp-02-51.pdf)

ZABA B., URASSA M., MARSTON M. *et al.*, *Survival following HIV infection in the pre-ART era in a rural tanzanian cohort*, XVIth International AIDS Conference, Toronto (CA), août, 2006, abstract n°MOPE0272.

ZABA B. W. et GREGSON S., « Measuring the impact of HIV on fertility in Africa », *AIDS*, n°12 Suppl 1, 1998, pages S41-S50.

ZABA B. W., CARPENTER L. M., BOERMA J. T. *et al.*, « Adjusting ante-natal clinic data for improved estimates of HIV prevalence among women in sub-Saharan Africa », *AIDS*, n°14(17), 2000, pages 2741-2750.

ZENILMAN J. M., GLASS G., SHIELDS T. *et al.*, « Geographic epidemiology of gonorrhoea and chlamydia on a large military installation: application of a GIS system », *Sexually Transmitted Infections*, n°78(1), 2002, pages 40-44.

Sigles employés

AFRO	Région Africaine de l'OMS, correspond approximativement à l'Afrique subsaharienne (l'Algérie fait néanmoins partie de cette région par exemple).
AIDS	Acquired Immune Deficiency Syndrome (SIDA en français).
AIM	AIDS Impact Model (module du logiciel Spectrum).
AIS	AIDS Impact Survey.
ANRS	Agence Nationale de Recherche sur le SIDA et les hépatites virales.
APP	Agence de Protection des Programmes.
ARV	Antirétroviraux, traitements contre le VIH/SIDA.
ASSA	Actuarial Society of South Africa.
BEH	*Bulletin Épidémiologique Hebdomadaire*, édité par la DGS (France).
BLUP	Best Linear Unbiased Predictor.
BSS	Behavioral Surevillance Surveys, coordonnées par FHI (http://www.fhi.org/en/topics/bss.htm).
BUCREP	Bureau Central des Recensements et des Études de Population - Cameroun.
CBS	Central Bureau of Statistics – Kenya.
CDC	Center for Disease prevention and Control (USA).
CDI	Consommateurs de Drogues Injectables.
CEA	Commissariat à l'Énergie Atomique.

CeCILL	Acronyme pour Ce(A)C(nrs)I(NRIA)L(ogiciel)L(ibre). Licence de logiciel libre de droit français (http://www.cecill.info/).
CEPED	Centre Population et Développement.
CIR	Center for International Research of the United States Bureau of the Census.
CNLS	Conseil National de Lutte contre le SIDA.
CNLS-IST	Conseil National de Lutte contre le SIDA et les Infections Sexuellement Transmissibles.
CNRS	Centre National de la Recherche Scientifique.
CPN	Clinique Prénatale.
CPS	Contraceptive Prevalence Survey.
DCW	Digital Chart of the World.
DGS	Direction Générale de la Santé (France).
DHS	Demographic and Health Survey (EDS en français).
DMA	US Defense Mapping Agency.
EDS	Enquête Démographique et de Santé (DHS en anglais).
EDSBF-III	Troisième EDS réalisée au Burkina Faso en 2003.
EDSC-III	Troisième EDS réalisée au Cameroun en 2004.
EIS	Enquête sur les Indicateurs du SIDA.
ELISA	Enzyme-Linked ImmunoSorbent Assay, littéralement dosage d'immunosorption liée à enzyme, c'est-à-dire dosage immunoenzymatique sur support solide. Technique utilisée pour la recherche des anticorps anti-VIH.
EMF	Enquête Mondiale de Fécondité (WFS en anglais).
EPP	Epidemic Projection Package qui deviendra Estimation and projection Package. Logiciel de modélisation d'ONUSIDA.
ESRI	Environmental Systems Research Institute, Inc.
ESV	Enquête de Sérologie VIH.
FHI	Family Health International.
GAA	Global Area of Affinité (aire géographique d'affinité).
GNU GPL	GNU General Public License.
GPA	Global Programme on AIDS de l'OMS.
GPS	Global Positioning System, un système américain de positionnement mondial par satellite.
GPW	Gridded Population of the World.
GRUMP	Global Rural-Urban Mapping Project.

HIV	Human Immunodeficiency Virus (VIH en français).
HSBS	HIV/AIDS Sero-Behavioural Survey.
HSH	Hommes ayant des rapports Sexuels avec d'autres Hommes.
HSV2	Herpes Simplex Virus 2, virus responsable de l'herpès génital (le HSV1 étant responsable de l'herpès buccal, neuro-méningé et ophtalmique).
HTLV-III	Human T-Lymphotropic Virus-III, ancien nom du VIH.
ICF	Indice Conjoncturel de Fécondité, également nommé ISF.
IDW	Inverse Distance Weighting.
INRIA	Institut National de Recherche en Informatique et en Automatique.
INS	Institut National de la Statistique – Cameroun.
INSD	Institut National de la Statistique et de la Démographie – Burkina Faso.
INSEE	Institut National de la Statistique et des Études Économiques.
IRD	Institut de Recherche pour le Développement.
IRIS	Îlots Regroupés pour l'Information Statistique, découpage du territoire français par l'INSEE.
ISF	Indice Synthétique de Fécondité, également nommé ICF.
IST	Infections Sexuellement Transmissibles.
IUSSP	International Union for the Scientific Study of Population (UIESP en français).
IWG	United States Interagency Working Group.
LAV	Lymphadenopathy Associated Virus, ancien nom du VIH.
LGBT	Lesbiennes, Gays, Bisexuels, Transsexuels.
LGBTIQ	Lesbiennes, Gays, Bisexuels, Transsexuels, Intersexués, Queer.
MDGs	Millenium Development Goals, connus en français sous le nom d'Objectifs de Millénaire.
MEASURE	Monitoring and Evaluation to Assess and use Results.
MICS	Multiple Indicator Cluster Surveys, coordonnées par l'UNICEF (http://www.childinfo.org/mics/index.htm).
MMWR	*Morbidity and Mortality Weekly Report*, édité par le CDC (USA).
MoT	Modes of Transmission, modèle d'ONUSIDA sous la forme d'une feuille Excel permettant de faire des estimations du nombre de nouvelles infections.
NASSEP IV	4th National Sample Survey and Evaluation Programme.
OMS	Organisation Mondiale de la Santé (WHO en anglais).
ONUSIDA	Programme Commun des Nations unies sur le VIH/SIDA (UNAIDS en anglais).

ORI	Odds Relatif de l'Infection (voir annexe 1).
PEPFAR	US President's Emergency Plan for AIDS Relief.
PNLS	Programme National de Lutte contre le SIDA.
PTME	Prévention de la Transmission Mère-Enfant.
REH	*Relevé Épidémiologique Hebdomadaire*, édité par l'OMS (WER en anglais).
REMUA	Réseau Migration et Urbanisation en Afrique de l'Ouest.
RGPH	Recensement Général de la Population et de l'Habitat.
RHS	Reproductive Health Surveys, coordonnées par le CDC (http://www.cdc.gov/reproductivehealth/index.htm).
SBS	Sexual Behavior Surveys, coordonnées par MEASURE Evaluation (http://www.cpc.unc.edu/measure/).
SIDA	Syndrome de l'Immunodéficience Acquise (AIDS en anglais).
SIG	Système d'Information Géographique.
STI	Sexually Transmitted Infections (IST en français).
TS	Travailleuses du Sexe.
UAS	Unlinked Anonymous Screening.
UIESP	Union Internationale pour l'Étude Scientifique de la Population (IUSSP en anglais).
UNAIDS	Joint United Nations Programme on HIV/AIDS (ONUSIDA en français).
UNGASS	United Nations General Assembly Special Session on HIV/AIDS.
UNHCR	United Nations High Commissioner for Refugees – Haut Commissariat des Nations Unies pour les Réfugiés.
UNICEF	United Nations International Children's Emergency Fund soit, en français, Fonds des Nations unies pour l'Enfance.
UNPOP	Division de la Population de l'Organisation des Nations unies.
USA	United States of America (États-Unis en français).
USAID	United States Agency for International Development.
VIH	Virus de l'Immunodéficience Humaine (HIV en anglais).
VIH-	Séronégatif au VIH.
VIH+	Séropositif au VIH.
WER	*Weekly Epidemiological Record*, édité par l'OMS (REH en français).
WFS	World Fertility Survey (EMF en français).
WHO	World Health Organization (OMS en français).

Table des matières

Liste des tableaux

Liste des figures

Liste des encadrés

Liste des équations

Liste des postulats

Liste des annexes

Prévalences du VIH en Afrique :
validité d'une mesure

Annexes

Sommaire des annexes

Annexe 1
Ajustement possible de la prévalence du VIH mesurée en clinique prénatale en corrigeant la sous-fécondité des femmes séropositives au VIH

Communication signée par Joseph LARMARANGE et Benoît FERRY et présentée à la Chaire Quételet 2004, *Santé de la Reproduction au Nord et au Sud : de la connaissance à l'action*, organisée par l'Université Catholique de Louvain, du 17 au 20 novembre 2004 à Louvain-la-Neuve (Belgique), lors de la séance *Infections Sexuellement Transmissibles / VIH et SIDA* du 19 novembre 2004.

Le titre original de cette communication était *Estimation des niveaux de prévalence du VIH dans les pays d'Afrique subsaharienne et ajustement possible à partir des femmes enceintes*.

1.1 Introduction

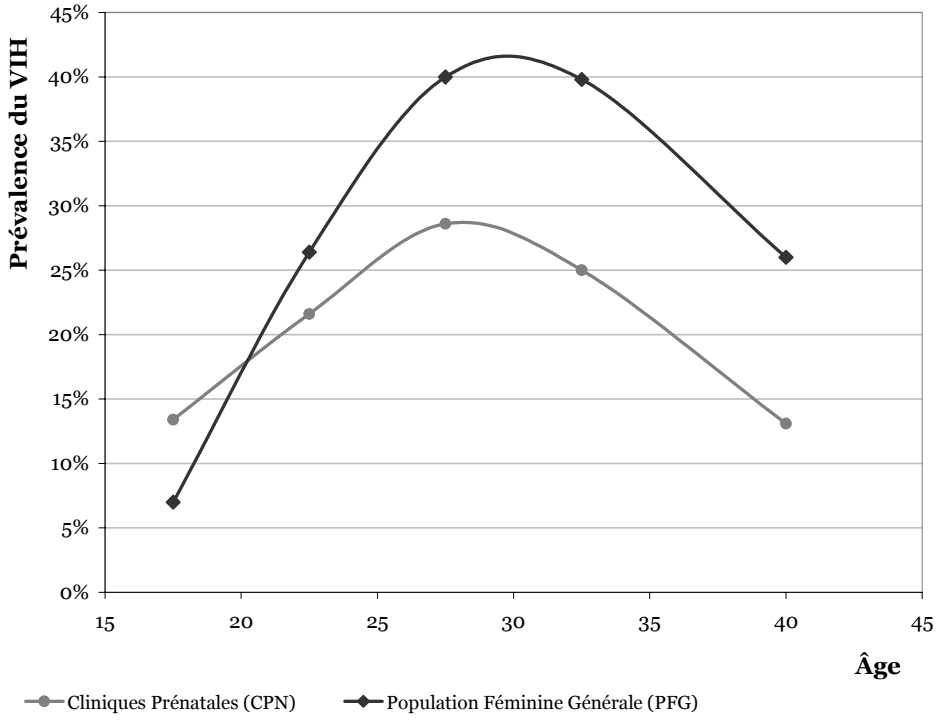
Dans les pays en développement et en Afrique subsaharienne en particulier, les données les plus fréquentes concernant la prévalence du VIH sont issues de sites sentinelles testant les femmes en suivi prénatal. Ces données sont à la base de la majorité des estimations de l'ONUSIDA (SCHWARTLANDER 1999), sous l'hypothèse que ces femmes enceintes sont représentatives de la population adulte sexuellement active[1] (BOISSON 1996). Ainsi en 2003, 118 pays dans le monde, incluant 39 pays d'Afrique subsaharienne, disposent d'un système de surveillance sentinelle qui s'appuie sur les cliniques prénatales (CPN) (UNAIDS/WHO 2003).

Concernant les estimations au niveau national, plusieurs biais interviennent et en premier lieu le choix des sites sentinelles retenus et leur représentativité, ainsi que le type de population qui les consultent (JACKSON 1999, SCHWARTLANDER 1999). La qualité des données est ainsi très variable d'un pays à l'autre (WALKER 2001, GARCIA-CALLEJA 2004). Les comparaisons dans le temps sont sujettes à caution étant donnée la variation importante dans un pays du nombre et de la localisation des sites sentinelles retenus. De nombreuses études ont également mis en évidence que les femmes qui consultaient en suivi prénatal n'avaient pas le même profil que les autres (GREGSON 1995, FYLKESNES 1998, FYLKESNES 2001, GLYNN 2001, GREGSON 2002), même si globalement il a été montré que les prévalences calculées à partir des données de femmes enceintes reflètent les tendances de la prévalence réelle (BORGDORFF 1993, KWESIGABO 2000, CRAMPIN 2003). Néanmoins, le fait même d'avoir recours aux femmes enceintes pour estimer la prévalence de la population générale induit des biais méthodologiques.

La prévalence des sites sentinelles en cliniques prénatales (CPN) sous-estime la prévalence de l'ensemble des femmes à tous les âges, excepté pour le groupe d'âges le plus jeune où il y a surestimation. Cela a été observé en Tanzanie (KWESIGABO 2000, CHANGALUCHA 2002), au Malawi (CRAMPIN 2003), en Zambie (FYLKESNES 1998), au Zimbabwe (GREGSON 1995, GREGSON 2002) et en Ouganda (CARPENTER 1997, GRAY 1998) où des données provenant à la fois de sites sentinelles et d'enquêtes en population générale sont disponibles (voir Figure annexe 1.1). Il s'agit d'un phénomène bien établi.

[1] L'hypothèse la plus commune consiste à considérer que les femmes enceintes sont représentatives de la population âgées de 15 à 49 ans, hommes et femmes confondus. Cependant, d'autres auteurs considèrent qu'elles sont représentatives seulement des femmes adultes.

Figure annexe 1.1
Prévalence du VIH selon l'âge des femmes en population générale et de celles consultant en clinique prénatale à Manicaland, Zimbabwe (1998-2000).



Sources : (GREGSON 2002)

Cette sous-estimation s'explique en partie par un différentiel de fécondité chez les femmes séropositives (ZABA 1998). Ces dernières sont moins fertiles (WIDY-WIRSKI 1988) et ont significativement plus d'avortements spontanés (BROCKLEHURST 1998). Elles ont, par ailleurs, plus fréquemment une stérilité préexistante des suites d'autres IST (ROSS 1999). Des facteurs comportementaux existent également mais leur effet est *a priori* faible. Ainsi, il a été montré que dans la région de Rakai (Ouganda), les femmes séropositives ont des rapports sexuels moins fréquents (GRAY 1998). À Kinshasa (Zaïre), elles utiliseraient plus fréquemment un moyen contraceptif (RYDER 1991). Une femme séropositive a plus souvent un partenaire séropositif, d'où moins de rapports sexuels du fait de la maladie du partenaire, et un risque de veuvage plus élevé (ZABA 1998). D'autre part, le VIH entraîne une baisse de la production de spermatozoïdes chez les hommes (KRIEGER 1991, MARTIN 1992). Bien que le statut sérologique soit rarement connu, une suspicion de séropositivité pourrait parfois entraîner une rupture du couple (NDINYA-ACHOLA 1990) et défavoriser le remariage des veuves et des divorcées (NTOZI 1997).

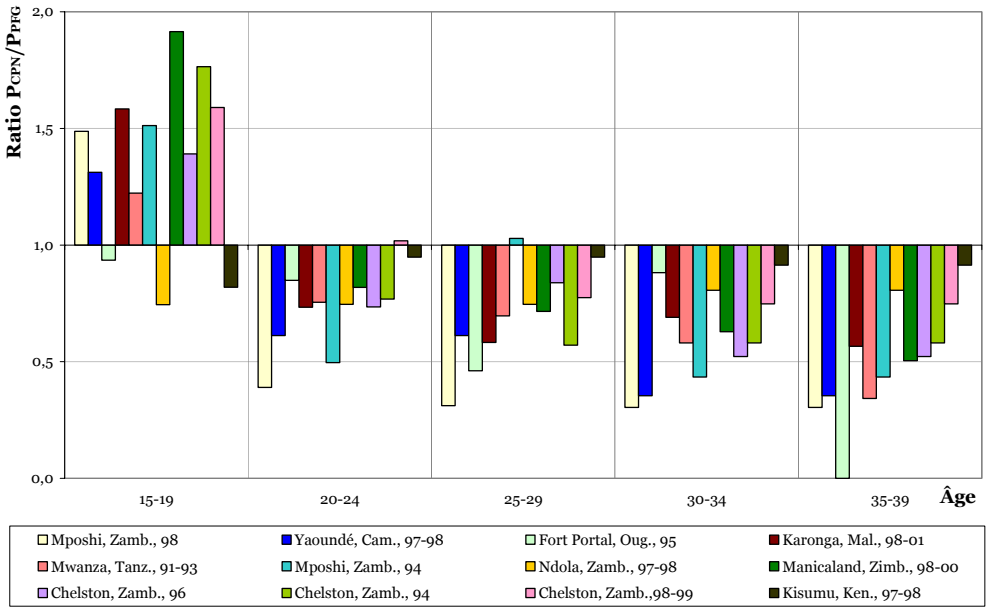
La sous-fécondité des femmes séropositives augmente avec l'âge de celles-ci. Seules les plus jeunes (15-19 ans) ont une fécondité plus élevée que les femmes séronégatives du même âge (CARPENTER 1997, GRAY 1998, CARPENTER 2002). Ceci

s'explique par un effet de sélection : les femmes de moins de vingt ans qui sont enceintes ont commencé leur vie sexuelle plus tôt et ont donc été plus soumises au risque d'être contaminées par le VIH.

Sur l'ensemble des sites que nous avons étudiés ici, nous avons retrouvé cette tendance d'une sous-estimation, augmentant avec l'âge, de la prévalence chez les femmes en suivi prénatale par rapport à celle observée en population générale (exception faite le plus souvent des 15-19 ans, voir des 20-24 ans, où il y a surestimation, voir Figure annexe 1.2).

Figure annexe 1.2

Ratio prévalence du VIH observée en clinique prénatale (CPN) sur prévalence en population féminine générale (PFG) selon l'âge



L'objectif de ce travail consiste à présenter une méthode d'ajustement décrite depuis quelques années et de l'appliquer sur une dizaine de sites en ayant recours à d'autres séries de coefficients que ceux précédemment utilisés. Cette méthode vise, en partant de la prévalence observée en clinique prénatale, à estimer la prévalence chez les femmes en population générale, en tenant compte des différentiels de fécondité entre femmes VIH+ et VIH-. Pour vérifier la pertinence de cette méthode, nous travaillerons sur des sites où l'on dispose à la fois des prévalences en cliniques prénatales et des prévalences en population générale, afin de pouvoir comparer les prévalences estimées aux valeurs qu'elles sont censées estimer. Il ne s'agit pour le moment que d'un travail préliminaire.

1.2 Méthode

Une méthode de correction des données de prévalence issues des sites sentinelles en CPN afin d'estimer la prévalence en population générale en tenant compte de cette moindre fécondité des femmes VIH+ a été proposée et appliquée au district de Gulu au nord de l'Ouganda (FABIANI 2001). Fabiani *et al.* ont procédé, en 2003, à une évaluation de cette méthode sur huit séries de données portant sur six sites[2] dans trois pays (Ouganda, Tanzanie et Zambie) où l'on dispose à la fois des données de prévalence en population générale et des données issues des centres locaux de suivi prénatal (FABIANI 2003).

Si l'on suppose que l'ensemble des femmes enceintes consultent en suivi prénatal, ou bien que le taux de consultation ne varie pas selon le statut sérologique à VIH, alors la prévalence $P_{CPN}^{x,x+a}$ observée à un âge donné en clinique prénatale est égale au nombre de femmes séropositives enceintes sur le nombre total de femmes enceintes. Le nombre de femmes enceintes séropositives correspond au nombre de femmes séropositives (soit $n^{x,x+a} \cdot P_{PFG}^{x,x+a}$ où $n^{x,x+a}$ est le nombre total de femmes âgées de *x* à *x*+*a* ans exacts dans la zone et $P_{PFG}^{x,x+a}$ la prévalence observée dans ce même groupe d'âges en population féminine générale) multiplié par $ENC_{VIH+}^{x,x+a}$ la proportion de femmes VIH+, toujours du même groupe d'âges, enceintes sur la période. De même, le nombre de femmes VIH-, âgées de *x* à *x*+*a* ans, enceintes, vaut $n^{x,x+a} \cdot \left(1 - P_{PFG}^{x,x+a}\right) \cdot ENC_{VIH-}^{x,x+a}$. On obtient alors l'Équation annexe 1.1 qui, une fois résolue, donne l'Équation annexe 1.2. La correction étant réalisée pour chaque classe d'âges, un dernier ajustement est nécessaire selon les $F_{PFG}^{x,x+a}$, c'est-à-dire la proportion de la classe d'âges en population féminine générale pour obtenir P_{PFG}, la prévalence en population féminine générale pour l'ensemble des femmes adultes (Équation annexe 1.3) (BOISSON 1996, FABIANI 2001).

Équation annexe 1.1

$$P_{CPN}^{x,x+a} = \frac{n^{x,x+a} \cdot P_{PFG}^{x,x+a} \cdot ENC_{VIH+}}{n^{x,x+a} \cdot P_{PFG}^{x,x+a} \cdot ENC_{VIH+}^{x,x+a} + n^{x,x+a} \cdot \left(1 - P_{PFG}^{x,x+a}\right) \cdot ENC_{VIH-}^{x,x+a}}$$

[2] Les données utilisées par Fabiani *et al.* proviennent d'enquêtes lourdes sur de petites zones d'études. Outre la mise en place d'un dépistage dans les cliniques prénatales de la zone d'étude, ces enquêtes comportent un recensement des populations locales et des passages plus ou moins répétés au cours desquels une sérologie VIH est effectuée, permettant la mesure de la prévalence en population générale. Pour la présente analyse, les données de Fabiani *et al.* ont été reprises et complétées par d'autres mesures issues d'enquêtes du même type.

Équation annexe 1.2

$$P_{PFG}^{x,x+a} = \frac{P_{CPN}^{x,x+a}}{\frac{ENC_{VIH+}^{x,x+a}}{ENC_{VIH-}^{x,x+a}} - P_{CPN}^{x,x+a} \cdot \frac{ENC_{VIH+}^{x,x+a}}{ENC_{VIH-}^{x,x+a}} + P_{CPN}^{x,x+a}}$$

Équation annexe 1.3

$$P_{PFG} = \sum_{classe\ d'âge} F_{PFG}^{x,x+a} \cdot P_{PFG}^{x,x+a} = \sum_{classe\ d'âge} F_{PFG}^{x,x+a} \cdot \left(\frac{P_{CPN}^{x,x+a}}{\frac{ENC_{VIH+}^{x,x+a}}{ENC_{VIH-}^{x,x+a}} - P_{CPN}^{x,x+a} \cdot \frac{ENC_{VIH+}^{x,x+a}}{ENC_{VIH-}^{x,x+a}} + P_{CPN}^{x,x+a}} \right)$$

Pour appliquer cette méthode, il est nécessaire de disposer d'une estimation du paramètre ENC_{VIH+}/ENC_{VIH-} pour chaque classe d'âges. Ce paramètre correspond à un indicateur nommé Odds Relatif de l'Infection (ORI) ou Relative Odds of Infection (ROI) en anglais (voir Équation annexe 1.4).

Équation annexe 1.4

$$ORI = \frac{femmes\ enceintes\ VIH+ \Big/ femmes\ enceintes\ VIH-}{ensemble\ des\ femmes\ VIH+ \Big/ ensemble\ des\ femmes\ VIH-}$$

pour chaque groupe d'âges, ce qui s'écrit également

$$ORI = \frac{femmes\ enceintes\ VIH+ \Big/ ensemble\ des\ femmes\ VIH+}{femmes\ enceintes\ VIH- \Big/ ensemble\ des\ femmes\ VIH-}$$

Le ORI est considéré comme approximativement équivalent au RRF (Ratio du Risque de Fécondité) ou FRR (Fertility Risk Ratio), voir Équation annexe 1.5, dans la mesure où les différences de mesure de la survenue d'une grossesse entre femmes VIH+ et VIH- seraient insignifiantes (ZABA 1998, LEWIS 2004).

Équation annexe 1.5

$$RRF = \frac{naissances\ de\ mères\ VIH+ \Big/ ensemble\ des\ femmes\ VIH+}{naissances\ de\ mères\ VIH- \Big/ ensemble\ des\ femmes\ VIH-}$$

Cependant, en raison d'un nombre plus élevé de mortalité intra-utérine tardive chez les femmes VIH+, le RRF est *a priori* légèrement plus faible que le ORI (LEWIS 2004). Ainsi, avec les RRF, on surestime la sous-fécondité des femmes VIH+ et donc, la correction apportée aux données issues des CPN serait trop élevée.

FABIANI *et al.* (2001, 2003) ont eu recours à des valeurs standards de type RRF calculées à partir d'une étude de CARPENTER *et al.* (1997) portant sur le district de Masaka, Ouganda, sur la période 1989-1996.

1.3 Données

Pour mesurer l'efficacité de cette méthode d'ajustement, nous l'avons appliquée à douze séries de mesures portant sur neuf sites répartis dans sept pays[3]. Pour chacun de ces sites, nous disposons à la fois de la prévalence observée en CPN par groupe d'âges et de la prévalence par âge pour l'ensemble des femmes de la même zone. Il s'agit de grosses enquêtes conçues pour que ce type de données soit comparable car portant sur les mêmes zones géographiques. Sur l'ensemble de ces sites, le taux de fréquentation des CPN pendant la grossesse est élevé, réduisant ainsi les biais qui seraient dus à une fréquentation différentielle selon le statut sérologique VIH.

Outre les coefficients utilisés par FABIANI *et al.* (2001, 2003) (nommés coefficients A), nous avons calculé d'autres séries de coefficients (voir Tableau annexe 1.1). Tout d'abord, en reprenant les données de CARPENTER *et al.* (1997), il apparaît que les coefficients A ont été calculés à partir d'un tableau donnant la prévalence par âge des femmes ayant eu une grossesse pendant la période d'observation et de celles n'en ayant pas eue. Cependant, il semble que les naissances comptabilisées dans ce tableau ne soient que celles enregistrées localement. Or, comme il y a des migrations dans la zone d'étude et comme ces migrations sont plus importantes parmi les femmes VIH+ (la migration étant un facteur de risque de contamination par l'infection à VIH), les RRF calculés sont *a priori* sous-estimés, les naissances non comptabilisées étant plus importantes pour les femmes VIH+. Nous avons alors rapporté les naissances observées non pas au nombre de femmes mais au nombre de femmes-années, pour ainsi tenir compte des migrations. Nous obtenons alors la série des coefficients B.

[3] Nous disposions également de trois séries de données sur Kagera, en milieu urbain, en Tanzanie : KWESIGABO G., KILLEWO J. Z. *et. al.*, « Monitoring of HIV-1 infection prevalence and trends in the general population using pregnant women as a sentinel population: 9 years experience from the Kagera region of Tanzania », *Journal of Acquired Immune Deficiency Syndromes*, n°23(5), 2000. Cependant, nous ne disposions d'une information que par groupe d'âges décennaux. Or, si cela ne constitue qu'un biais mineur pour les femmes âgées, aux jeunes âges il est nécessaire de disposer de données au minimum par groupes quinquennaux, les coefficients d'ajustement variant très sensiblement entre les 15-19 ans et les 20-24 ans. D'autre part, ces groupes d'âges sont également les plus représentés dans la structure par âge et donc ceux dont les biais auront le plus d'impact sur l'estimation finale. Pour ces raisons, nous avons préféré écarter ces trois séries de l'étude.

Tableau annexe 1.1

Coefficients d'ajustement selon différentes sources

Séries	A	B	C	D	*(B+C)/2*
Âges					
15-19 ans	1,350	1,535	0,750	1,372	1,142
20-24 ans	0,629	0,737	0,480	0,541	0,609
25-29 ans	0,621	0,718	0,810	0,524	0,764
30-34 ans	0,355	0,426	0,680	0,324	0,553
35-39 ans	0,569	0,691	0,680	0,294	0,685
Type	RRF	RRF	ORI	RRF	
Zone	Rural, Masaka, Uganda 1990-96	Rural, Masaka, Uganda 1990-96	Rural, Rakai, Uganda 1989-1992	Rural, Kisesa, Tanzanie 1994-1997	
Source	(CARPENTER 1997)	(CARPENTER 1997)	(GRAY 1998)	(HUNTER 2003)	

RRF : Ratio du Risque de Fécondité.
ORI : Odds Relatif de l'Infection.
La série A est issue des calculs effectués par Fabiani (2001, 2003) à partir des données de Carpenter (1997).
La série B est issue de nos propres calculs, à partir de la même source, en tenant compte des migrations.
Les séries C et D ont également été calculées par nos soins à partir de Gray 1998 et de Hunter 2003.
La série (B+C)/2 est obtenue en faisant la moyenne arithmétique des séries B et C pour chaque groupe d'âges.

Une étude de GRAY *et al.* (GRAY 1998) portant sur Rakai, Ouganda (1989-1992), fournit les taux de femmes enceintes par âge et statut sérologique VIH. Nous avons donc pu calculer des coefficients de type ORI (série C). Cependant, cette étude ne porte que sur les femmes sexuellement actives. Ainsi, aux jeunes âges, la part des femmes n'ayant pas entamé leur vie sexuelle étant non négligeable, les coefficients sont sous-estimés. Aux âges plus élevés, ce biais devient négligeable (la quasi-totalité des femmes ayant déjà eu des rapports sexuels) et les coefficients s'avèrent être plus justes car de type ORI (contrairement aux autres coefficients qui sont de type RRF).

Les coefficients C sont sous-estimés aux jeunes âges. Dans le même temps, les coefficients B, étant de type RRF, sont légèrement surestimés. Nous pouvons donc envisager qu'une série de coefficients situés entre ces deux mesures pourrait s'avérer meilleure. Plusieurs hypothèses peuvent être posées pour calculer une série intermédiaire. En l'absence d'éléments pour en définir une, nous avons calculé une série de coefficients de la manière la plus simple, c'est-à-dire en prenant la moyenne des coefficients B et C. Cette série de coefficients est notée (B+C)/2. Nous avons procédé aux calculs pour cette série à titre indicatif.

Enfin, nous avons eu recours à une dernière étude de HUNTER *et al.* (HUNTER 2003) sur la période 1994-1997 à Kisesa, Tanzanie pour calculer la série D. Cette étude a la particularité de comparer la fécondité des femmes non infectées aux deux passages de l'enquête (1994-95 et 1996-97) aux femmes infectées ou au statut inconnu à la première et infectées à la seconde. Ne sont donc pas prises en compte les femmes pour lesquelles on a observé une séroconversion entre les deux passages. De plus, parmi les femmes VIH- au premier passage et non testées au

second, il est probable qu'il y ait une surreprésentation des femmes ayant pris des risques pendant la période entre les deux passages. Or, il a été montré, y compris dans cette étude, que la sous-fécondité des femmes VIH+ augmentait avec la durée de l'infection. Ainsi, les coefficients calculés sont sous-estimés puisqu'on ne prend en compte que des femmes contaminées depuis plusieurs années.

À partir des travaux de GRAY *et al.* (GRAY 1998), nous ne pouvons calculer de coefficients au-delà de 40 ans. D'autre part, sur les douze séries de mesure utilisées, nous ne disposons pas toujours des données de prévalence au-delà de 40 ans. Nous avons donc fait le choix de limiter ce travail aux 15-39 ans[4].

1.4 Résultats

Si l'on regarde l'ensemble des résultats, il s'avère que l'ordre des prévalences estimées est très souvent le même. Ainsi, une prévalence estimée avec les coefficients D sera toujours plus élevée qu'une prévalence estimée avec les B. Si le choix des coefficients prime sur les différences de structure par âge, c'est que l'ensemble des sites possède des structures par âge, que ce soit en PFG ou en CPN, de grandeur relativement équivalente. Il en résulte que le principal facteur qui détermine quel ajustement sera le plus efficace[5] est l'écart initial entre la prévalence observée en CPN ajustée selon la structure par âge de la PFG et la prévalence réelle en population féminine générale.

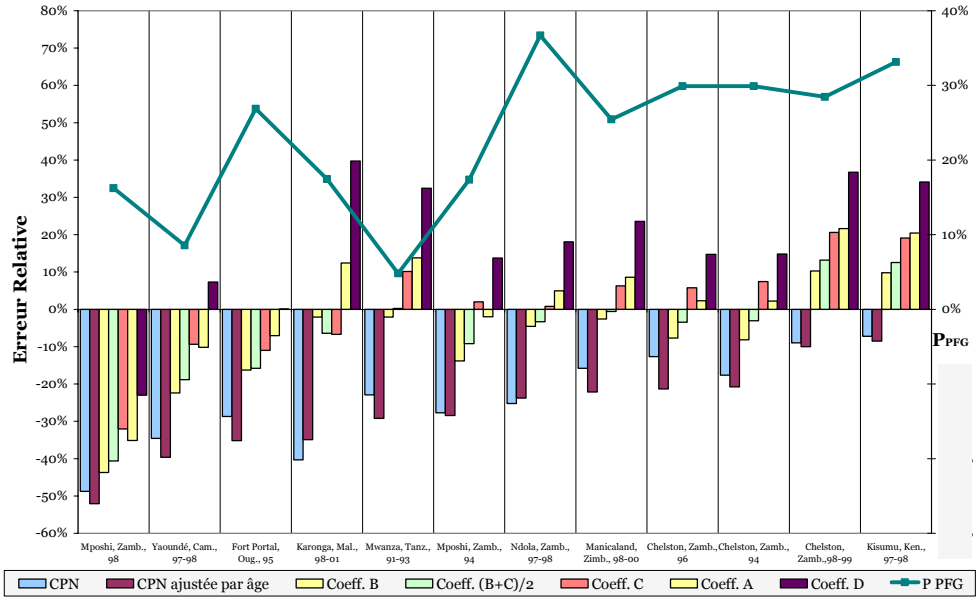
Sur la Figure annexe 1.3, pour les trois premières séries de données, cet écart est très important. Ce sont donc les coefficients D qui s'avèrent être les plus efficaces. Pour les deux dernières, cet écart est très faible et un simple ajustement par âge s'avère alors être plus efficace que les ajustements à partir des coefficients de sous-fécondité.

[4] Cependant, étant donné que la majorité des femmes (du fait de la pyramide des âges de ces pays) ont moins de 40 ans, ceci ne constitue qu'un biais relativement faible pour évaluer l'efficacité de cette méthode d'ajustement. Les résultats obtenus sur les 15-49 ans, pour les coefficients et les sites pour lesquels le calcul est possible, ne diffèrent quasiment pas des résultats présents.

[5] Nous utiliserons le terme d'*efficace* pour désigner un ajustement permettant d'obtenir une prévalence estimée plus proche de P_{PFG} que ne l'est P_{CPN}. On peut alors mesurer l'efficacité par l'erreur absolue (à savoir *prévalence estimée – prévalence PFG*) ou par l'erreur relative ([*prévalence estimée – prévalence PFG*] / *prévalence PFG*). Lorsque la comparaison porte sur un même site, les résultats sont identiques que l'on utilise l'erreur absolue ou l'erreur relative. Entre deux ajustements, celui ayant l'erreur la plus faible (en valeur absolue) sera considéré comme étant le plus efficace.

Figure annexe 1.3

Erreur relative des différents ajustements et prévalence en population féminine générale sur douze séries de données



Pour l'ensemble des autres séries, nous observons, d'une part, que le meilleur ajustement est obtenu avec la série des coefficients B ou C (et dans deux cas la série A) et, d'autre part, que les ajustements obtenus avec A, B et C sont systématiquement meilleurs que la prévalence observée en CPN brute ou ajustée par âge. Enfin, le plus souvent un simple ajustement par âge est encore moins efficace que la prévalence brute observée en CPN.

On retrouve ces résultats lorsque l'on regarde la moyenne sur les douze séries de la valeur absolue de l'erreur relative[6] (Tableau annexe 1.2). En moyenne, l'ajustement par âge s'avère moins efficace que la prévalence brute. Les coefficients C sont les meilleurs, avec une erreur moyenne de 10,9 %, suivi de près par les coefficients A. Nous avons, par ailleurs, testé une série de coefficients obtenus en prenant pour chaque classe d'âges la moyenne arithmétique des coefficients B et C. Cette série notée (B+C)/2 s'avère être en moyenne encore plus efficace que B et C.

[6] Pour le calcul d'un indicateur global d'efficacité sur les douze sites, nous avons préféré utiliser l'erreur relative plutôt que l'erreur absolue afin de ne pas donner plus de poids aux sites ayant une prévalence plus élevée, dans la mesure où le niveau de prévalence n'est pas corrélé avec l'écart relatif initial. D'autre part, nous avons utilisé la moyenne arithmétique des valeurs absolues des erreurs relatives afin d'avoir un indicateur aisément compréhensible. Ainsi, le 12,0 % obtenu pour les coefficients B signifie qu'en ajustant les prévalences CPN avec ces coefficients, on obtient une prévalence estimée qui diffère, en moyenne sur ces douze sites, de 12 % avec la prévalence réellement observée en PFG.

Tableau annexe 1.2

Prévalences observées et estimées, erreurs relatives et moyenne des valeurs absolues des erreurs relatives pour chaque série de coefficients

Site	Prévalence du VIH (%)								Erreur Relative (ER en %)						Effectifs	
	PFG	CPN	CPN aj.	A	B	C	D	(B+C)/2	CPN	CPN aj.	A	B	C	D	(B+C)/2	PFG	CPN
Rural, Mposhi, Zambie, 1998 (Fylkesnes 2001)	16	8,3	7,8	11	9,1	11	13	9,6	-48,7	-52	-35,1	-43,7	-32	-23	-40,6	425	300
Urbain, Yaoundé, Cameroun, 1997-98 (Glynn 2001)	8,6	5,6	5,2	7,7	6,6	7,8	9,2	7	-34,6	-39,6	-10,2	-22,4	-9,3	7,3	-18,8	829	1 525
Urbain, Fort Portal, Ouganda, 1995 (Kilian 1999)	27	19	17	25	23	24	27	23	-28,7	-35,2	-7,1	-16,2	-11	0,1	-15,8	470	458
Rural, Karonga, Malawi, 1998-01 (Crampin 2003)	18	10	11	20	17	16	24	16	-40,3	-34,9	12,4	-2,1	-6,7	39,7	-6,4	287	2 908
Rural, Mwanza, Tanzanie, 1991-93 (Changalucha 2002)	4,8	3,7	3,4	5,5	4,7	5,3	6,4	4,8	-22,9	-29,2	13,8	-2	10,1	32,4	0,2	5 089	2 153
Rural, Mposhi, Zambie, 1994 (Fylkesnes 2001, Fylkesnes 1998)	17	13	12	17	15	18	20	16	-27,7	-28,4	-2	-13,8	2	13,7	-9,2	426	422
Urbain, Ndola, Zambie, 1997-98 (Glynn 2001)	37	27	28	39	35	37	43	36	-25,2	-23,8	5	-4,6	0,8	18,1	-3,3	730	1 002
Rural, Manicaland, Zimbabwe, 1998-2000 (Gregson 2002)	25	21	20	28	25	27	31	25	-15,8	-22,1	8,6	-2,6	6,3	23,6	-0,6	4 452	1 162
Urbain, Chelston, Zambie, 1996 (Fylkesnes 2001)	30	26	24	31	28	32	34	29	-12,6	-21,4	2,3	-7,7	5,8	14,7	-3,4	1 211	532
Urbain, Chelston, Zambie, 1994 (Fylkesnes 2001, Fylkesnes 1998)	30	25	24	31	27	32	34	29	-17,6	-20,8	2,2	-8,2	7,5	14,8	-3,1	1 211	443
Urbain, Chelston, Zambie, 1998-99 (Fylkesnes 2001)	29	26	26	35	31	34	39	32	-9	-10	21,7	10,3	20,6	36,7	13,2	1 206	776
Urbain, Kisumu, Kenya, 1997-98 (Glynn 2001)	33	31	30	40	36	40	44	37	-7,2	-8,5	20,5	9,8	19,1	34,1	12,5	739	1 447
Moyenne des valeurs absolues des ER*									24,20	27,20	11,70	12,00	10,90	21,50	10,60		

PFG : Population Féminine Générale – CPN : femmes consultant en Clinique Prénatale – CPN aj. : prévalence observée en CPN ajustée par la structure par âge en PFG.
A, B, C et D : prévalence en population générale estimée à partir des coefficients A, B, C ou D. ER : Erreur relative = (prévalence estimée – prévalence PFG) / prévalence PFG.
* Lecture : L'écart moyen des prévalences observées en CPN par rapport à celles en PFG est de 24,2 %. L'écart des prévalences estimées avec la série de coefficients A par rapport à la prévalence observée en PFG est en moyenne de 11,7 %.

Les coefficients D, par contre, s'avèrent en moyenne encore moins efficace qu'un ajustement par âge. Nous avons vu précédemment que ces coefficients surestimaient la sous-fécondité des femmes VIH+, ce qui induit qu'en les utilisant on surestime de manière importante la prévalence en population générale.

Comme l'indique la Figure annexe 1.3, il semblerait que le niveau de l'épidémie de VIH ne joue pas sur l'efficacité de telle ou telle série de coefficients d'ajustement.

On retiendra que la méthode d'ajustement par les coefficients A, B, C ou (B+C)/2 permet d'obtenir une prévalence en population féminine générale estimée plus proche du réel pour dix séries de prévalence sur douze.

1.5 Discussion

Les deux séries pour lesquelles la méthode s'avère inefficace (Chelston, Zambie 1998-99 et Kisumu Kenya 1997-98) ont des niveaux d'utilisation de méthodes contraceptives non négligeables. Or, la contraception constitue une source importante de biais. Il a déjà été montré que les méthodes d'ajustement étaient moins efficaces dans les zones où les pratiques contraceptives étaient courantes (ZABA 2000, GREGSON 2002, FABIANI 2003). Les interférences de la contraception sur les données CPN sont mal connues. Si par endroit il a été observé que les femmes VIH+ utilisaient plus souvent une méthode contraceptive (RYDER 1991), le type de contraception utilisé peut également intervenir. Un usage important du préservatif aura un effet à la fois de protection contre le VIH et de contraception, à condition qu'il soit utilisé notamment par les femmes VIH-. Il est donc primordial d'étudier et de clarifier la part de la contraception sur la sous-fécondité des femmes VIH+ et sur les différentiels de fréquentation des CPN.

Les résultats des estimations ont été comparés à la prévalence observée en population générale mesurée par enquête. Cependant, cette prévalence est également sujette à des biais et en particulier à la variation importante du taux de non réponses par âge (CARPENTER 1997). Or, les personnes refusant de se faire tester seraient plus souvent séropositives que les autres (BUVE 2001, GLYNN 2001). Ces variations différentielles des taux de non réponses induisent des biais à la fois pour établir les coefficients correcteurs et pour déterminer les prévalences à estimer. Il est très difficile de pouvoir déterminer dans nos résultats la part de la sous-fécondité des femmes VIH+ et celle des biais.

Néanmoins, il apparaît que cette méthode d'ajustement, consistant à corriger la prévalence de chaque groupe d'âges en tenant compte de la fécondité différentielle des femmes VIH+, est efficace, à condition de disposer de données par âge au minimum quinquennal, de se situer dans une zone où l'épidémie est généralisée et où la contraception est faible. Le problème majeur reste le choix des coefficients d'ajustement. Il serait nécessaire d'affiner ceux-ci à partir de données plus importantes et issues d'un plus grand nombre de sites. De même, il faudrait pouvoir tester cette méthode sur d'autres zones d'Afrique pour voir si elle y est également efficace.

Il reste que cette méthode offre une piste pour l'organisation du suivi épidémiologique dans les pays à épidémie généralisée. En effet, dans la majorité des pays, des enquêtes nationales de séroprévalence commencent à se mettre en place. Celles-ci pourraient éventuellement permettre de calculer des coefficients d'ajustement propres au pays, afin d'appliquer ces coefficients sur les données prénatales entre deux enquêtes nationales. Cela présuppose néanmoins que les données en CPN soient de bonne qualité et que les sites sentinelles choisis soient bien représentatifs du pays. Une piste de recherche consisterait donc à valider cette méthode à un niveau national, en effectuant des comparaisons entre les données sentinelles d'une année donnée et les données nationales en population générale de la même année. Il serait alors possible de mettre en place un suivi sentinelle plus efficace à moindre coût[7].

[7] Néanmoins, cela supposerait de régler le problème de représentativité des sites sentinelles retenus, problème non résolu par cette méthode et qui se manifeste, entre autres, dans les écarts observés dans certains pays entre les estimations faites à partir de données CPN et les récentes mesures réalisées lors d'enquêtes en population générale telles que les enquêtes démographiques et de santé (EDS).

Annexe 2
Estimation de la proportion de personnes infectées situées dans la fenêtre sérologique

2.1 Modèle simple

Suite aux recommandations du Groupe de Référence de l'ONUSIDA (THE UNAIDS REFERENCE GROUP ON ESTIMATES MODELLING AND PROJECTIONS 2002), nous modélisons la courbe de survie, *S(d)*, en fonction de la durée d'infection *d* par une courbe de Weibull. Ce type de courbe prend en compte deux paramètres, *k* et λ. La courbe de survie sera donc de la forme :

$$S(d) = e^{-(d/\lambda)^k}$$

Dans la version 2007 d'EPP[8], le logiciel d'estimation d'ONUSIDA, la valeur par défaut retenue pour le paramètre *k* est 2. Nous appelons *med* la durée médiane de survie. Nous avons alors la relation suivante :

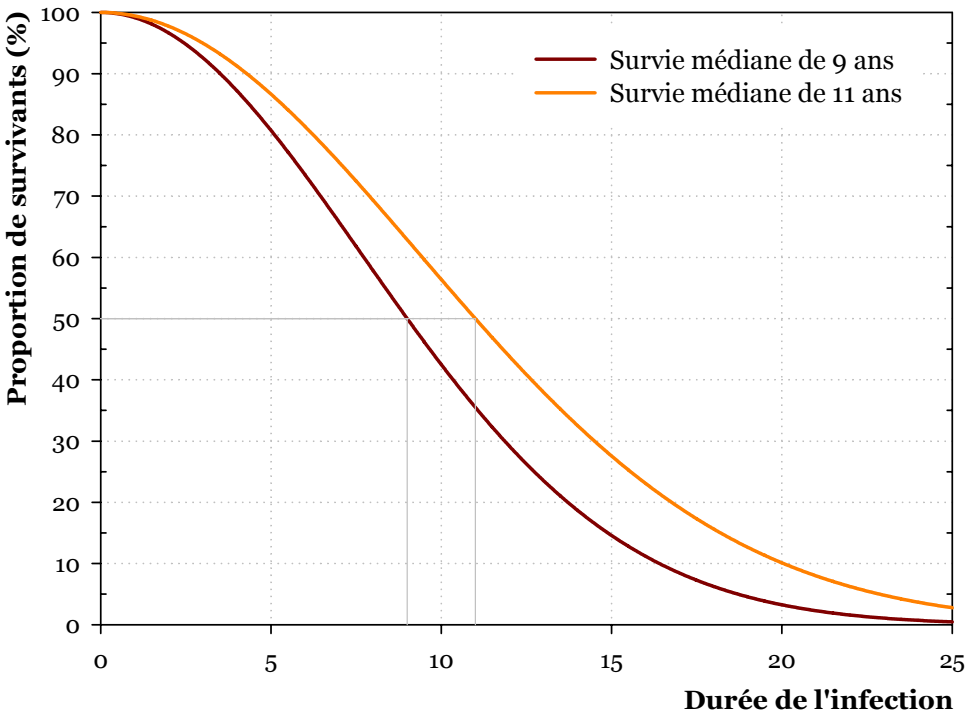
[8] *EPP software release 9*, de juin 2007, disponible sur le site d'ONUSIDA : http://www.unaids.org/en/HIV_data/Epidemiology/epi_software2007.asp.

$S(med) = e^{-(med/\lambda)^2} = \frac{1}{2} \Rightarrow \lambda = \frac{med}{\sqrt{\ln(2)}}$ (NB : med>0 par définition) d'où $S(d) = 2^{-\frac{d^2}{med^2}}$

En 2002, le Groupe de Référence recommandait de prendre une durée de vie médiane de 9 ans (THE UNAIDS REFERENCE GROUP ON ESTIMATES MODELLING AND PROJECTIONS 2002) puis, plus récemment, une valeur de 11 ans (THE UNAIDS REFERENCE GROUP ON ESTIMATES MODELLING AND PROJECTIONS 2006). Nous obtenons, pour ces deux hypothèses, les courbes de survie suivantes :

Figure annexe 2.1

Courbes de survie en fonction de la durée d'infection (en années) selon deux hypothèses de durée médiane de survie



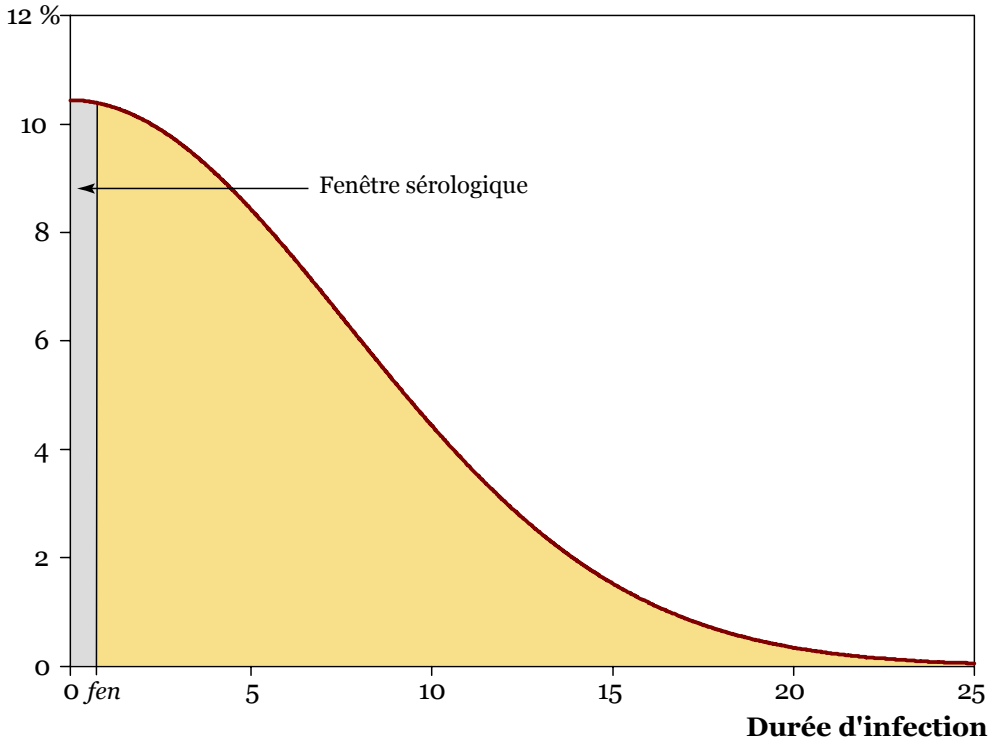
Sous l'hypothèse d'une incidence constante, la distribution des personnes infectées en fonction de la durée d'infection est égale à la répartition des décès, moyennant un facteur d'échelle afin que la superficie de la surface située sous la courbe soit égale à 1. Il suffit pour cela de diviser $S(d)$ par son intégrale entre 0 et l'infini, soit :

$$\int_0^{+\infty} 2^{-\frac{d^2}{med^2}} \partial d = \frac{1}{2} med \sqrt{\frac{\pi}{\ln 2}}$$

Nous obtenons ainsi la fonction de distribution *D(d)* des personnes infectées en fonction de la durée d'infection.

$$D(d) = \frac{2^{-\frac{d^2}{med^2}}}{\frac{1}{2} med \sqrt{\frac{\pi}{\ln 2}}}$$

Figure annexe 2.2
Fenêtre sérologique et distribution des personnes infectées en fonction de la durée d'infection (en années) sous l'hypothèse d'une incidence constante.



La proportion de personnes infectées situées dans la fenêtre sérologique correspond alors à la surface grise représentée sur le graphique. Si nous notons *fen* la durée de la fenêtre sérologique, la proportion p_{nobs} de personnes non observées en raison de leur présence dans la fenêtre sérologique correspond à :

$$p_{nobs} = \int_0^{fen} D(d) \partial d$$

Pour le calcul du risque résiduel dans le domaine du don de sang, la majorité des études utilisent la valeur de 22 jours (BUSCH 1995, SCHREIBER 1996, BUSCH 2000), avec des marges d'incertitude entre 6 et 38 jours, à la fois dans des pays occidentaux comme la France (PILLONEL 2002), l'Italie (VELATI 2002), l'Espagne (ALVAREZ 2002), l'Australie ou les États-Unis (GLYNN 2002), mais également dans

des pays africains comme la Guinée (LOUA 2004), la Côte d'Ivoire (OUATTARA 2006) ou l'Afrique du Sud (FANG 2003), ou bien encore en Chine (SHANG 2007). Concernant les tests de recherche d'anticorps anti-VIH de quatrième génération (incluant la recherche de l'antigène p24), la fenêtre sérologique est estimée à 17 jours (BUSCH 2005).

Nous retiendrons donc comme hypothèse centrale de durée de la fenêtre sérologique la valeur de 22 jours. L'hypothèse haute portera sur 38 jours et l'hypothèse basse sur 17 jours. Les valeurs numériques des intégrales ont été calculées à l'aide de la version 5.2.0.0 du logiciel Mathematica développé par la société Wolfram Research.

Tableau annexe 2.1
Proportion de personnes infectées situées dans la fenêtre sérologique selon plusieurs hypothèses de survie et de durée de la fenêtre sérologique

Fenêtre sérologique	Durée médiane de survie	
	9 ans	11 ans
17 jours	0,49 %	0,40 %
22 jours	0,63 %	0,51 %
38 jours	1,09 %	0,89 %

Ce modèle simple présuppose, pour être vérifié, une incidence constante depuis au moins 25 ans. Or, les différentes épidémies africaines ont débuté vraisemblablement entre 1979 et 1982. De ce fait, le temps que l'épidémie se développe, les nouvelles infections ont été plus nombreuses au cours des dix dernières années que pendant la décennie 1980. Le modèle simple représente donc une estimation *a minima* de la proportion de personnes infectées non observables car situées dans la fenêtre sérologique, correspondant à la valeur vers laquelle une épidémie stabilisée devrait tendre.

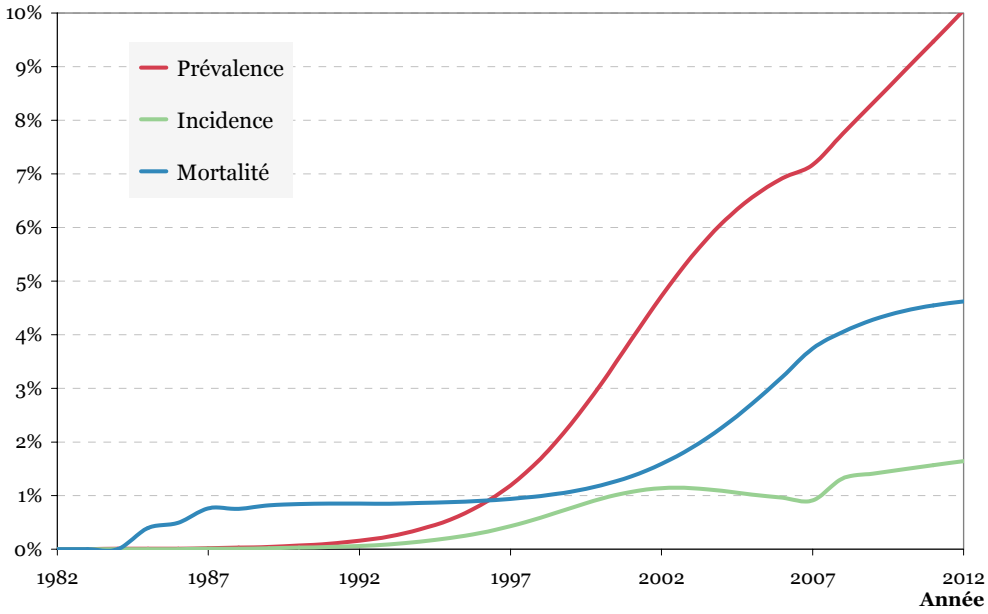
Au démarrage d'une épidémie, cette proportion devrait *a priori* diminuer dans la mesure où se constitue une population de personnes anciennement infectées. Par la suite, une fois l'épidémie développée et la première génération de personnes infectées presque totalement décédées, à mortalité constante, une incidence croissante devrait induire une augmentation de la proportion non observable et inversement. À incidence constante, une diminution de la mortalité (ce qui correspond à une augmentation de la durée médiane de survie) induit une réduction de cette proportion.

2.2 Modèle Spectrum

Pour estimer l'évolution de la proportion de non observables au cours du développement d'une épidémie, nous avons utilisé la population et l'épidémie fictive correspondant à la projection exemple fournie avec le logiciel de projection démographique et épidémiologique Spectrum, version 3.13 Beta 6 d'août 2007, développé par Futures Institute, sur financement USAID et téléchargeable sur http://www.unaids.org.

Cette épidémie exemple correspond aux caractéristiques classiques d'une épidémie généralisée et les paramètres utilisés pour la projection sont les paramètres standards pour les pays d'Afrique subsaharienne, selon les recommandations du Groupe de Référence d'ONUSIDA sur les estimations, la modélisation et les projections. C'est une épidémie qui a démarré en 1981 et dont la prévalence du VIH a augmenté continuellement pour atteindre 10 % à l'horizon 2012 (voir Figure annexe 2.3).

Figure annexe 2.3
Prévalence, incidence et mortalité liée au VIH des 15-49 ans, de 1982 à 2012, de la projection exemple du logiciel Spectrum



Pour chaque année, le modèle fournit le nombre de personnes infectées par le VIH ainsi que le nombre de nouvelles infections. Nous pouvons alors calculer la proportion d'infections de moins de un an parmi les personnes infectées. En supposant que les nouvelles infections se répartissent uniformément au cours de l'année écoulée, la proportion de personnes infectées non observables s'obtient en

multipliant la proportion des infections de moins de un par la durée de la fenêtre sérologique exprimée en année, soit 22 jours / 365.25 = 6,023 %.

Les calculs ont été effectués pour l'ensemble de la population adulte, soit les 15-49 ans, ainsi que pour les groupes d'âges 15-19 ans et 20-24 ans (voir Tableau annexe 2.2, Tableau annexe 2.3 et Tableau annexe 2.4).

Tableau annexe 2.2

Nouvelles infections, personnes infectées et proportion de personnes non observables dans la projection exemple de Spectrum pour les 15-49 ans

Année	Nouvelles infections annuelles (1)	Nombre de personnes VIH+ (2)	Proportion de nouvelles infections (1)/(2)	Proportion de personnes non observables
1982	61	61	100,00%	6,02%
1983	37	98	37,76%	2,27%
1984	60	157	38,22%	2,30%
1985	97	252	38,49%	2,32%
1986	157	406	38,67%	2,33%
1987	254	657	38,66%	2,33%
1988	412	1 063	38,76%	2,33%
1989	665	1 712	38,84%	2,34%
1990	1 062	2 734	38,84%	2,34%
1991	1 698	4 346	39,07%	2,35%
1992	2 651	6 830	38,81%	2,34%
1993	4 092	10 633	38,48%	2,32%
1994	6 259	16 449	38,05%	2,29%
1995	9 466	25 281	37,44%	2,25%
1996	14 225	38 605	36,85%	2,22%
1997	20 760	58 079	35,74%	2,15%
1998	29 220	85 462	34,19%	2,06%
1999	39 125	121 891	32,10%	1,93%
2000	49 196	167 159	29,43%	1,77%
2001	57 668	219 241	26,30%	1,58%
2002	62 986	274 505	22,95%	1,38%
2003	64 658	328 713	19,67%	1,18%
2004	63 475	378 481	16,77%	1,01%
2005	61 115	422 149	14,48%	0,87%
2006	59 073	459 645	12,85%	0,77%
2007	57 909	491 733	11,78%	0,71%
2008	86 298	547 861	15,75%	0,95%
2009	94 031	607 215	15,49%	0,93%
2010	102 013	669 922	15,23%	0,92%
2011	110 269	736 242	14,98%	0,90%
2012	118 746	806 315	14,73%	0,89%

Hypothèse : fenêtre sérologique de 22 jours.

Tableau annexe 2.3

Nouvelles infections, personnes infectées et proportion de personnes non observables dans la projection exemple de Spectrum pour les 15-19 ans

Année	Nouvelles infections annuelles (1)	Nombre de personnes VIH+ (2)	Proportion de nouvelles infections (1)/(2)	Proportion de personnes non observables
1982	9	9	100,00%	6,02%
1983	7	14	49,21%	2,96%
1984	11	21	51,71%	3,11%
1985	15	32	47,94%	2,89%
1986	24	48	49,48%	2,98%
1987	37	74	49,85%	3,00%
1988	60	116	51,56%	3,11%
1989	91	178	50,92%	3,07%
1990	134	266	50,39%	3,04%
1991	212	407	52,02%	3,13%
1992	326	623	52,28%	3,15%
1993	488	943	51,70%	3,11%
1994	721	1408	51,22%	3,09%
1995	1 067	2089	51,06%	3,08%
1996	1 494	2965	50,39%	3,03%
1997	2 244	4343	51,68%	3,11%
1998	3 217	6317	50,92%	3,07%
1999	4 407	8978	49,09%	2,96%
2000	5 813	12 387	46,93%	2,83%
2001	7 204	16 145	44,62%	2,69%
2002	8 532	20 167	42,30%	2,55%
2003	9 499	24 060	39,48%	2,38%
2004	10 336	27 682	37,34%	2,25%
2005	12 082	32 043	37,71%	2,27%
2006	11 595	34 788	33,33%	2,01%
2007	13 195	38 304	34,45%	2,07%
2008	14 518	42 227	34,38%	2,07%
2009	17 430	47 923	36,37%	2,19%
2010	18 838	53 623	35,13%	2,12%
2011	20 855	59 783	34,88%	2,10%
2012	22 597	66 245	34,11%	2,05%

Hypothèse : fenêtre sérologique de 22 jours.

Tableau annexe 2.4

Nouvelles infections, personnes infectées et proportion de personnes non observables dans la projection exemple de Spectrum pour les 20-24 ans

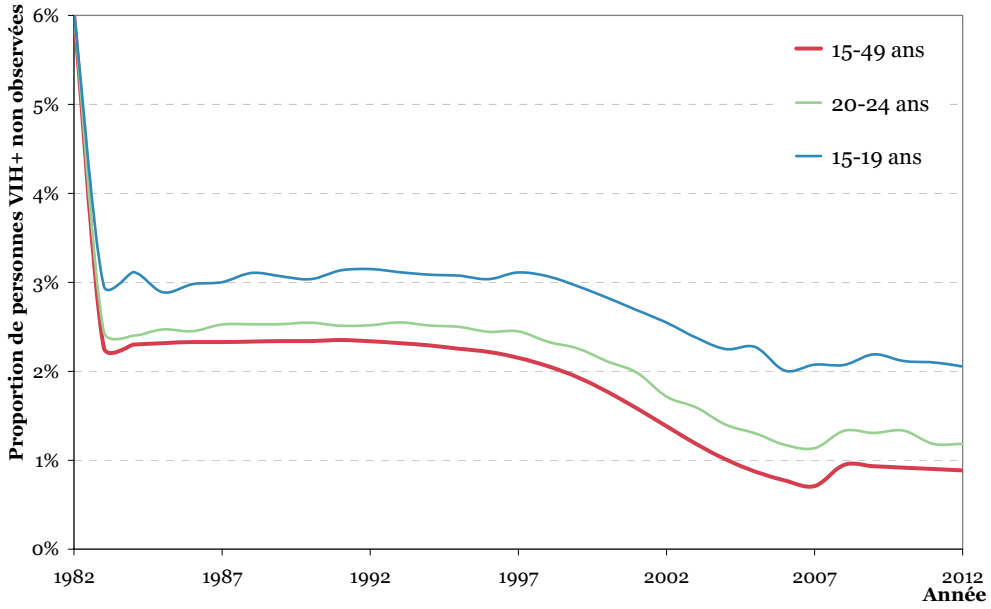
Année	Nouvelles infections annuelles (1)	Nombre de personnes VIH+ (2)	Proportion de nouvelles infections (1)/(2)	Proportion de personnes non observables
1982	9	9	100,00%	6,02%
1983	6	15	40,67%	2,45%
1984	10	25	39,84%	2,40%
1985	16	40	41,03%	2,47%
1986	26	65	40,69%	2,45%
1987	44	106	41,96%	2,53%
1988	72	172	41,98%	2,53%
1989	117	278	42,03%	2,53%
1990	187	442	42,25%	2,54%
1991	286	685	41,71%	2,51%
1992	437	1046	41,80%	2,52%
1993	676	1598	42,32%	2,55%
1994	1 009	2417	41,74%	2,51%
1995	1 519	3661	41,50%	2,50%
1996	2 253	5554	40,56%	2,44%
1997	3 415	8401	40,66%	2,45%
1998	4 754	12278	38,72%	2,33%
1999	6 542	17460	37,47%	2,26%
2000	8 311	23702	35,06%	2,11%
2001	10 115	30681	32,97%	1,99%
2002	10 684	37507	28,48%	1,72%
2003	11 870	44857	26,46%	1,59%
2004	12 040	51753	23,26%	1,40%
2005	12 709	58800	21,61%	1,30%
2006	12 551	64442	19,48%	1,17%
2007	13 211	70020	18,87%	1,14%
2008	17 533	79316	22,10%	1,33%
2009	19 325	89032	21,71%	1,31%
2010	22 235	100378	22,15%	1,33%
2011	21 538	109569	19,66%	1,18%
2012	23 496	119527	19,66%	1,18%

Hypothèse : fenêtre sérologique de 22 jours.

Les résultats de ces trois tableaux sont représentés graphiquement par la figure ci-dessous.

Figure annexe 2.4

Proportion de personnes infectées non observables, sous l'hypothèse d'une fenêtre sérologique de 22 jours, pour trois groupes d'âges, selon la projection exemple du logiciel Spectrum



Pour les groupes d'âges les plus jeunes, 15-19 ans et 20-24 ans, la proportion de personnes infectées non observables est plus importante que dans le reste de la population dans la mesure où les individus n'ont débuté leur sexualité que récemment (voir Tableau annexe 2.5), ce qui modifie la structure par durée d'infection des personnes infectées à ces âges (notamment par l'absence de personnes infectées depuis plus de 5 ou 10 ans).

Tableau annexe 2.5

Âge médian au premier rapport sexuel selon différentes EDS, par groupes d'âges et sexe, en Afrique subsaharienne

	Femmes								Hommes									
EDS pays / année	**20 24**	**25 29**	**30 34**	**35 39**	**40 44**	**45 49**	**25 49**	**20 49**	**25 54**	**25 29**	**30 34**	**35 39**	**40 44**	**45 49**	**50 54**	**55 59**	**60 64**	**65+**
Afrique du Sud 1998	17.8	18.1	18.2	18.5	18.7	18.7	18.4	18.2	-	-	-	-	-	-	-	-	-	-
Bénin 1996	17.2	17.2	17.3	17.3	17.5	17.4	17.3	17.3	18.1	17.2	17.2	18.2	18.4	18.7	18.8	19.6	20.5	-
Bénin 2001	17.2	17.4	17.1	17.4	17.3	17.8	17.3	17.3	17.8	17.2	17.6	17.4	18.3	18.3	19.1	19.9	18.8	-
Botswana 1988	17.4	17.4	17.6	17.5	17.6	18.3	17.6	17.5	-	-	-	-	-	-	-	-	-	-
Burkina Faso 1998/99	17.3	17.5	17.3	17.5	17.6	17.6	17.5	17.4	20.4	20.1	20.3	20.2	20.4	20.6	20.8	22.3	-	-
Burkina Faso 2003	17.5	17.4	17.5	17.5	17.6	17.7	17.5	17.5	20.7	20.4	20.6	20.6	20.9	21.0	21.6	21.3	-	-
Burundi 1987	-	19.2	19.4	19.3	19.5	18.9	19.3	19.6	-	-	-	-	-	-	-	-	-	-
Cameroun 1991	16.2	16.2	15.9	16.0	16.1	15.8	16.0	16.1	-	-	-	-	-	-	-	-	-	-
Cameroun 1998	16.3	15.9	15.8	15.8	15.7	15.6	15.8	15.9	18.2	17.6	17.9	18.2	18.4	18.9	19.0	19.4	-	-
Cameroun 2004	16.7	16.5	16.3	16.2	16.4	16.3	16.4	16.5	18.6	17.9	18.3	18.6	19.0	19.3	19.9	20.2	-	-
Comores 1996	-	19.7	18.2	18.1	17.4	17.7	18.3	18.8	18.3	16.8	17.0	18.3	18.3	20.6	20.4	20.9	25.0	-
Congo 2005	16.2	15.9	15.7	15.8	15.9	15.8	15.8	15.9	16.8	16.2	16.2	17.1	16.7	17.5	18.9	18.2	-	-
Côte d'Ivoire 1994	15.8	15.7	15.8	15.7	15.9	16.0	15.8	15.8	-	-	-	-	-	-	-	-	-	-
Côte d'Ivoire 1998/99	16.2	16.2	15.9	16.3	16.0	16.3	16.1	16.1	18.5	17.3	18.7	18.3	19.5	18.9	20.2	18.9	-	-
Érythrée 1995	17.9	17.7	17.1	16.8	16.4	16.0	16.8	17.0	-	-	-	30.6	-	-	-	-	-	-
Érythrée 2002	18.3	18.3	17.7	18.2	18.1	16.4	17.9	18.0	-	-	-	-	-	-	-	-	-	-
Éthiopie 2000	18.1	17.0	15.8	15.8	15.7	15.8	16.0	16.4	20.3	21.3	19.7	19.1	20.2	20.4	20.7	20.4	-	-
Éthiopie 2005	18.2	16.6	16.4	16.1	15.7	15.7	16.1	16.5	21.2	22.0	21.0	21.3	20.8	20.8	22.0	21.0	-	-
Gabon 2000	16.2	16.3	16.1	16.3	15.8	15.8	16.1	16.2	16.9	15.9	16.6	16.8	17.5	17.6	17.8	18.3	-	-
Ghana 1988	16.8	16.7	16.4	16.6	16.5	16.5	16.5	16.6	-	-	-	-	-	-	-	-	-	-
Ghana 1993	16.9	17.0	16.8	17.0	17.4	17.6	17.1	17.0	19.2	18.6	19.1	20.0	19.2	19.6	20.1	20.6	-	-
Ghana 1998	17.5	18.0	17.7	17.6	17.5	17.5	17.7	17.6	19.3	19.1	18.9	19.0	19.5	19.4	20.0	20.4	-	-
Ghana 2003	18.4	18.3	17.9	18.0	18.2	18.4	18.2	18.2	20.1	19.6	20.1	20.0	20.3	20.4	20.8	21.8	-	-

	Femmes								Hommes									
EDS pays / année	**20 24**	**25 29**	**30 34**	**35 39**	**40 44**	**45 49**	**25 49**	**20 49**	**25 54**	**25 29**	**30 34**	**35 39**	**40 44**	**45 49**	**50 54**	**55 59**	**60 64**	**65+**
Guinée 1999	16.0	15.9	15.9	16.4	15.9	16.2	16.0	16.0	18.9	17.6	18.1	18.9	20.1	20.2	20.7	21.0	-	-
Guinée 2005	16.4	16.0	15.9	15.8	16.1	16.0	16.0	16.0	19.9	18.6	18.9	19.6	19.9	20.6	21.2	22.3	-	-
Kenya 1989	17.1	16.7	16.4	16.4	16.5	16.6	16.5	16.7	-	-	-	-	-	-	-	-	-	-
Kenya 1993	17.4	17.2	16.7	16.5	16.4	16.9	16.8	16.9	16.9	16.4	17.0	16.8	17.3	17.1	18.0	-	-	-
Kenya 1998	17.3	16.8	16.7	16.3	16.2	16.1	16.5	16.7	17.0	16.3	16.6	16.9	17.1	17.7	18.1	-	-	-
Kenya 2003	18.1	18.0	17.7	17.7	17.1	16.9	17.6	17.8	17.2	17.4	16.9	17.1	17.2	17.6	17.8	-	-	-
Lesotho 2004	18.7	18.7	18.8	18.9	18.4	18.4	18.6	18.7	20.0	18.6	19.3	20.1	20.5	20.8	22.1	22.4	-	-
Libéria 1986	15.5	15.5	15.5	15.4	15.2	15.5	15.5	15.5	-	-	-	-	-	-	-	-	-	-
Madagascar 1992	17.0	17.0	16.6	16.6	16.6	16.0	16.7	16.8	-	-	-	-	-	-	-	-	-	-
Madagascar 1997	17.0	17.1	17.1	16.7	16.6	16.8	16.9	16.9	-	-	-	-	-	-	-	-	-	-
Madagascar 2003/2004	17.3	17.4	17.6	17.6	17.4	17.5	17.5	17.4	18.0	17.6	17.6	18.0	18.1	18.4	18.7	18.0	-	-
Malawi 2000	17.1	16.9	16.7	16.8	16.7	16.9	16.8	16.9	18.4	18.2	18.3	18.1	18.5	19.5	19.6	-	-	-
Malawi 2004	17.4	17.5	17.1	17.2	17.1	17.6	17.3	17.3	18.6	18.6	18.4	18.4	18.8	19.0	19.0	-	-	-
Mali 1987	16.1	16.0	15.7	15.7	15.8	15.8	15.8	15.8	-	-	-	-	-	-	-	-	-	-
Mali 1995/96	15.9	15.8	15.8	15.8	15.9	15.7	15.8	15.8	20.6	19.6	19.8	20.3	20.7	22.2	22.5	23.3	-	-
Mali 2001	15.9	15.9	15.8	15.8	15.9	15.9	15.8	15.9	20.2	19.0	20.0	20.2	20.5	20.5	20.8	21.1	-	-
Mauritanie 2000/01	-	19.0	16.8	16.5	15.7	16.0	17.0	17.8	-	-	24.5	25.4	24.3	25.3	25.3	25.3	-	-
Mozambique 1997	16.0	15.9	16.1	15.7	16.3	16.0	15.9	16.0	18.2	18.5	17.8	17.6	17.8	18.9	18.7	20.2	19.4	-
Mozambique 2003	16.0	16.0	16.0	16.1	16.1	16.6	16.1	16.1	-	-	-	-	-	-	-	-	-	-
Namibie 1992	18.6	19.2	18.9	18.9	19.5	20.2	19.2	19.0	-	-	-	-	-	-	-	-	-	-
Namibie 2000	18.2	18.8	18.9	19.6	19.9	20.3	19.2	18.9	18.5	17.8	18.3	18.7	18.9	19.6	20.3	20.1	-	-
Niger 1992	15.2	15.1	15.1	15.2	15.1	15.0	15.1	15.1	-	-	-	-	-	-	-	-	-	-
Niger 1998	-	-	-	-	-	-	-	-	20.4	20.3	20.4	20.2	20.3	20.9	20.5	20.8	-	-
Nigéria 1990	17.0	16.9	16.6	17.2	17.0	17.2	16.9	17.0	-	-	-	-	-	-	-	-	-	-
Nigéria 1999	18.1	18.1	17.3	17.7	17.7	18.1	17.8	17.9	20.1	19.8	19.8	19.8	20.2	20.4	20.9	21.0	20.8	-
Nigéria 2003	17.6	17.3	16.1	15.9	15.6	15.5	16.2	16.7	20.7	20.4	20.3	20.8	20.8	21.8	21.3	25.2	-	-
Ouganda 1988	15.9	15.6	15.6	15.8	15.4	15.3	15.6	15.7	-	-	-	-	-	-	-	-	-	-
Ouganda 1995	16.4	16.0	16.1	15.8	15.9	15.9	16.0	16.1	-	-	-	-	-	-	-	-	-	-

EDS pays / année	Femmes								Hommes									
	20 24	25 29	30 34	35 39	40 44	45 49	25 49	20 49	25 54	25 29	30 34	35 39	40 44	45 49	50 54	55 59	60 64	65+
Ouganda 2000/01	16.7	16.8	16.5	16.7	16.4	16.6	16.6	16.7	18.8	19.4	19.0	18.6	18.8	18.7	18.5	-	-	-
République centrafricaine 1994/95	16.0	15.9	16.0	16.0	15.8	15.9	15.9	15.9	17.8	17.4	17.5	17.7	18.1	18.2	18.4	18.9	-	-
Rwanda 1992	-	20.2	20.0	19.7	19.2	18.5	19.7	19.9	-	-	-	-	-	-	-	-	-	-
Rwanda 2000	20.0	20.3	20.3	19.9	19.9	19.8	20.1	-	-	-	-	-	-	-	-	-	-	-
Rwanda 2005	-	20.0	20.6	20.5	20.1	20.1	20.3	-	20.8	20.6	21.5	21.0	20.6	20.8	19.9	19.8	-	-
Sénégal 1992/93	17.8	16.7	16.2	16.2	15.9	15.9	16.2	16.6	-	-	-	-	-	-	-	-	-	-
Sénégal 1997	19.3	18.3	17.3	17.1	16.8	16.7	17.3	17.8	22.1	21.0	21.1	21.3	22.2	23.8	25.2	25.4	25.7	26.2
Sénégal 2005	19.6	19.3	18.7	18.1	17.7	17.6	18.4	18.7	-	-	-	-	-	-	-	-	-	-
Tanzanie 1992	17.3	17.3	16.4	16.4	16.4	16.6	16.7	16.8	17.5	16.8	17.5	18.1	16.9	18.1	17.5	17.5	18.7	-
Tanzanie 1996	17.4	16.9	16.9	16.4	16.5	16.6	16.7	16.9	18.0	17.6	18.1	18.2	18.1	18.1	18.4	19.2	-	-
Tanzanie 1999	16.9	16.8	16.9	16.3	16.0	16.2	16.6	16.7	17.8	17.6	17.6	17.7	17.9	18.1	18.4	18.1	-	-
Tanzanie 2004	17.1	17.3	17.0	17.0	17.1	16.6	17.0	17.0	18.5	18.2	18.4	18.7	18.7	18.8	-	-	-	-
Tchad 1996/97	16.0	15.5	15.6	15.6	15.3	15.4	15.5	15.6	18.6	18.7	18.5	18.3	18.7	18.9	19.1	19.2	-	-
Tchad 2004	15.9	15.8	15.7	15.6	15.7	15.8	15.7	15.8	18.7	18.1	18.3	18.9	18.8	18.7	20.1	21.2	-	-
Togo 1988	16.5	16.4	16.0	16.9	16.5	17.2	16.5	16.5	-	-	-	-	-	-	-	-	-	-
Togo 1998	17.2	17.1	17.2	17.2	17.6	17.7	17.3	17.3	-	-	-	-	-	-	-	-	-	-
Zambie 1992	16.6	16.5	16.1	16.2	16.4	16.2	16.3	16.4	-	-	-	-	-	-	-	-	-	-
Zambie 1996	16.6	16.4	16.4	16.3	16.6	16.3	16.4	16.4	16.7	16.4	16.0	16.6	17.3	18.6	18.0	16.6	-	-
Zambie 2001/02	17.0	17.0	16.9	16.5	16.7	16.5	16.8	16.8	-	-	-	-	-	-	-	-	-	-
Zimbabwe 1988	18.3	17.6	17.4	17.5	16.8	17.0	17.4	17.6	-	-	-	-	-	-	-	-	-	-
Zimbabwe 1994	18.8	18.4	18.0	18.1	18.6	18.3	18.3	18.4	19.6	19.0	19.1	19.6	19.8	20.0	21.0	-	-	-
Zimbabwe 1999	18.9	19.1	19.1	18.0	18.4	18.5	18.7	18.8	19.7	19.1	19.7	19.2	20.2	19.9	20.8	-	-	-

Source : ORC Macro, MEASURE DHS STATcompiler. http://www.statcompiler.com, 29 août 2007.

Annexe 3
Intervalle de confiance bilatéral d'une proportion

Le principe général d'un intervalle de confiance consiste à déterminer, à partir de ce qui a été observé dans un sous-échantillon, un intervalle dans lequel la grandeur que l'on étudie, au sein de la population dont est extrait l'échantillon, a de fortes chances de se situer. En l'occurrence, il s'agit de déterminer un intervalle, connaissant la proportion p observée dans l'échantillon, au sein duquel la proportion π réelle de la population étudiée se situe avec une probabilité égale à une valeur fixée à l'avance, usuellement 95 %, et notée 1-α.

Il s'agit donc de rechercher a et b tels que $p[a \leq \pi \leq b]=1-\alpha$. α correspond au risque d'erreur que nous acceptons de prendre. Nous pouvons interpréter cet intervalle de confiance par "*il y a une probabilité de 1-α pour que π soit compris entre a et b*".

Dans la suite de notre propos, nous nous situerons dans le cadre d'un échantillonnage aléatoire simple issu d'une population infinie. Dans les faits, les populations que nous étudions ne sont pas infinies mais sont suffisamment grandes comparées à la taille de nos échantillons (le plus souvent, il s'agit de populations de plusieurs millions d'individus tandis que nos enquêtes portent sur quelques milliers de personnes au plus) pour que nous puissions considérer qu'il s'agit d'une population infinie. Des corrections sont techniquement possibles. En l'espèce, nous pouvons négliger cet aspect.

3.1 Loi Binomiale

Nous noterons n la taille de notre échantillon. Le nombre d'individus observés, parmi les n enquêtés, présentant le caractère étudié est donc égale à np. Ce nombre np suit ce que l'on nomme une loi Binomiale de paramètre n et π. Il s'agit d'une loi discrète et non continue. En effet, np ne peut être qu'un nombre entier. Si 100 personnes ont été enquêtées, 12 ou encore 58 parmi elles peuvent présenter présenteer un caractère donné, mais cela ne peut être 12,5 ou 58,3. Le caractère discret de la loi binomiale implique que le calcul d'un intervalle de confiance exact est relativement compliqué et a amené à de vives discussions parmi les statisticiens sur la manière adéquate de procéder (CLOPPER 1934, CASSIGNOL 1954, DUMAS 1955, VESSEREAU 1978, BRENNER 1990, COPAS 1992).

Cependant, dès lors que l'échantillon est suffisamment grand, il est possible d'avoir recours au théorème central limite qui permet de simplifier le calcul, p pouvant être alors approximé par une loi Normale.

De plus, dans le cadre d'un tirage aléatoire simple avec remise (ce qui correspond à une population-mère infinie), nous pouvons considérer que la sélection de chaque individu de l'échantillon est indépendante des autres individus sélectionnés. Il en résulte alors que le meilleur estimateur possible de la proportion réelle π est la proportion observée p. D'un point de vue plus formel, p est un estimateur sans biais, efficace et convergent de π et correspond à l'estimateur obtenu par la méthode du maximum de vraisemblance (BOUZITAT 1990, p. 164, 166, 169 et 175).

Plusieurs méthodes ont été développées pour calculer l'intervalle de confiance d'une proportion. Nous ne présenterons ici que quatre d'entre elles :

- la méthode standard traditionnelle, nommée méthode asymptotique ou bien encore méthode WALD par VOLLSET (1993) et d'autres auteurs à sa suite ;
- la méthode de score ou méthode WILSON (1927), encore appelée méthode de l'ellipse ;
- la méthode WALD avec correction de continuité (BLYTH 1983) ;
- la méthode de score de WILSON avec correction de continuité (GOSH 1979, VOLLSET 1993).

Un calculateur en ligne pour ces quatre méthodes est disponible à l'adresse suivante : http://www.ac-poitiers.fr/math/prof/resso/cali/ic_phrek.html.

Des méthodes plus complexes ont également été proposées mais elles sont plus difficiles à mettre en œuvre.

3.2 Méthode standard

Selon le théorème central limite, la moyenne expérimentale d'une répétition d'expériences identiques converge, quand n augmente, vers une loi Normale. Il en résulte que, pour n suffisamment grand, nous pouvons considérer que p suit une loi Normale de moyenne π et d'écart-type $\sqrt{\frac{\pi(1-\pi)}{n}}$. Usuellement, on considère que cette approximation est valable pour n supérieur à 30 (JOLION 2006, p. III-8). D'autres auteurs préconisent que l'on ait observé au moins 5 succès et 5 échecs : soit $np \geq 5$ et $n(1\text{-}p) \geq 5$ (WONNACOTT 1990, p. 310).

La différence $p\text{-}\pi$ suit donc une loi Normale de moyenne nulle et de même écart-type. Comme π est inconnue, nous ne connaissons pas la valeur exacte de l'écart-type. Il existe deux possibilités pour l'approximer. On utilise le majorant (c'est-à-dire la plus grande valeur possible) de $\pi(1\text{-}\pi)$ à savoir ¼. Ou bien, on remplace π par son estimation p pour le calcul de l'écart-type, ce qui est la méthode la plus courante.

Si l'on note z la valeur pour laquelle la fonction de répartition de la loi Normale centrée réduite est égale à $1\text{-}\alpha/2$, les bornes de l'intervalle de confiance de π sont alors égales à :

$$p \pm z\sqrt{\frac{p(1-p)}{n}}$$

Cette méthode est la plus courante. Si elle produit des intervalles de confiance parfaitement centrés autour de la proportion observée, elle peut induire, pour des valeurs de p proches de 0 ou de 1, des intervalles dont une partie est inférieure à 0 ou supérieure à 1. Il est alors nécessaire de tronquer l'intervalle pour supprimer les valeurs aberrantes. Cette approximation n'est donc valable que pour une valeur de p proche de 50 %.

3.3 Méthode de score

Comme pour la méthode standard, la méthode de score utilise la convergence de la loi Binomiale vers la loi Normale et approxime l'écart type en remplaçant π par p. Cette approche suggère que l'on peut obtenir un intervalle de confiance en prenant en compte les valeurs de π *telles que* :

$$\left|\frac{p-\pi}{\sqrt{p(1-p)/n}}\right| < z$$

En élevant cette inégalité au carré puis en résolvant l'équation au second degré obtenue, on détermine alors les bornes de l'intervalle de confiance par la formule :

$$\frac{2np + z^2}{2(n + z^2)} \pm \frac{z\sqrt{z^2 + 4np(1-p)}}{2(n + z^2)}$$

Cet intervalle de confiance n'est plus centré sur p et présente l'avantage de ne pas produire de valeur aberrante (inférieure à 0 ou supérieure à 1).

Pour plus de détails sur la résolution de l'équation, voir le mémoire de Patrick GAGNON (2006, p. 12-13).

3.4 Correction de continuité

Il s'agit d'une correction, initialement proposée par YATES (1934), pour tenir compte du passage d'une loi discrète à une loi continue. Chaque nombre entier x sera considéré comme couvrant l'intervalle allant de x-½ à x+½. Cela induit une légère modification des formules pour le calcul des intervalles de confiance.

Pour la méthode de WALD on obtient ainsi :

$$p \pm \left[z\sqrt{\frac{p(1-p)}{n}} + \frac{1}{2n} \right]$$

et pour la méthode de WILSON :

$$B_{-} = \frac{2np + z^2 - 1 - z\sqrt{z^2 - 2 - 1/n + 4p\left(n(1-p)+1\right)}}{2(n + z^2)}$$

$$B_{+} = \frac{2np + z^2 + 1 + z\sqrt{z^2 + 2 - 1/n + 4p\left(n(1-p)-1\right)}}{2(n + z^2)}$$

3.5 Choix d'une méthode

Différents auteurs ont comparé l'efficacité de plusieurs méthodes, dont les quatre présentées ici, ainsi que des méthodes de calcul plus complexes (NEWCOMBE 1998, TOBI 2005, TSAI 2007). Si la méthode classique doit être évitée parce qu'il s'agit de la moins performante et qu'elle produit des valeurs aberrantes, la méthode WILSON de score avec correction de continuité est recommandée dans la mesure où ses performances sont presque équivalentes à celles de méthodes dites exactes et où son calcul est relativement aisé.

Nous privilégierons donc cette méthode pour les analyses de cette thèse.

3.6 Valeurs courantes de *z*

Nous avons noté *z* la valeur pour laquelle la fonction de répartition de la loi Normale centrée réduite est égale à $1-\alpha/2$. Le plus souvent, les intervalles de confiance calculés sont les intervalles à 95 %, parfois 90 % et plus rarement 99 %.

Voici les valeurs de *z* correspondantes pour ces cas les plus fréquents :

- intervalle de confiance à 90 % : $z = 1{,}64485362695 \approx 1{,}645$
- intervalle de confiance à 95 % : $z = 1{,}95996398454 \approx 1{,}960$
- intervalle de confiance à 99 % : $z = 2{,}57582930355 \approx 2{,}576$

Annexe 4
Code des fonctions implémentées dans prevR

4.1 calcul.dist.cities

```
`calcul.dist.cities` <-
function (clust, cities, dist=15, type="ask", lang="en",
var.cities=c("x","y","city.name"), var.clust=c("x","y","residence"),
urban.code="Urban", dist.func = "rdist.earth", miles=FALSE)
{
    require(fields)
    #Vérification langue
    if (lang=="fr") lg <- 2
    else lg <- 1
    #Vérifications
    if (!is.numeric(dist)) stop(c("dist must be a numeric number","dist doit être un
nombre numérique.")[lg])
    if (dist<=0) stop(c("dist must be positive.","dist doit être positif")[lg])
    if (type!="ask" & type!="all") stop(c("type must be 'ask' or 'all'. See help.","ask
doit être 'ask' ou 'all'. Voir l'aide.")[lg])
    if (length(clust[[var.clust[3]]])==0) stop(c("residence must be specified in
clust.","le milieu de résidence doit être spécifié dans clust.")[lg])
    if (dist.func!="rdist" & dist.func!="rdist.earth") stop(c("dist.func must be 'rdist'
or 'rdist.earth'.","dist.func doit être égal à 'rdist' ou à 'rdist.earth'.")[lg])
```

```
    if (dist.func=="rdist.earth" & !is.logical(miles)) stop(c("'miles' must be TRUE or
FALSE.","'miles' doit être TRUE or FALSE.")[lg])
    #Préparation du nouveau fichier
    result <- clust
    result$dist.city <- 0
    result$city.name<- "NA"
    result$urban.area <- "NA"
    #Sélection des villes
    if (type=="all") {
        villes <- cities
    } else {
        liste_villes<-as.vector(cities[[var.cities[3]]])
        message(c("A window will open. Please select the cities for U parameter (use
CTRL and SHIFT).","Une fenêtre va s'ouvrir. Veuillez sélectionner les villes retnues
pour le paramètre U (utilisez les touches CTRL et SHIFT).")[lg])
        selected <- select.list(liste_villes, multiple=TRUE, title=c("Cities for U
parameter","Villes pour le paramètre U")[lg])
        villes <- cities[cities[[var.cities[3]]]==selected[1],]
        for (k in 2:length(selected)) {
            villes <- rbind (villes, cities[cities[[var.cities[3]]]==selected[k],])
        }
    }
    #Calcul des distances entre les clusters et les villes
    coord.cities <- data.frame(villes[[var.cities[1]]],villes[[var.cities[2]]])
    coord.clust <- data.frame(clust[[var.clust[1]]],clust[[var.clust[2]]])
    names(coord.cities) <- c("x","y")
    names(coord.clust) <- c("x","y")
    if (dist.func=="rdist.earth")
        distances <- rdist.earth(coord.clust,coord.cities, miles=miles)
    else
        distances <- rdist(coord.clust,coord.cities)
    #Transformation du nom de la ville en character
    villes[[var.cities[3]]] <- as.character (villes[[var.cities[3]]])
    #On attribue la ville correcte à chaque cluster
    for (i in 1:length(result$dist.city)) {
        distances.cluster <- distances[i,]
        result$dist.city[i] <- min(distances.cluster)
        result$city.name[i] <- villes[[var.cities[3]]][which.min(distances.cluster)]
        if(result$dist.city[i]<=dist & result[[var.clust[3]]][i]==urban.code)
            result$urban.area[i] <- "in urban area"
        else
            result$urban.area[i] <- "outside urban area"
    }
    #On transforme in.urb.aggl en facteur
    result$urban.area <- as.factor (result$urban.area)
    #On renvoit le résultat final
    return(result)
}
```

4.2 check.names

```
`check.names` <-
function (data, lang='en', size.max = 10) {
    #Vérification langue
    if (lang=="fr") lg <- 2
    else lg <- 1
    #vérifications
    if (!is.data.frame(data)) stop(c("data must be a data.frame.","data doit être un
tableau de données.")[lg])
    if (!is.numeric(size.max)) stop(c("size.max must be an integer.","size.max doit
être un entier.")[lg])
    #Récupération des noms
    noms <- as.data.frame(names(data), stringsAsFactors=FALSE)
    names(noms) <- 'variable'
    noms$size <- nchar(noms$variable)
    # Modification des noms
    while(max(noms$size)>size.max) {
        message(c("Certain variables have a name exceeding ","Certaines variables
ont un nom dépassant ")[lg],size.max,c(" characters. Please enter new names.","
caractères. Veuillez entrer de nouveaux noms.")[lg])
        message(c("To remove a variable, enter NULL.","Pour supprimer une
variable, entrez NULL.")[lg])
        noms <- edit(noms)
        noms$size <- nchar(noms$variable)
    }
    result <- data
    names(result) <- noms$variable
    #Supression des colonnes inutiles
    for (i in 1:length(noms[noms$variable=='NULL',]$variable)) result[['NULL']]
<- NULL
    #Renvoi final
    return(result)
}
```

4.3 estimate.prev

```
`estimate.prev` <-
function (clust, N=seq(100,500,50), R=Inf, U=FALSE, dist.func="rdist.earth",
miles=FALSE, lang="en", progression=TRUE, merge.result=FALSE,
var.clust=c("x","y","n","nweight","obs.prevalence","urban.area","city.name"),
urban.area.code="in urban area")
{
    begin.t <- Sys.time()
    require(fields)
    #Vérification langue
    if (lang=="fr") lg <- 2
    else lg <- 1
    if (progression) message(c("Beginning of calculations: ","Début des calculs :
")[lg],format(begin.t,"%X"))
    #Vérifications
```

```
    if (!is.numeric(clust[[var.clust[1]]])) stop(var.clust[1],c(" (x coordinate) must be
numeric."," (coordonnées x) doit être numérique.")[lg])
    if (!is.numeric(clust[[var.clust[2]]])) stop(var.clust[2],c(" (y coordinate) must be
numeric."," (coordonnées y) doit être numérique.")[lg])
    if (!is.numeric(clust[[var.clust[3]]])) stop(var.clust[3],c(" (number of persons by
cluster) must be numeric."," (nombre de personnes par cluster) doit être
numérique.")[lg])
    if (!is.numeric(clust[[var.clust[4]]])) stop(var.clust[4],c(" (weighted total of
persons by cluster) must be numeric."," (total pondérés du nombre de personnes
par cluster) doit être numérique.")[lg])
    if (!is.numeric(clust[[var.clust[5]]])) stop(var.clust[5],c(" (observed prevalence
by cluster) must be numeric."," (prévalence observée par cluster) doit être
numérique.")[lg])
    if (!is.logical(U) & U!=2) stop(c("U must be TRUE, FALSE or 2.","U doit être
TRUE, FALSE ou 2.")[lg])
    if ((U | U==2) & length(clust[[var.clust[6]]])==0) stop(var.clust[6],c(" (urban
area) must be specified."," (milieu de résidence recodé) doit être spécifiée.")[lg])
    if ((U | U==2) & length(clust[[var.clust[7]]])==0) stop(var.clust[7],c(" (city
name) must be specified."," (nom des villes) doit être spécifiée.")[lg])
    if ((U | U==2) & !is.factor(clust[[var.clust[6]]])) clust[[var.clust[6]]] <-
as.factor(clust[[var.clust[6]]])
    if ((U | U==2) &
length(clust[clust[[var.clust[6]]]==urban.area.code,][[var.clust[6]]])==0)
stop(var.clust[6],c(" must contain at least a cluster where "," doit contenir au moins
un cluster tel que ")[lg],var.clust[[6]],c(" is equal to '"," est égal à
'")[lg],urban.area.code,"'.")
    if ((U | U==2) & !is.character(clust[[var.clust[7]]])) clust[[var.clust[7]]] <-
as.character(clust[[var.clust[7]]])
    if(!is.numeric(R)) stop(c("R must be a number or a list of numbers.","R doit être
un nombre ou une liste de nombres.")[lg])
    if(!is.numeric(N)) stop(c("N must be a number or a list of numbers.","N doit
être un nombre ou une liste de nombres.")[lg])
    if (dist.func!="rdist" & dist.func!="rdist.earth") stop(c("dist.func must be 'rdist'
or 'rdist.earth'.","dist.func doit être égal à 'rdist' ou à 'rdist.earth'.")[lg])
    if (dist.func=="rdist.earth" & !is.logical(miles)) stop(c("'miles' must be TRUE or
FALSE.","'miles' doit être TRUE or FALSE.")[lg])
    N <- sort(N,decreasing=TRUE)
    R <- sort(R,decreasing=TRUE)
    #Calcul de la matrice des distances
    coord.clust <- data.frame(clust[[var.clust[1]]],clust[[var.clust[2]]])
    names(coord.clust) <- c("x","y")
    if (dist.func=="rdist.earth")
        distances <- rdist.earth(coord.clust, miles=miles)
    else
        distances <- rdist(coord.clust)
    #Préparation des nouvelles variables
    clust$est.prevalence <- as.numeric(NA)
    clust$circle.count <- as.numeric(NA)
    clust$circle.radius <- as.numeric(NA)
    clust$circle.nb.clusters <- as.integer(NA)
    clust$quality.indicator <- as.numeric(NA)
    clust$N.parameter <- as.numeric(NA)
    clust$R.parameter <- as.numeric(NA)
    clust$U.parameter <- as.numeric(NA)
    result <- data.frame()
    if (U | U==2) U.param <- length(table(clust[[var.clust[7]]]))
```

```
    #Calculs pour chaque cluster
    for (i in 1:length(clust[[1]])) {
        one.clust <- clust[i,]
        temp <- clust
        temp$dist <- distances[i,]
        temp <- temp[order(temp$dist),]
        #Restriction au même milieu de résidence, et même ville si dans agglo
urbaine
        if (U | U==2) {
            temp.U <- temp[temp[[var.clust[6]]]==one.clust[[var.clust[6]]][1],]
            if (one.clust[[var.clust[6]]][1]==urban.area.code)
                temp.U <-
temp.U[temp.U[[var.clust[7]]]==one.clust[[var.clust[7]]][1],]
            temp.U$cum.n <- cumsum (temp.U[[var.clust[3]]])
        }
        #Calculs sans paramètre U
        if (!U | U==2) {
            temp$cum.n <- cumsum (temp[[var.clust[3]]])
            for (j in N) {
                if(length(temp[temp$cum.n>=j,]$cum.n)==0)
                    maxi <- Inf
                    else maxi <- min(temp[temp$cum.n>=j,]$cum.n)
                temp <- temp[temp$cum.n<=maxi,]
                for (k in R) {
                    temp2 <- temp[temp$dist<=k,]
                    one.result <- one.clust
                    one.result$est.prevalence <-
weighted.mean(temp2[[var.clust[5]]],temp2[[var.clust[4]]],na.rm=TRUE)
                    one.result$circle.count <- sum(temp2[[var.clust[3]]])
                    one.result$circle.radius <- max(temp2$dist)
                    one.result$circle.nb.clusters <- length(temp2$dist)
                    one.result$quality.indicator <-
one.result$circle.radius^2/sqrt(one.result$circle.count)
                    one.result$N.parameter <- j
                    one.result$R.parameter <- k
                    one.result$U.parameter <- 0
                    result <- rbind(result,one.result)
                }
            }
        }
        #Calculs pour le paramètre U
        if (U | U==2) {
            for (j in N) {
                if(length(temp.U[temp.U$cum.n>=j,]$cum.n)==0)
                    maxi <- Inf
                    else maxi <- min(temp.U[temp.U$cum.n>=j,]$cum.n)
                temp.U <- temp.U[temp.U$cum.n<=maxi,]
                for (k in R) {
                    temp2.U <- temp.U[temp.U$dist<=k,]
                    one.result <- one.clust
                    one.result$est.prevalence <-
weighted.mean(temp2.U[[var.clust[5]]],temp2.U[[var.clust[4]]],na.rm=TRUE)
                    one.result$circle.count <- sum(temp2.U[[var.clust[3]]])
                    one.result$circle.radius <- max(temp2.U$dist)
                    one.result$circle.nb.clusters <- length(temp2.U$dist)
```

```
                    one.result$quality.indicator <-
one.result$circle.radius^2/sqrt(one.result$circle.count)
                    one.result$N.parameter <- j
                    one.result$R.parameter <- k
                    one.result$U.parameter <- U.param
                    result <- rbind(result,one.result)
                }
            }
        }
        if (progression & i%%25==0) message("Cluster ",i,c(" of "," sur
")[lg],length(clust[[1]]),c(" finished."," terminé.")[lg],"
(",format(Sys.time(),"%X"),")")
    }
    #Trier les résultats
    result <-
result[order(result$U.parameter,result$N.parameter,result$R.parameter),]
    #Affichage progression
    end.t <- Sys.time()
    if (progression) message(c("\nEnd of calculations: ","\nFin des calculs :
")[lg],format(end.t,"%X"))
    if (progression) message(c("Time of calculations: ","Temp de calcul :
")[lg],round(as.numeric(end.t-begin.t),digits=2)," ",attr(end.t-begin.t,"units"),".")
    #Résultats
    if (merge.result) result <- merge.prev(result)
    return(result)
}
```

4.4 extract.col

```
`extract.col` <-
function (x,value='.pred')
{
    #Récupération des noms
    noms <- names(x)
    #Recherche des colonnes à supprimer (soit celles non gardées)
    for (i in 1:length(noms)) if (length(grep(value,noms[i]))==0) noms[i]<-'NULL'
    #Réaffectation des noms
    result <- x
    names(result) <- noms
    #Suppression des colonnes inutiles
    for (i in 1:length(noms[noms=='NULL'])) result[['NULL']] <- NULL
    #Renvoi final
    return(result)
}
```

4.5 extract.data

```
`extract.data` <-
function (data, value="ask", lang="en",
var.data=c("N.parameter","R.parameter","U.parameter"))
{
    #Vérification langue
    if (lang=="fr") lg <- 2
    else lg <- 1
    #Vérifications
    ask=FALSE
    if (length(value)==1) if (value=="ask") ask=TRUE
    if (!ask & length(value)!=length(var.data)) stop(c("value and var.data must have
the same number of elements.","value et var.data doivent avoir le même nombre
d'éléments.")[lg])
    if (length(value)!=length(na.exclude(value))) stop(c("NA is not allowed for
value.","Les valeurs NA ne sont pas permises dans value.")[lg])
    if (length(var.data)!=length(na.exclude(var.data))) stop(c("NA is not allowed for
var.data.","Les valeurs NA ne sont pas permises dans var.data.")[lg])
    for (i in 1:length(var.data)) {
        if (length(data[[var.data[i]]])==0) message(var.data[i],c(" is not present in
data."," n'est pas présent dans data.")[lg])
    }
    #Préparation fichier
    result <- data
    #Pour chaque variable
    for (i in 1:length(var.data)) {
        if (length(result[[var.data[i]]])!=0) { #Si le paramètre n'est pas spécifié, on
passe au paramètre suivant.
            if (ask) {
                choix <- levels(as.factor(result[[var.data[i]]]))
                if (length(choix)==1) {
                    message(c("Only one value of ","Une seule valeur de
")[lg],var.data[i],c(" was observed. It is : "," a été observée. Il s'agit de :
")[lg],choix,".")
                    val.param <- choix
                } else {
                    message(c("Choose the value of ","Choisissez la valeur de
")[lg],var.data[i]," :")
                    val.param <- choix[menu(choix)]
                }
            }else { #La valeur du paramètre est fournie
                val.param <- value[i]
            }
            result <- result[result[[var.data[i]]]==val.param,]
            if (length(result[[1]])==0) stop(c("This choice for ","ce choix pour
")[lg],var.data[i],c(" induces an empty selection."," induit une sélection vide.")[lg])
        }
    }
    #Résultat final
    return(result)
}
```

4.6 infos.prev

```
`infos.prev` <-
function (prev, lang="en", var.n='n', var.nweight='nweight',
var.obs.prevalence='obs.prevalence', var.city.name='city.name',
var.circle.radius='circle.radius') {
    #Vérification langue
    if (lang=="fr") lg <- 2
    else lg <- 1
    #Calculs
    nb.clusters <- length(prev[[1]])
    n.total <- sum(prev[[var.n]])
    global.prevalence <-
round(weighted.mean(prev[[var.obs.prevalence]],prev[[var.nweight]],na.rm=TRU
E),digits=2)
    message(c("Statistics of this file:","Statistiques du fichier :")[lg])
    message("* ",nb.clusters,c(" clusters."," clusters.")[lg])
    message("* ",n.total,c(" valid observations."," observations valides.")[lg])
    message("* ",c("Global prevalence of ","Prévalence globale de
")[lg],global.prevalence,"%.")
    message("* ",c("Value of Noptimal proposed: ","Valeur de Noptimal proposée :
")[lg],N.optim(n.total,global.prevalence,nb.clusters))
    if (length(prev[[var.city.name]])!=0) {
        villes <- levels(as.factor(prev[[var.city.name]]))
        message("* ",c("Number of found urban centers: ","Nombre
d'agglomérations urbaines trouvées : ")[lg],length(villes))
        message("* ",c("Names of found urban centers:","Noms des agglomérations
urbaines trouvées :")[lg])
        message(paste(villes,collapse=', '),'.')
    }
    if (length(prev[[var.circle.radius]])!=0) {
        message("* ",c("Quantiles of the radius of the smoothing circles:","Quantiles
des rayons des cercles de lissage :")[lg])
        print(round(quantile(prev[[var.circle.radius]],
probs=c(0.5,0.75,0.80,0.85,0.90,0.95,0.99)),digits=1))
    }
}
```

4.7 krige.prev

```
`krige.prev` <-
function (data, formula=est.prevalence~1, locations=~x+y, type="ask",
boundary=NULL, cell.size=0.05, ask.cell.size=TRUE, lang="en",
model=vgm(1,"Exp",1), idp=2, show.variogram=TRUE, ...)
{
    #NB : pour voir les différents types de modèles de variogrammes :
show.vgms()
    #Variable type : si ask, essaye de fitter le variogramme expérimental à partir
de model et propose à l'utilsateur, si auto, essaye de fitter et prend directement le
variogramme fitté, si model, prend directement model, si idw, réalise une
interpolation idw.
```

```
    #Variable formula : peut-être une liste de formules afin de réaliser plusieurs
estimations à la suite avec la même grille pour pouvoir les représenter ensuite. Si
plusieurs formules sont proposées, le type model sera désactivé.
    require('gstat')
    #Vérification langue
    if (lang=="fr") lg <- 2
    else lg <- 1
    #Vérifications et préparation des données
    if (class(formula)=="formula") formula <- c(formula)
    if (!is.list(formula)) stop(c("formula must be a formula or a list of
formulas.","formula doit être une formule ou une liste de formules.")[lg])
    for (i in 1:length(formula)) if(class(formula[[i]])!="formula") stop(c("formula
must be a formula or a list of formulas.","formula doit être une formule ou une liste
de formules.")[lg])
    termes <- attr(terms(locations),"term.labels")
    if (length(termes)!=2) stop(c("Only two dimensions are dealt with. Please,
modify formula.","Seules deux dimensions sont prises en charge. Merci de modifier
formula.")[lg])
    if (!is.null(boundary)) {
        #Si boundary contient les mêmes variables que la première formule, alors
on prend ces variables. Sinon, on prend x et y. Sinon erreur.
        if (length(boundary[[termes[1]]])>0 & length(boundary[[termes[2]]])>0)
            boundary <- data.frame(boundary[[termes[1]]],boundary[[termes[2]]])
        else if (length(boundary[['x']])>0 & length(boundary[['y']])>0) {
            boundary <- data.frame(boundary[['x']],boundary[['y']])
            message(termes[1]," & ",termes[2],c(" are not informed in boundary. X
and y were taken."," sont non renseignés dans boundary. x et y ont été pris à la
place.")[lg])
        } else
            stop(c("boundary contains neither the terms specified in formula, nor
variables X and Y.","boundary ne contient ni les termes spécifiés dans formula, ni
de variables x et y.")[lg])
        names(boundary) <- termes
    }
    if (!any(type==c("ask","auto","model","idw"))) stop(c("type can take only one of
the following values: ask, auto, model, idw.","type ne peut prendre qu'une des
valeurs suivantes : ask, auto, model, idw.")[lg])
    nb.krige <- length(formula)
    if (type!= "idw" & any(class(model)=="variogramModel")) model <-
rep(list(model),nb.krige)
    if (type!= "idw") for (i in 1:nb.krige) if
(!any(class(model[[i]])=='variogramModel')) stop(c("model must be a object or a
list of objects of class 'variogramModel'. See help (vgm).","model doit être un objet
ou une liste d'objets de la classe 'variogramModel'. Voir help(vgm).")[lg])
    if (type!='idw' & length(model)!=nb.krige) stop(c("formula and model must
have the same number of objects.","formula et model doivent avoir le même
nombre d'objets.")[lg])
    if (type== 'idw' & length(idp)==1) {
        if (!is.numeric(idp)) stop(c("idp must be numeric.","idp doit être
numérique.")[lg])
        idp <- rep(idp, nb.krige)
    }
    if (type=='idw' & length(idp)!=nb.krige) stop(c("formula and idp must have the
same number of objects.","formula et idp doivent avoir le même nombre
d'objets.")[lg])
    #Définition de la grille
```

```
    if (is.null(boundary)) {
        ok <- FALSE

        max.x <- max(data[[termes[1]]])
        min.x <- min(data[[termes[1]]])
        max.y <- max(data[[termes[2]]])
        min.y <- min(data[[termes[2]]])
    } else {
        max.x <- max(boundary[[1]])
        min.x <- min(boundary[[1]])
        max.y <- max(boundary[[2]])
        min.y <- min(boundary[[2]])
    }
    repeat {
        nb.cells.x = (max.x-min.x)%/%cell.size+1
        nb.cells.y = (max.y-min.y)%/%cell.size+1
        message(c("A cell size of ","Une taille de cellule de ")[lg],cell.size,c(" induces
a grid of "," induit une grille de ")[lg],nb.cells.x,"x",nb.cells.y,c(" cells."," 
cellules.")[lg])
        if (!ask.cell.size) break
        message(c("Is that appropriate to you?","Cela vous convient-il ?")[lg])
        ok <- menu(c(c("Yes","Oui")[lg],c("No","Non")[lg]))
        if (ok==1) break
        message(c("Enter a new value for cell size :","Entrez une nouvelle valeur de
taille de cellule :")[lg])
        cell.size <- edit(cell.size)
        next
    }
    grid_topo <-
GridTopology(c(min.x,min.y),c(cell.size,cell.size),c(nb.cells.x,nb.cells.y))
    locations.data <- as.data.frame(coordinates(grid_topo))
    names(locations.data) <- termes
    #Préparation data
    coordinates(data) <- locations
    coordinates(locations.data) <- locations
    #Interpolation
    for (i in 1:length(formula)) {
        nom <- paste(formula[[i]][2],formula[[i]][1],formula[[i]][3])
        message("\n-------- ",nom," --------\n")
        if (type=="idw") {
            #IDW
            result.one <- idw(formula[[i]], data,locations.data, idp=idp[[i]], ...)
        } else {
            #KRIGEAGE
            sample.vario <- variogram(formula[[i]],data)
            if (any(type==c("ask","auto")))
                model[[i]] <- fit.variogram(sample.vario,model=model[[i]])
            if (type=="ask") {
                X11()
                repeat {

    plot(x=sample.vario$dist,y=sample.vario$gamma,xlab="Distance",ylab="Semi-
variance",main=paste(c("Semi-variogram","Semi-variogramme")[lg],nom))

    lines.default(variogramLine(model[[i]],maxdist=max(sample.vario$dist)))
                    message("\n--- ",nom," ---\n")
```

```
                print(model[[i]])
                message(c("\nIs this variogram model appropriate?","\nCe modèle
de variogramme convient-il ?")[lg])
                ok <- menu(c(c("Yes","Oui")[lg],c("No","Non")[lg]))
                if (ok==1) break
                message(c("Input a new model.","Entrez un nouveau
modèle.")[lg])
                model[[i]] <- edit(model[[i]])
                class(model[[i]]) <- c('variogramModel', 'data.frame')
                next
            }
            dev.off()
        }
        message(c("Variogram model used:","Modèle de variogramme utilisé
:")[lg])
        print(model[[i]])
        result.one <- krige(formula[[i]], data,locations.data, model=model[[i]],
...)
        if (i==1) {
            liste.noms <- list(nom)
            liste.sample.vario <- list(sample.vario)
            liste.model <- list(model[[i]])
        } else {
            liste.noms[[i]] <- nom
            liste.sample.vario[[i]] <- sample.vario
            liste.model[[i]] <- model[[i]]
        }
    }
    gridded(result.one) <- TRUE
    temp <- attr(result.one, 'data')
    nom.var <- attr(terms(formula[[i]]),'variables')[[2]]
    names(temp) <- c(paste(nom.var,'.pred',sep=''),paste(nom.var,'.var',sep=''))
    if (!is.null(boundary)) {
        if (i==1) {
            temp.coord <- attr(result.one, 'coords')
            is.inside <-
point.in.polygon(temp.coord[,1],temp.coord[,2],boundary[,1],boundary[,2])
        }
        temp[is.inside==0,1]<-NA
        temp[is.inside==0,2]<-NA
    }
    if (i==1) {
        attr(result.one, 'data') <- temp
        result <- result.one
    } else {
        data.result <- attr(result, 'data')
        data.result <- cbind(data.result,temp)
        attr(result, 'data') <- data.result
    }
}
if (show.variogram & type!="idw") {
    for (k in 1:length(liste.noms)) {
        X11()

    plot(x=liste.sample.vario[[k]]$dist,y=liste.sample.vario[[k]]$gamma,xlab="Dis
tance",ylab="Semi-variance",main=liste.noms[[i]])
```

```
    lines.default(variogramLine(liste.model[[k]],maxdist=max(liste.sample.vario[[k
]]$dist)))
        }
    }
    return(result)

}
```

4.8 make.boundary.dcw

```
`make.boundary.dcw` <-
function (file, progression=TRUE, lang="en")
{
    #Vérification langue
    if (lang=="fr") lg <- 2
    else lg <- 1
    #Lecture du fichier
    boundary <- readLines(file)
    country <- boundary[1]
    boundary <- boundary[-1]

    #On calcule le nombre de polygones contenus dans le fichier
    temp <- boundary[regexpr(" ",boundary)!=1]
    temp <- temp[temp!="END"]
    nb.polygons <- length(temp)

    #On créé un data.frame avec les limites de l'ensemble des polygones
    nb <- 0
    type.point <- "limit"
    coord <- data.frame()
    for (i in 1:length(boundary)) {
        if (regexpr(" ",boundary[i])!=1 & boundary[i]!="END") {
            nb <- as.integer(boundary[i])
            type.point <- "center" #Les premières coordonnées de la liste sont celles
du centroïde
        } else if (boundary[i]!="END") {
            #Extraire les coordonnées et les mettre dans un dataframe.
            if (type.point=="limit") {
                point <-
data.frame(as.numeric(substr(boundary[i],0,20)),as.numeric(substr(boundary[i],2
0,40)),nb)
                names(point) <- c("x","y","poly")
                coord <- rbind(coord, point)
            }
            type.point <- "limit"
            if (progression & i%%100==0) {message(i,c(" on "," sur
")[lg],length(boundary),c(" points analyzed."," points traités.")[lg])}
        }
    }
    #On affiche les différentes zones
    plot.new()
```

```
    plot.window(xlim=c(min(coord$x),max(coord$x)),ylim=c(min(coord$y),max(c
oord$y)),asp=1)
    title(main=country,xlab="longitude",ylab="latitude")
    col <- c("#E41A1C", "#377EB8", "#4DAF4A", "#984EA3", "#FF7F00",
"#FFFF33", "#A65628", "#F781BF", "#999999")
    lab.legend <- c()
    col.legend <- c()
    counter <- 0
    counter2 <- 0
    id <- vector("integer",12)
    for (j in 1:nb.polygons) {
        poly <- coord[coord$poly==j,]
        #Si trop de polygones on n'affiche pas les petits polygones
        if(nb.polygons<10 || length(poly$x)>50)
            {
            counter <- counter+1
            polygon(poly$x, poly$y, col=col[counter%%9+1])
            lab.legend <- c(lab.legend,paste(c("Polygon ","Polygone ")[lg],counter))
            col.legend <- c(col.legend, col[counter%%9+1])
            id[counter] <- j
        } else {counter2 <- counter2 + 1}
    }
    #Informations sur le fichier
    message("\n-------------------------------------------------------\n")
    message(c("Maximum longitude observed in the file: ","Longitude maximale
observée dans le fichier : ")[lg],max(poly$x))
    message(c("Minimum longitude observed in the file: ","Longitude minimale
observée dans le fichier : ")[lg],min(poly$x))
    message(c("Maximum latitude observed in the file: ","Latitude maximale
observée dans le fichier : ")[lg],max(poly$y))
    message(c("Minimum latitude observed in the file: ","Latitude minimale
observée dans le fichier : ")[lg],min(poly$y))
    message("\n-------------------------------------------------------\n")
    #Extraction de la base finale
    if (nb.polygons==1) {
        message(c("This file contains only one polygon.","ce fichier ne contient qu'un
seul polygone.")[lg])
        message(c("The limit of the country are presented in a new window.","Les
limites du pays ont été affichées dans une nouvelle fenêtre.")[lg])
        k <- 1
    } else {
        legend("topleft", legend=lab.legend, fill=col.legend)
        message(c("This file contains ","Ce fichier contient ")[lg],nb.polygons,c("
polygons."," polygones.")[lg])
        message(c("A new window presents the polygons contained in the file and
their number.","Une nouvelle fenêtre vous montre les différents polygones
contenus dans le fichier et leur numéro.")[lg])
        message(c("WARNING: some polygons are not visible (small island for
example).","ATTENTION : certains polygones sont invisibles à l'oeil nu (petites îles
par exemple).")[lg])
        if(counter2>0) message(c("WARNING: ","ATTENTION : ")[lg],counter2,c("
small polygons have not been drawned (less than 50 points)."," petits polygones
n'ont pas été dessinés (moins de 50 points).")[lg])
        message(c("Please, select the main polygon which will be used as limits of
the country.","Veuillez sélection le polygone principal qui sera utilisé comme
limites du pays.")[lg])
```

```
        k <- menu(lab.legend)
    }
    final <- coord[coord$poly==id[k],]
    final <- final[-3]
}
```

4.9 make.cities.csv

```
`make.cities.csv` <-
function (file, lang="en") {
    #Vérification langue
    if (lang=="fr") lg <- 2
    else lg <- 1
    #Lecture du fichier
    temp <- read.csv (file)
    temp.var <- attr(temp,"names")
    #Choix des variables
    message(c("A window will open presenting the data contained in the file. Thank
you to identify the following variables (all needed): \n- Names of cities \n-
Longitude (decimal format in degrees) \n- Latitude (decimal format in degrees) \n-
Population count \n Once the names of these variables identified, close the window
so that the program can continue. \n\n Are you ready?","Une fenêtre va s'ouvrir
présentant les données contenues dans le fichier. Merci de repérer les variables
suivantes (toutes nécessaires) :\n- Nom des villes \n- Longitude (en degrés au
format décimal) \n- Latitude (en degrés au format décimal) \n- Population
(effectif) \n Une fois les noms de ces variables identifiés, fermez la fenêtre pour que
le programme puisse continuer. \n\n Êtes-vous prêt ?")[lg])
    menu(c("Yes","Oui")[lg])
    edit(temp)
    ok <- 0
    while (ok!=1) {
        message("\n-------------------------------------------------------\n")
        message(c("Please indicate the following variables:\n","Veuillez indiquer les
variables suivantes :\n")[lg])
        message(c("\nNames of cities:","\nNom des villes :")[lg])
        city.name <- menu(temp.var)
        message(c("\nLongitude (decimal value):","\nLongitude (valeur décimale)
:")[lg])
        long <- menu(temp.var)
        message(c("\nLatitude (decimal value):","\nLatitude (valeur décimale)
:")[lg])
        lat <- menu(temp.var)
        message(c("\nPopulation of cities:","\nPopulation des villes :")[lg])
        pop <- menu(temp.var)
        #Affichage des données entrées
        message("\n-------------------------------------------------------\n")
        message(c("CAUTION: please check the following
informations:\n","ATTENTION : veuillez vérifier les informations suivantes
:\n")[lg])
        message(c("* Names of city:","* Nom des villes :")[lg])
```

```
        message(temp.var[city.name],if(city.name==0) c("Not available -
WARNING: this variable must be specified!!","Non renseignée - ATTENTION :
cette variable doit être renseignée !!")[lg])
        message(c("* Longitude (decimal value):","* Longitude (valeur décimale)
:")[lg])
        message(temp.var[long],if(long==0) c("Not available - WARNING: this
variable must be specified!!","Non renseignée - ATTENTION : cette variable doit
être renseignée !!")[lg])
        message(c("* Latitude (decimal value):","* Latitude (valeur décimale) :")[lg])
        message(temp.var[lat],if(lat==0) c("Not available - WARNING: this variable
must be specified!!","Non renseignée - ATTENTION : cette variable doit être
renseignée !!")[lg])
        message(c("* Population of cities:","* Population of cities :")[lg])
        message(temp.var[pop],if(pop==0) c("Not available - WARNING: this
variable must be specified!!","Non renseignée - ATTENTION : cette variable doit
être renseignée !!")[lg])
        message("\n-----------------------------------------------------\n")
        #Vérification des données entrées
        alarm()
        if (city.name==0) message(c("WARNING: Names of city must be
specified!!","ATTENTION : le nom des villes doit être spécifié !!")[lg])
        if (long==0) message(c("WARNING: Longitude must be
specified!!","ATTENTION : la longitude doit être spécifiée !!")[lg])
        if (lat==0) message(c("WARNING: Latitude number must be
specified!!","ATTENTION : la latitude doit être spécifiée !!")[lg])
        if (pop==0) message(c("WARNING: Population of city must be
specified!!","ATTENTION : la population des villes doit être spécifiée !!")[lg])
        if (city.name==0 | long==0 | lat==0 | pop==0) {
            message("\n-----------------------------------------------------\n")
            message(c("WARNING: some problem was found (see above). You have
to start again. Are you ready?","ATTENTION : un problème a été détecté (voir ci-
dessus). Vous devez recommencer. Êtes-vous prêt ?")[lg])
            menu(c("Yes","Oui")[lg])
            }
        else
            {
            message(c("Are these data correct?","Ces données sont-elles correctes
?")[lg])
            ok <- menu(c(c("Yes","Oui")[lg],c("No","Non")[lg]))
            }
    }
    #Création de la base
    villes <- data.frame(temp[city.name],temp[long],temp[lat],temp[pop])
    names(villes)[1] <- "city.name"
    names(villes)[2] <- "x"
    names(villes)[3] <- "y"
    names(villes)[4] <- "population"
    #Tri décroissant selon la taille de la ville
    villes <- villes[order(villes$population, decreasing = TRUE),]
    return(villes)
}
```

4.10 make.clust.dbf

```
`make.clust.dbf` <-
function (file, ind, lang="en")
{
    #Vérification langue
    if (lang=="fr") lg <- 2
    else lg <- 1
    #Vérification des variables de ind
    if (!is.integer(ind$cluster)) ind$cluster <- as.integer(as.character(ind$cluster))
    if (!is.numeric(ind$weight)) ind$weight <-
as.numeric(as.character(ind$weight))
    if (levels(ind$result)[1]!="Negative" | levels(ind$result)[2]!="Positive")
        stop(c("Result variable from ind is not correctly formatted. It must be a
factor format, with levels 'Negative' and 'Positive'.","La colonne result de ind n'est
pas renseignée correctement. Cela doit être au format factor, avec pour levels les
valeurs 'Negative' et 'Positive'")[lg])
    ind$statut <- NA
    ind$statut[ind$result=="Negative"] <- 1
    ind$statut[ind$result=="Positive"] <- 2
    #Lecture du fichier de données et du nom des variables
    require(foreign)
    temp.clust <- read.dbf(file)
    temp.var <- attr(temp.clust,"names")
    #Identification de chaque variable
    message(c("A window will open presenting the data contained in the file. Thank
you to identify the following variables: \n- Cluster number (needed) \n- Longitude
(decimal format in degrees, needed) \n- Latitude (decimal format in degrees,
needed) \n- Place of residence (urban/rural, needed for U parameter) \n- Code
number of DHSregions (optional) \n- Name of regions (optional) \n Once the
names of these variables identified, close the window so that the program can
continue. \n\n Are you ready?","Une fenêtre va s'ouvrir présentant les données
contenues dans le fichier. Merci de repérer les variables suivantes :\n- Numéro du
cluster (nécessaire) \n- Longitude (en degrés au format décimal, nécessaire) \n-
Latitude (en degrés au format décimal, nécessaire) \n- Milieu de résidence
(urbain/rural, nécessaire si utilisation du paramètre U) \n- Code numérique des
régions (optionnel) \n- Nom des régions (optionnel) \n Une fois les noms de ces
variables identifiés, fermez la fenêtre pour que le programme puisse continuer.
\n\n Êtes-vous prêt ?")[lg])
    menu(c("Yes","Oui")[lg])
    edit(temp.clust)
    ok <- 0
    while (ok!=1) {
        message("\n-----------------------------------------------------\n")
        message(c("Please indicate the following variables:\n","Veuillez indiquer les
variables suivantes :\n")[lg])
        message(c("\nCluster number:","\nNuméro des clusters :")[lg])
        nb.clust <- menu(temp.var)
        message(c("\nLongitude (decimal value):","\nLongitude (valeur décimale)
:")[lg])
        long <- menu(temp.var)
        message(c("\nLatitude (decimal value):","\nLatitude (valeur décimale)
:")[lg])
        lat <- menu(temp.var)
        message(c("\nType of residence:","\nMilieu de résidence :")[lg])
```

```
        residence <- menu(temp.var)
        message(c("\nCode number of regions (0 if doesn't exist):","\nCode
numérique des régions (0 si non renseigné) :")[lg])
        region <- menu(temp.var)
        message(c("\nRegion names (0 if doesn't exist):","\nNom des régions (0 si
non renseigné) :")[lg])
        region.name <- menu(temp.var)
        #Affichage des données entrées
        message("\n-------------------------------------------------\n")
        message(c("CAUTION: please check the following
informations:\n","ATTENTION : veuillez vérifier les informations suivantes
:\n")[lg])
        message(c("* Cluster number:","* Numéro de cluster :")[lg])
        message(temp.var[nb.clust],if(nb.clust==0) c("Not available - WARNING:
this variable must be specified!!","Non renseignée - ATTENTION : cette variable
doit être renseignée !!")[lg])
        message(c("* Longitude (decimal value):","* Longitude (valeur décimale)
:")[lg])
        message(temp.var[long],if(long==0) c("Not available - WARNING: this
variable must be specified!!","Non renseignée - ATTENTION : cette variable doit
être renseignée !!")[lg])
        message(c("* Latitude (decimal value):","* Latitude (valeur décimale) :")[lg])
        message(temp.var[lat],if(lat==0) c("Not available - WARNING: this variable
must be specified!!","Non renseignée - ATTENTION : cette variable doit être
renseignée !!")[lg])
        message(c("* Type of residence:","* Milieu de résidence :")[lg])
        message(temp.var[residence],if(residence==0) c("Not available -
WARNING: needed to use U parameter","Non renseignée - ATTENTION : cette
variable est nécessaire pour pouvoir utiliser le paramètre U")[lg])
        message(c("* Code number of regions:","* Code numérique des régions
:")[lg])
        message(temp.var[region],if(region==0) c("Not available","Non
renseignée")[lg])
        message(c("* Region names:","* Nom des régions :")[lg])
        message(temp.var[region.name],if(region.name==0) c("Not available","Non
renseignée")[lg])
        message("\n-------------------------------------------------\n")
        #Vérification des données entrées
        alarm()
        if (nb.clust==0) message(c("WARNING: Cluster number must be
specified!!","ATTENTION : le numéro des clusters doit être spécifié !!")[lg])
        if (long==0) message(c("WARNING: Longitude must be
specified!!","ATTENTION : la longitude doit être spécifiée !!")[lg])
        if (lat==0) message(c("WARNING: Latitude number must be
specified!!","ATTENTION : la latitude doit être spécifiée !!")[lg])
        if (residence==0) message(c("WARNING: if type of residence is not
specified, you will be not able to use U parameter!!","ATTENTION : si le milieu de
résidence n'est pas précisé, vous ne pourrez utiliser le paramètre U !!")[lg])
        if (nb.clust==0 | long==0 | lat==0) {
            message("\n-------------------------------------------------\n")
            message(c("WARNING: some problem was found (see above). You have
to start again. Are you ready?","ATTENTION : un problème a été détecté (voir ci-
dessus). Vous devez recommencer. Êtes-vous prêt ?")[lg])
            menu(c("Yes","Oui")[lg])
            }
        else
```

```
            {
            message(c("Are these data correct?","Ces données sont-elles correctes
?")[lg])
            ok <- menu(c(c("Yes","Oui")[lg],c("No","Non")[lg]))
            }
    }
    #Création de la base
    clust <-
data.frame(temp.clust[nb.clust],temp.clust[long],temp.clust[lat],temp.clust[reside
nce],temp.clust[region],temp.clust[region.name])
    #Renommage des variables
    i <- 1
    if (nb.clust!=0)
        {
        names(clust)[i] <- "cluster"
        i <- i+1
        }
    if (long!=0)
        {
        names(clust)[i] <- "x"
        i <- i+1
        }
    if (lat!=0)
        {
        names(clust)[i] <- "y"
        i <- i+1
        }
    if (residence!=0)
        {
        names(clust)[i] <- "residence"
        i <- i+1
        }
    if (region!=0)
        {
        names(clust)[i] <- "region"
        i <- i+1
        }
    if (region.name!=0)
        {
        names(clust)[i] <- "region.name"
        i <- i+1
        }
    #Restriction selon l'age ou le sexe
    info <- ""
    if (!is.null(ind$sex))
        {
        message("\n-------------------------------------------------------\n")
        message(c("The sex variable was detected in the ind file.\nThis variable has
the following modalities:","La variable sexe a été détectée dans le fichier
ind.\nVoici ses modalités :")[lg])
        modalites <- attr(ind$sex,'levels')
        message("- ",paste(modalites,collapse="\n- "),"\n")
        message(c("Do you want to restrict this analyse at one or several of this
modalities ?","Voulez-vous restreindre l'analyse à l'une ou plusieurs de ces
modalités ?")[lg])
        choix <- menu(c(c("Yes","Oui")[lg],c("No","Non")[lg]))
```

```
        if (choix==1)
            {
            message(c("A window will open. Select the modalities you want to keep (use CTRL to select several modalities).","Une fenêtre va s'ouvrir. Sélectionnez la ou les modalités à retenir (utilisez la touche CTRL pour un choix multiple).")[lg])
            mod <- select.list(modalites, multiple=TRUE, title=c("Select modalities to keep.","Sélectionnez les modalités à conserver.")[lg])
            ind$weight2 <- ind$weight
            ind$weight <- 0
            for (i in 1:length(mod)) ind$weight[ind$sex==mod[i]] <- ind$weight2[ind$sex==mod[i]]
            info <- paste(info, c("The analyse was restricted with the folowing modalities:\n- ","L'analyse a été restreinte aux modalités suivantes :\n- ")[lg], paste(mod,collapse="\n- "),"\n")
            }
        }
    if (!is.null(ind$age))
        {
        message("\n-------------------------------------------------\n")
        message(c("The age variable was detected in the file.\nHere its characteristics:","La variable âge a été détectée dans le fichier ind.\nVoici ses caractéristiques :")[lg])
        print(summary(ind$age[ind$weight>0]))
        message(c("\nDo you want to restrict this analyse on an interval of this variable ?","\nVoulez-vous restreindre l'analyse sur un intervalle de cette variable ?")[lg])
        choix <- menu(c(c("Yes","Oui")[lg],c("No","Non")[lg]))
        if (choix==1)
            {
            minimum <- as.numeric(min(ind$age[ind$weight>0]))
            maximum <- as.numeric(max(ind$age[ind$weight>0]))
            message(c("Two windows will open. Modify the minimum value to keep, close the window, modify the maximum value, close the window.","Deux fenêtre vont s'ouvrir. Modifiez la valeur minimum à garder, fermez la fenêtre, modifiez la valeur maximum, fermez la fenêtre.")[lg])
            minimum <- edit(minimum)
            maximum <- edit(maximum)
            ind$weight[ind$age<minimum] <- 0
            ind$weight[ind$age>maximum] <- 0
            info <- paste(info,c("The analyse was restricted with people aged ","L'analyse a été restreinte aux personnes âgées de ")[lg],minimum,"-",maximum,c("years old."," ans.")[lg])
            }
        }
    #Vérification type des variables
    if (!is.integer(clust$cluster)) clust$cluster <- as.integer(as.character(clust$cluster))
    if (!is.numeric(clust$x)) clust$longitude <- as.numeric(as.character(clust$x))
    if (!is.numeric(clust$y)) clust$latitude <- as.numeric(as.character(clust$y))
    if (residence !=0)
        if (!is.factor(clust$residence)) clust$residence <- as.factor(clust.residence)
    #Création des nouvelles variables
    clust$n <- 0
    clust$nweight <- 0
    clust$obs.prevalence <- NA
    #Calcul, pour chaque cluster, de ces variables à partir de la base ind
```

```
    for (i in 1:length(clust$cluster)) {
        j <- clust$cluster[i]
        temp.ind <- ind[ind$cluster==j,,drop=FALSE]
        temp.ind$weight[is.na(temp.ind$statut)] <- 0
        #calcul de n, nbre d'observations valides, càd avec un résultat et poids non
nul
        clust$n[i] <- sum(table(temp.ind$weight,temp.ind$statut, exclude = c(NA,
NaN, 0)))
        clust$nweight[i] <- sum(temp.ind$weight)
        clust$obs.prevalence[i] <- weighted.mean(temp.ind$statut-
1,temp.ind$weight,na.rm=TRUE)*100
    }
    #Codification du milieu de résidence
    if (residence!=0) {
        if (!is.factor(clust$residence)) clust$residence <- as.factor(clust$residence)
        message(c("A window will open. Please select the urban item and the rural
item.","Une fenêtre va s'ouvrir. Veuillez spécifier l'item urbain et l'item rural.")[lg])
        modalite <- attr(clust$residence,'levels')
        urb<-select.list(modalite, multiple=FALSE, title=c("Urban item?","Item
urbain ?")[lg])
        rur<-select.list(modalite, multiple=FALSE, title=c("Rural item?","Item rural
?")[lg])
        modalite[modalite==urb] <- "Urban"
        modalite[modalite==rur] <- "Rural"
        attr(clust$residence,'levels') <- modalite
    }
    #Récap infos sur le fichier créé.
    message("\n-------------------------------------------------\n")
    message(info)
    infos.prev(clust, lang=lang)
    return(clust)
}
```

4.11 make.ind.spss

```
`make.ind.spss` <-
function (file, lang="en")
{
    # Vérification langue
    if (lang=="fr") lg <- 2
    else lg <- 1
    # Lecture du fichier de données et des noms de variable
    require(foreign)
    temp.ind <- read.spss2(file, use.value.labels=TRUE, to.data.frame=TRUE)
    temp.var <- paste(attr(temp.ind, "names"), attr(temp.ind,"variable.labels"),
sep=" - ")
    # Identification de chaque variable
    ok <- 0
    while (ok!=1) {
        message("\n-------------------------------------------------\n")
        message(c("Please indicate the following variables:\n","Veuillez indiquer les
variables suivantes :\n")[lg])
```

```
        message(c("\nIdentification number (0 if doesn't exist):","\nIdentifiant (0
s'il n'existe pas) :")[lg])
        id <- menu(temp.var)
        message(c("\nCluster number (0 if doesn't exist. The cluster number will be
calculated with identification number.):","\nNuméro du cluster (0 s'il n'existe pas.
Le numéro de cluster sera alors calculé à partir des identifiants.) :")[lg])
        clust <- menu(temp.var)
        message(c("\nAge (0 if doesn't exist):","\nAge (0 s'il n'existe pas) :")[lg])
        age <- menu(temp.var)
        message(c("\nSex (0 if doesn't exist):","\nSexe (0 s'il n'existe pas) :")[lg])
        sex <- menu(temp.var)
        message(c("\nAnalyzed variable (for example, result of HIV
testing):","\nVariable analysée (par exemple, résultat du test VIH) :")[lg])
        result <- menu(temp.var)
        message(c("\nStatistical weight (0 if doesn't exist. All persons will have the
same weight of 1.):","\nPoids statistique (0 s'il n'existe pas. Tous les individus
auront un poids égal à 1.) :")[lg])
        weight <- menu(temp.var)
        #Affichage des données entrées
        message("\n-------------------------------------------------\n")
        message(c("CAUTION: please check the following
informations:\n","ATTENTION : veuillez vérifier les informations suivantes
:\n")[lg])
        message(c("* Identification number:","* Identifiant des individus :")[lg])
        message(temp.var[id],if(id==0) c("Not available","Non renseigné")[lg])
        message(c("* Cluster number:","* Numéro de cluster :")[lg])
        message(temp.var[clust],if(clust==0) c("Not available - It will be calculated
with identification number.","Non renseigné - Il sera calculé à partir du numéro
d'identifiants")[lg])
        message(c("* Age:","* Age :")[lg])
        message(temp.var[age],if(age==0) c("Not available","Non renseigné")[lg])
        message(c("* Sex:","* Sexe :")[lg])
        message(temp.var[sex],if(sex==0) c("Not available","Non renseigné")[lg])
        message(c("* Analyzed variable:","* Variable analysée :")[lg])
        message(temp.var[result],if(result==0) c("Not available - WARNING: this
variable must be specified!!","Non renseignée - ATTENTION : cette variable doit
être renseignée !!")[lg])
        message(c("* Statistical weight:","* Poids statistique :")[lg])
        message(temp.var[weight],if(weight==0) c("Not available - All persons will
have a weight of 1.","Non renseigné - Tous les individus auront un poids égal à
1.")[lg])
        message("\n-------------------------------------------------\n")
        #Vérification des données entrées
        alarm()
        if (id==0 & clust==0)
            message(c("WARNING: if cluster number not specified, you need to
specify identification number!!","ATTENTION : si le numéro de cluster n'est pas
renseigné, vous devez spécifier l'identifiant des individus!!")[lg])
        if (result==0)
            message(c("WARNING: you have to specify the analyzed
variable!!","ATTENTION : vous devez spécifier la variable à analyser!!")[lg])
        if ((id==0 & clust==0) | result==0)
            {
            message("\n-------------------------------------------------\n")
```

```
            message(c("WARNING: some problem was found (see above). You have
to start again. Are you ready?","ATTENTION : un problème a été détecté (voir ci-
dessus). Vous devez recommencer. Êtes-vous prêt ?")[lg])
            menu(c("Yes","Oui")[lg])
            }
        else
            {
            message(c("Are these data correct?","Ces données sont-elles correctes
?")[lg])
            ok <- menu(c(c("Yes","Oui")[lg],c("No","Non")[lg]))
            }
    }
    #Création de la base
    ind <-
data.frame(temp.ind[id],temp.ind[clust],temp.ind[age],temp.ind[sex],temp.ind[re
sult],temp.ind[weight])
    #Renommage des variables
    i <- 1
    if (id!=0)
        {
        names(ind)[i]<-"id"
        i <- i+1
        }
    if (clust!=0)
        {
        names(ind)[i]<-"cluster"
        i <- i+1
        }
    if (age!=0)
        {
        names(ind)[i]<-"age"
        i <- i+1
        }
    if (sex!=0)
        {
        names(ind)[i]<-"sex"
        i <- i+1
        }
    if (result!=0)
        {
        names(ind)[i]<-"original.result"
        i <- i+1
        }
    if (weight!=0)
        {
        names(ind)[i]<-"weight"
        i <- i+1
        }
    #Recodage de la variable result
    ok <- 0
    message("\n-------------------------------------------------\n")
    message(c("Three windows will open in order to re-code the analysed
variable.\nYou will have to specify the modalities corresponding to a positive result
(the analysed phenomenon occured), \na negative result (not occured) and an
undetermined result (considered as a missing value).\nYou can select several
modalities with CTRL.\nAre you ready?","Trois fenêtres vont s'ouvrir pour recoder
```

```
la variable analysée.\nVous devrez spécifier les modalités correspondant à un
résultat positif (le phénomène étudié a eu lieu), \nnégatif (n'a pas eu lieu) ou
indéterminé (considéré alors comme valeur manquante).\nVous pouvez
sélectionner plusieurs modalités à l'aide de la touche CTRL.\nÊtes-vous-prêt
?")[lg])
    menu(c("Yes","Oui")[lg])
    modalites <- attr(ind$original.result,'levels')
    while (ok!=1) {
        pos <- select.list(modalites, multiple=TRUE, title=c("Positive
result","Résultat positif")[lg])
        neg <- select.list(modalites, multiple=TRUE, title=c("Negative
result","Résultat négatif")[lg])
        und <- select.list(modalites, multiple=TRUE, title=c("Undetermined
result","Résultat indéterminé")[lg])
        message("\n-------------------------------------------------\n")
        if (length(neg)+length(pos)+length(und)!=length(modalites)) {
            alarm()
            message(c("You specified the same modality two times or you forgot one.
Please start again.\nAre you ready?","Vous avez sélectionné une modalité deux
fois, ou bien vous en avez oublié une. Veuillez recommencer.\nÊtes-vous prêt
?")[lg])
            menu(c("Yes","Oui")[lg])
        } else {
            message(c("CAUTION: please check the following
informations:\n","ATTENTION : veuillez vérifier les informations suivantes
:\n")[lg])
            message(c("\n* Positive result:","\n* Résultat positif :")[lg])
            message("- ",paste(pos,collapse="\n- "))
            message(c("\n* Negative result:","\n* Résultat négatif :")[lg])
            message("- ",paste(neg,collapse="\n- "))
            message(c("\n* Undetermined result:","\n* Résultat indéterminé :")[lg])
            message("- ",paste(und,collapse="\n- "))
            message(c("\nAre these data correct?","\nCes données sont-elles
correctes ?")[lg])
            ok <- menu(c(c("Yes","Oui")[lg],c("No","Non")[lg]))
        }
    }
    ind$result <- NA
    for (i in 1:length(neg)) ind$result[ind$original.result==neg[i]] <- 1
    for (i in 1:length(pos)) ind$result[ind$original.result==pos[i]] <- 2
    ind$result <- as.factor(ind$result)
    levels(ind$result)[1] <- "Negative"
    levels(ind$result)[2] <- "Positive"
    #Vérification de la variable weight
    if (weight==0) {
        ind$weight <- 1
    } else {
        message("\n-------------------------------------------------\n")
        message(c("Often, in DHS, the weight variable have to be divided by a factor,
usually 1'000'000. Its mean value is:","Souvent, dans les EDS, la variable poids doit
être divisée par un facteur, usuellement 1 000 000. Sa valeur moyenne est de
:")[lg])
        message(round(mean(ind$weight[ind$weight>0]),digits=2))
        message(c("\nIf this value is close to 1, the variable does not have to be
modified. If it is close to 1'000'000, then it must be divided by this factor; Else
consult the survey documentation.","\nSi cette valeur est proche de 1, a priori la
```

```
variable n'a pas à être modifiée. Si elle est proche de 1 000 000, alors elle doit être
divisée par ce facteur, sinon consultez la documentation de l'enquête.")[lg])
        message(c("Does the weight variable have to be divided by a factor or staying
like actually ?","La variable doit-elle être gardée telle quelle ou divisée par un
facteur ?")[lg])
        choix <- menu(c(c("No modification","Pas de modification")[lg],c("Divided
by 1'000'000","Division par 1 000 000")[lg],c("Divided by an other
factor","Division par un autre facteur")[lg]))
        if (choix > 1) {
            if (choix==2) division.factor <- 1000000
            else {
                message(c("Wich factor?","Quel facteur de division ?")[lg])
                division.factor <- 1
                division.factor <- as.integer(de(division.factor))
            }
            ind$weight <- ind$weight/division.factor
        }
    }
    #Vérification de la variable cluster
    if (clust!=0) {
        if (is.character(ind$cluster)) ind$cluster<-as.integer(ind$cluster)
        if (is.factor(ind$cluster)) ind$cluster<-as.integer(as.character(ind$cluster))
    } else {
        #Création de la variable cluster si non renseignée
        if (!is.character(ind$id)) ind$id <- as.character(ind$id)
        exemple1 <- ind$id[1]
        exemple2 <- ind$id[length(ind$id)%/%2]
        exemple3 <- ind$id[length(ind$id)]
        message("\n-------------------------------------------------\n")
        message(c("The cluster variable is not specified. It must be calculated from
the identification number. Usually in DHS, the cluster number corresponds to the
first three digits of the identification number. You can see here three identification
number selected at the beginning, the medium and the end of the file:","La variable
cluster n'est pas renseignée et doit donc être calculée à partir du numéro
d'identifiant. Usuellement, dans les EDS, le numéro de cluster correspond aux trois
premiers chiffres du numéro d'identifiant. Voici trois numéros d'identifiants, pris
au début, au mileu et à la fin du fichier :")[lg])
        message(c("- Identification number 1: ","Numéro d'identification 1 :
")[lg],exemple1)
        message(c("- Identification number 2: ","Numéro d'identification 2 :
")[lg],exemple2)
        message(c("- Identification number 3: ","Numéro d'identification 3 :
")[lg],exemple3)
        message(c("\nLocate for each one of them the number of cluster. Among the
various proposals below, which extracts the good cluster numbers?\n","\nRepérez
pour chacun d'eux le numéro de cluster. Parmi les différentes propositions ci-
dessous, laquelle extrait les bons numéros de cluster ?\n")[lg])
        choix <- c()
        for (i in 1:(nchar(exemple1)-2))
            choix <- c(choix,paste("Cluster 1: ",substr(exemple1,i,i+2),"  -  Cluster 2:
",substr(exemple2,i,i+2),"  -  Cluster 3: ",substr(exemple3,i,i+2),"          ."))
        lim <- menu(choix)
        ind$cluster <- as.integer(substr(ind$id,lim,lim+2))
    }
    #Vérification de la variable sex
    if (sex!=0)
```

```
        if(!is.factor(ind$sex)) ind$sex <- as.factor(ind$sex)
    #Vérification de la variable age
    if (age!=0)
        if(!is.integer(ind$age)) ind$age <- as.integer(as.character(ind$age))
    #On renvoie le résultat final
    return(ind)
}
```

4.12 map.cities

```
`map.cities` <-
function (cities, boundary, var.cities=c("x","y","city.name","population"),
min.population=NULL, new.window=TRUE, ...) {
    if (new.window) X11()
    plot(boundary,asp=1,type='l', axes=FALSE,xlab=NA,ylab=NA)
    if (!is.null(min.population))
        cities <- cities[cities[[var.cities[4]]]>=min.population,]
    points(x=cities[[var.cities[1]]], y=cities[[var.cities[2]]], pch=22, bg='red')
    text(x=cities[[var.cities[1]]], y=cities[[var.cities[2]]],
labels=as.character(cities[[var.cities[3]]]), adj=c(1,1))
    title(...)
}
```

4.13 map.clust

```
`map.clust` <-
function (clust, boundary, type='urb',lang="en", var.urb='residence', var.n='n',
var.obs.prevalence='obs.prevalence', var.coords=c('x','y'), inverse=FALSE,
add.legend=TRUE, legend.location='bottomright', factor.size=0.2,
new.window=TRUE, ...) {
    #Si type='urb', on fait la carte des clusters selon milieu de résidence, si
type='flower', on fait la carte du nombre de personnes HIV+ par cluster, si
type='count' la carte du nombre de personnes testées par cluster.
    #Vérification langue
    if (lang=="fr") lg <- 2
    else lg <- 1
    #Vérifications
    if(!any(type==c('urb','flower','count'))) stop(c("type must be 'urb','flower' or
'count'.","type doit être 'urb', 'flower' ou 'count'.")[lg])
    if(type=='urb' & length(clust[[var.urb]])==0) stop(c("var.urb must be
specified.","var.urb doit être spécifée")[lg])
    if(any(type==c('flower','count')) & length(clust[[var.n]])==0) stop(c("var.n
must be specified.","var.n doit être spécifée")[lg])
    if(type=='flower' & length(clust[[var.obs.prevalence]])==0)
stop(c("var.obs.prevalence must be specified.","var.obs.prevalence doit être
spécifée")[lg])
    if(length(clust[[var.coords[1]]])==0 | length(clust[[var.coords[2]]])==0)
stop(c("var.coords must be specified.","var.coords doit être spécifée")[lg])
```

```
    if(any(names(boundary)!=c('x','y'))) stop(c("Variables names of boundary must
be x and y.","Les noms de variables de boundary doivent être x et y.")[lg])
    #Préparation carte
    if (new.window) X11()
    plot(boundary,asp=1,type='l', axes=FALSE,xlab=NA,ylab=NA)
    if (type=='urb') {
        if(inverse) niveaux <- sort(levels(clust[[var.urb]]),decreasing=TRUE)
        else niveaux <- sort(levels(clust[[var.urb]]),decreasing=FALSE)
        rur <- clust[clust[[var.urb]]==niveaux[1],]
        urb <- clust[clust[[var.urb]]==niveaux[2],]
        points(x=rur[[var.coords[1]]], y=rur[[var.coords[2]]], pch=21, bg='green')
        points(x=urb[[var.coords[1]]], y=urb[[var.coords[2]]], pch=22, bg='red')
        title(...)
        if (add.legend)
            legend(legend.location, legend=niveaux, pch=c(21,22),
pt.bg=c('green','red'))
    }
    if (type=='flower') {
        n.positif <- round(clust[[var.obs.prevalence]]*clust[[var.n]]/100)
        aucun <- clust[n.positif==0,]
        points(x=aucun[[var.coords[1]]],y=aucun[[var.coords[2]]], pch=22,
bg='green3', cex=0.7)

    sunflowerplot(x=clust[[var.coords[1]]],y=clust[[var.coords[2]]],number=n.posi
tif,add=TRUE,seg.lwd=1.5,seg.col='red3',cex=0.8, bg='VioletRed4', pch=21)
        title(...)
        if (add.legend)
            legend(legend.location, legend=c(c("No positive case","Aucun cas
positif")[lg],c("Only one positive case","Un seul cas positif")[lg],c("One positive
case by 'petal'","Un cas positif par 'pétale'")[lg]), pch=c(22,21,8),
pt.bg=c('green3','VioletRed4','red3'), col=c('black','black','red3'))
    }
    if (type=='count') {
        points(x=clust[[var.coords[1]]],y=clust[[var.coords[2]]], pch=21,
bg='skyblue3', cex=clust[[var.n]]*factor.size)
        title(...)
        if(add.legend)
            legend(legend.location, legend=c("5      ","10      ","25"), pch=c(21,21,21),
pt.bg=c('skyblue3','skyblue3','skyblue3'),
pt.cex=c(5*factor.size,10*factor.size,25*factor.size), horiz=TRUE)
    }
}
```

4.14 merge.prev

```
`merge.prev` <-
function (data) {
    #Vérifications
    if (length(data$N.parameter)==0) stop("data must contain the N.parameter
column.")
    if (length(data$R.parameter)==0) stop("data must contain the R.parameter
column.")
```

```
    if (length(data$U.parameter)==0) stop("data must contain the U.parameter
column.")
    temp <- by(data,
list(data$N.parameter,data$R.parameter,data$U.parameter),rename.variables.par
ameters)
    result <- temp[[1]]
    for (i in 2:length(temp)) {
        result <- merge(result, temp[[i]])
    }
    return(result)
}
```

4.15 N.optim

```
`N.optim` <-
function (n.total, global.prevalence, nb.clusters, b0=14.172, b1=0.419, b2=-0.361,
b3=0.037, c=-91.011) {
    return(round(b0*n.total^b1*global.prevalence^b2*nb.clusters^b3+c))
}
```

4.16 prevR.colors.blue.inverse

```
`prevR.colors.blue.inverse` <-
function (n) {
    if ((n <- as.integer(n[1])) > 0) {
        j <- n%/%2.75
        i <- n - 2*j
        c(if (j > 0) hsv(h = 33/48, v = seq(to = 1 - 1/(2 * j), from = 1/(2 * j), length =
j), s = 1) , hsv(h = seq(33/48,26/48, length = i)) , if (j > 0) hsv(h = 26/48, s =
seq(from = 1 - 1/(2 * j), to = 1/(2 * j), length = j),v = 1))
    }
    else character(0)
}
```

4.17 prevR.colors.blue

```
`prevR.colors.blue` <-
function (n) {
    if ((n <- as.integer(n[1])) > 0) {
        j <- n%/%2.75
        i <- n - 2*j
```

```
        c(if (j > 0) hsv(h = 26/48, s = seq(to = 1 - 1/(2 * j), from = 1/(2 * j), length =
j),v = 1) , hsv(h = seq(26/48,33/48, length = i)) , if (j > 0) hsv(h = 33/48, v =
seq(from = 1 - 1/(2 * j), to = 1/(2 * j), length = j), s = 1))
    }
    else character(0)
}
```

4.18 prevR.colors.gray.inverse

```
`prevR.colors.gray.inverse` <-
function (n) {
    if ((n <- as.integer(n[1])) > 0) {
        i <- n-1
        gray(0:i / i)
    }
    else character(0)
}
```

4.19 prevR.colors.gray

```
`prevR.colors.gray` <-
function (n) {
    if ((n <- as.integer(n[1])) > 0) {
        i <- n-1
        gray(i:0 / i)
    }
    else character(0)
}
```

4.20prevR.colors.green.inverse

```
`prevR.colors.green.inverse` <-
function (n)
{
    if ((n <- as.integer(n[1])) > 0) {
        j <- n%/%2
        i <- n - 2 * j
        c(if (j > 0) hsv(h = 1/3, v = seq(to = 1 - 1/(2 * j), from = 1/(2 * j), length = j), s
= 1), hsv(h = seq(1/3, 1/3, length = i)), if (j > 0) hsv(h = 1/3, s = seq(from = 1 - 1/(2
* j), to = 1/(2 * j), length = j), v = 1))
    }
```

```
    else character(0)
}
```

4.21 prevR.colors.green

```
`prevR.colors.green` <-
function (n)
{
    if ((n <- as.integer(n[1])) > 0) {
        j <- n%/%2
        i <- n - 2 * j
        c(if (j > 0) hsv(h = 1/3, s = seq(to = 1 - 1/(2 * j), from = 1/(2 * j), length = j), v
= 1), hsv(h = seq(1/3, 1/3, length = i)), if (j > 0) hsv(h = 1/3, v = seq(from = 1 - 1/(2
* j), to = 1/(2 * j), length = j), s = 1))
    }
    else character(0)
}
```

4.22 prevR.colors.red.inverse

```
`prevR.colors.red.inverse` <-
function (n) {
    if ((n <- as.integer(n[1])) > 0) {
        j <- n%/%3.5
        i <- n - 2*j
        c(if (j > 0) hsv(h = 0, v = seq(to = 1 - 1/(2 * j), from = 1/(2 * j), length = j), s =
1) , hsv(h = seq(0,1/6, length = i)) , if (j > 0) hsv(h = 1/6, s = seq(from = 1 - 1/(2 * j),
to = 1/(2 * j), length = j),v = 1))
    }
    else character(0)
}
```

4.23 prevR.colors.red

```
`prevR.colors.red` <-
function (n) {
    if ((n <- as.integer(n[1])) > 0) {
        j <- n%/%3.5
        i <- n - 2*j
        c(if (j > 0) hsv(h = 1/6, s = seq(to = 1 - 1/(2 * j), from = 1/(2 * j), length = j),v
= 1) , hsv(h = seq(1/6,0, length = i)) , if (j > 0) hsv(h = 0, v = seq(from = 1 - 1/(2 * j),
to = 1/(2 * j), length = j), s = 1))
    }
```

```
    else character(0)
}
```

4.24 prevR.demo.pal

```
`prevR.demo.pal` <-
function(n, border = if (n<32) "light gray" else NA, main = paste("Palettes prevR
n=",n),          ch.col =
c("prevR.colors.red.inverse(n)","prevR.colors.blue.inverse(n)",
"prevR.colors.green.inverse(n)","prevR.colors.gray.inverse(n)","prevR.colors.red(n
)", "prevR.colors.blue(n)", "prevR.colors.green(n)", "prevR.colors.gray(n)")) {
    #Fonction basée sur demo.pal - cf. aide de la fonction rainbow : ?rainbow
    nt <- length(ch.col)
    i <- 1:n; j <- n / nt; d <- j/6; dy <- 2*d
    plot(i,i+d, type="n", yaxt="n", ylab="", main=main)
    for (k in 1:nt) {
        rect(i-.5, (k-1)*j+ dy, i+.4, k*j,col = eval(parse(text=ch.col[k])), border =
border)
        text(2*j,  k * j +dy/4, ch.col[k],  cex=0.8)
    }
}
```

4.25 read.spss2

```
`read.spss2` <-
function (file, use.value.labels=TRUE, to.data.frame=FALSE, max.value.labels=Inf,
trim.factor.names=FALSE)
{
    #Cette fonction reprend la fonction read.spss du package foreign en la
modifiant légèrement
    require("foreign") #Ajout
    trim <- function(strings) {
    if (trim.factor.names)
        gsub(" +$", "", strings)
    else strings
    }
    rval <-  .Call("do_read_SPSS",file,PACKAGE="foreign")
    vl <- attr(rval, "label.table")
    has.vl <- which(!sapply(vl, is.null))
    for (v in has.vl) {
        nm <- names(vl)[[v]]
        nvalues <- length(na.omit(unique(rval[[nm]])))
        nlabels <- length(vl[[v]])
        #Modification de la condition ci-dessous
        if (use.value.labels && (!is.finite(max.value.labels) ||
            nvalues <= max.value.labels) )
            rval[[nm]] <- factor(rval[[nm]], levels = rev(vl[[v]]),
            labels = rev(trim(names(vl[[v]]))))
```

```
        else attr(rval[[nm]], "value.labels") <- vl[[v]]
    }
    if (to.data.frame) {
        varlab <- attr(rval, "variable.labels")
        rval <- as.data.frame(rval)
        attr(rval, "variable.labels") <- varlab
    }
    rval
}
```

4.26 rename.variables.parameters

```
`rename.variables.parameters` <-
function (data) {
    #Vérifications
    if (length(data$N.parameter)==0) stop("data must contain the N.parameter
column.")
    if (length(data$R.parameter)==0) stop("data must contain the R.parameter
column.")
    if (length(data$U.parameter)==0) stop("data must contain the U.parameter
column.")
    #Récupération des paramètres
    N <- data$N.parameter[1]
    R <- data$R.parameter[1]
    U <- data$U.parameter[1]
    #Suppression des colonnes inutiles
    data$N.parameter <- NULL
    data$R.parameter <- NULL
    data$U.parameter <- NULL
    #Récupération des noms
    noms <- names(data)
    #Modifications des noms des variables produites par estimate.prev
    noms[noms=='est.prevalence'] <- paste('est.prevalence.N',N,'.R',R,'.U',U,sep='')
    noms[noms=='circle.count'] <- paste('circle.count.N',N,'.R',R,'.U',U,sep='')
    noms[noms=='circle.radius'] <- paste('circle.radius.N',N,'.R',R,'.U',U,sep='')
    noms[noms=='circle.nb.clusters'] <-
paste('circle.nb.clusters.N',N,'.R',R,'.U',U,sep='')
    noms[noms=='quality.indicator'] <-
paste('quality.indicator.N',N,'.R',R,'.U',U,sep='')
    #Mise à jour du fichier
    names(data) <- noms
    #Renvoi final
    return(data)
}
```

4.27 verif.urb

```
`verif.urb` <-
```

```
function (clust, conf.level=0.90, add=FALSE, lang="en",
var.clust=c("n","nweight","obs.prevalence","urban.area","city.name"),
urban.area.code="in urban area")
{
    #Vérification langue
    if (lang=="fr") lg <- 2
    else lg <- 1
    #Vérifications
    if (!is.numeric(clust[[var.clust[1]]])) stop(var.clust[1],c(" (number of persons by
cluster) must be numeric."," (nombre de personnes par cluster) doit être
numérique.")[lg])
    if (!is.numeric(clust[[var.clust[2]]])) stop(var.clust[2],c(" (weighted total of
persons by cluster) must be numeric."," (total pondérés du nombre de personnes
par cluster) doit être numérique.")[lg])
    if (!is.numeric(clust[[var.clust[3]]])) stop(var.clust[3],c(" (observed prevalence
by cluster) must be numeric."," (prévalence observée par cluster) doit être
numérique.")[lg])
    if (length(clust[[var.clust[4]]])==0) stop(var.clust[4],c(" (urban area) must be
specified."," (milieu de résidence recodé) doit être spécifiée.")[lg])
    if (length(clust[[var.clust[5]]])==0) stop(var.clust[5],c(" (city name) must be
specified."," (nom des villes) doit être spécifiée.")[lg])
    if (!is.factor(clust[[var.clust[4]]])) clust[[var.clust[4]]] <-
as.factor(clust[[var.clust[4]]])
    if (length(clust[clust[[var.clust[4]]]==urban.area.code,][[var.clust[4]]])==0)
stop(var.clust[4],c(" must contain at least a cluster where "," doit contenir au moins
un cluster tel que ")[lg],var.clust[[4]],c(" is equal to '"," est égal à
'")[lg],urban.area.code,"'.")
    if (!is.character(clust[[var.clust[5]]])) clust[[var.clust[5]]] <-
as.character(clust[[var.clust[5]]])
    if (!is.logical(add)) stop(c("Add must be TRUE or FALSE.","Add doit être TRUE
ou FALSE.")[lg])
    #Préparation des données
    result<-data.frame()
    liste.villes <- levels(as.factor(clust[[var.clust[5]]]))
    #Pour chaque villes
    for (i in 1:length(liste.villes)) {
        city <- liste.villes[i]
        temp <-
clust[clust[[var.clust[4]]]==urban.area.code&clust[[var.clust[5]]]==city,]
        city.nb.cluster <- length(temp[[1]])
        if (city.nb.cluster > 0) {
            city.prevalence <-
weighted.mean(temp[[var.clust[3]]],temp[[var.clust[2]]])
            city.n <- sum(temp[[var.clust[1]]], na.rm=TRUE)
            city.nweight <- sum(temp[[var.clust[2]]], na.rm=TRUE)
            test <- prop.test(city.prevalence*city.n/100,city.n,conf.level=conf.level)
            city.low <- test$conf.int[1]*100
            city.high <- test$conf.int[2]*100
        } else {
            city.prevalence <- NA
            city.n <- 0
            city.nweight <- 0
            city.low <- NA
            city.high <- NA
        }
```

```
        one.result <- data.frame(city,city.nb.cluster,city.n, city.nweight,
city.prevalence, city.low, city.high)
        result <- rbind(result, one.result)
    }
    #Si on rajoute des données
    if (add) {
        result$add.prevalence <- 0
        result$add.n <- 0
        message(c("A window will open. Input additive data in columns
'add.prevalence' and 'add.n' for each city. Then close.\nFor 'add.prevalence', input
data in %. For example, for 0.034=3.4%, input 3.4.","Une fenêtre va s'ouvrir.
Compléter les colonnes 'add.prevalence' et 'add.n' pour chaque ville, puis
fermer.\nPour 'add.prevalence', rentrer les données en %. Par exemple, pour
0,034=3,4%, saisir 3.4.")[lg])
        result <- edit(result)
        result$add.low <- NA
        result$add.high <- NA
        result$p.value.comparison <- NA
        for (i in 1:length(result$city)) {
            if (result$add.n[i]>0) {
                test <-
prop.test(round(result$add.prevalence[i]*result$add.n[i]/100),result$add.n[i],con
f.level=conf.level)
                result$add.low[i] <- test$conf.int[1]*100
                result$add.high[i] <- test$conf.int[2]*100
                matrice <-
matrix(c(round(result$city.prevalence[i]*result$city.n[i]/100),round(result$add.p
revalence[i]*result$add.n[i]/100),result$city.n[i]-
round(result$city.prevalence[i]*result$city.n[i]/100),result$add.n[i]-
round(result$add.prevalence[i]*result$add.n[i]/100)),nc=2)
                test <- fisher.test(matrice, conf.level = conf.level)
                result$p.value.comparison[i] <- test$p.value
            }
        }
    }
    result$conf.level <- conf.level
    #Résultats
    return(result)
}
```

4.28write.boundary.shp

```
`write.boundary.shp` <-
function (boundary, file, country=file) {
  require(maptools)
    boundary <- Polygon(boundary)
    boundary <- Polygons(list(boundary),list(1))
    boundary <- SpatialPolygons(list(boundary))
    data <- data.frame(1, country)
    names(data) <- c("id","name")
    boundary <- SpatialPolygonsDataFrame(boundary, data)
    writePolyShape(boundary,file)
```

```
}
```

4.29write.prev.shp

```
`write.prev.shp` <-
function (x, file, coords=~x+y, check=TRUE, lang='en') {
    require(maptools)
    if (check) x <- check.names(x, lang=lang, size.max=10)
    coordinates(x) <- coords
    writePointsShape(x, file)
}
```

4.30write.txt

```
`write.txt` <-
function (x, file, dec=".") {
    #Cette fonction permet d'écrire un fichier txt délimité pardes tabulations.
    write.table(x, file = file, sep="\t", row.names = FALSE, quote = FALSE, dec =
dec)
    }
```

Annexe 5
Description des différentes fonctions implémentées dans prevR

Package 'prevR'

October 10, 2007

Type Package

Title Prevalence estimation with DHS data - Estimation des prévalences à partir des données EDS

Version 1.1

Date 2007-10-10

Author Joseph LARMARANGE <joseph@larmarange.net> IRD - Centre Muraz with supports of ANRS

Maintainer Joseph LARMARANGE <joseph@larmarange.net>

Depends fields, sp, gstat, maptools

Description Ce package permet d'importer des données de type EDS, de les formater, puis de cartographier les variations spatiales de la prévalence d'un phénomène par estimation de la prévalence de chaque point enquêté et interpolation spatiale. Les résultats peuvent ensuite être exportés vers d'autres logiciels de statistique ou de cartographie. La documentation n'est disponible pour le moment qu'en français.

License CeCILL-C - http://www.cecill.info/

URL http://ceped.cirad.fr/prevR/, http://joseph.larmarange.net/prevR/

R topics documented:

`N.optim` *Propose une valeur optimale pour le paramètre N.*

Description

Calcule et propose une valeur optimale pour le paramètre *N* à partir des caractéristiques de l'échantillon (nombre total d'observations, prévalence globale et nombre de clusters).

Usage

```
N.optim(
   n.total,
   global.prevalence,
   nb.clusters,
   b0 = 14.172,
   b1 = 0.419,
   b2 = -0.361,
   b3 = 0.037
   c = -91.011)
```

Arguments

`n.total`	numeric. Nombre total d'observations valides.
`global.prevalence`	numeric. Prévalence globale de l'échantillon. En pourcents.
`nb.clusters`	numeric. Nombre total de clusters.
`b0`	Paramètre b_0.
`b1`	Paramètre b_1.
`b2`	Paramètre b_2.
`b3`	Paramètre b_3.
`c`	Constante c.

Details

La valeur de $N_{optimal}$ est obtenue à partir de l'équation suivante :

$$N_{optimal} = b_0 * n.total^{b_1} * global.prevalence^{b_2} * nb.clusters^{b_3} + c$$

Les coefficients b_0, b_1, b_2 et b_3 ont été obtenus à partir de la simulation de 24500 enquêtes sur un pays modèle, enquêtes présentant une prévalence globale de 1%, 2%, 5%, 10%, 15%, 30% ou 45%, un effectif total de 5000, 6000, 7200, 8640, 10368, 12442 et 14930, réparti selon 300, 360, 432, 518 ou 622 clusters.

Value

Integer.

References

Joseph LARMARANGE, *Prévalences du VIH en Afrique : validité d'une mesure*, thèse de doctorat en démographie, sous la direction de Benoît FERRY, université Paris Descartes, 2007.
Disponible en ligne sur (http://joseph.larmarange.net/).

Examples

```
data(alicante)

infos.prev(alicante.clust)
N.optim(8000,10.16,401)
```

---

alicante *Données issues d'une simulation d'EDS.*

Description

Ces données sont issues de la simulation d'une Enquête Démographique et de Santé (EDS) sur un pays fictif présentant une prévalence nationale de 10 %. 8000 personnes ont été enquêtées, réparties en 401 clusters. Le pays virtuel a été appelé Alicante, d'où le nom du fichier de données.

Usage

```
data(alicante)
```

Value

Charge 5 objets :

`alicante.bounds`
: data.frame. Frontières d'Alicante. Résultat type de `make.boundary.dcw`.

`alicante.cities`
: data.frame. Principales villes d'Alicante. Résultat type de `make.cities.csv`.

`alicante.clust`
: data.frame. Résultats d'enquêtes. Résultat type de `make.clust.dbf` et de la fonction `calcul.dist.cities`.

`alicante.prev`
: data.frame. Estimation des prévalences à partir des données d'enquêtes. Résultat type de `estimate.prev`.

`alicante.krige`
: SpatialPixelsDataFrame. Interpolations spatiales réalisées à partir des estimations de prévalence. Résultat type de `krige.prev`.

Source

- Joseph Larmarange et al., 2006, 'Cartographier les données des enquêtes démographiques et de santé à partir des coordonnées des zones d'enquête', Chaire Quételet, 29 novembre au 1er décembre 2006, Université Catholique de Louvain, Louvain-la-Neuve, Belgique.
Disponible en ligne à (http://www.uclouvain.be/13881.html).

- Joseph LARMARANGE, *Prévalences du VIH en Afrique : validité d'une mesure*, thèse de doctorat en démographie, sous la direction de Benoît FERRY, université Paris Descartes, 2007.
Disponible en ligne sur (http://joseph.larmarange.net/).

`calcul.dist.cities` *Calcule la distance à la ville la plus proche et recode le milieu de résidence.*

Description

Cette fonction calcule la distance de chaque cluster à la ville la plus proche et recode le milieu de résidence selon l'appartenance ou non à une agglomération urbaine. Un cluster est considéré comme appartenant à une agglomération urbaine s'il est à la fois urbain et situé à une distance inférieure à `dist` de la ville la plus proche.

Usage

```
calcul.dist.cities(
   clust,
   cities,
   dist = 15,
   type = "ask",
   lang = "en",
   var.cities = c("x", "y", "city.name"),
   var.clust = c("x", "y", "residence"),
   urban.code = "Urban",
   dist.func = "rdist.earth",
   miles = FALSE)
```

Arguments

`clust`	data.frame. Une observation par cluster. Typiquement le résultat de la fonction `make.clust.dbf`.
`cities`	data.frame. Une observation par ville. Typiquement le résultat de la fonction `make.cities.csv`.
`dist`	numeric. Distance en dessous de laquelle un cluster urbain est considéré comme appartenant à une agglomération urbaine. Il ne s'agit pas de la taille d'une ville mais d'un paramètre permettant de distinguer les clusters urbains d'une agglomération urbaine des autres clusters urbains. Voir les détails pour l'unité à utiliser.
`type`	character. Prend la valeur `'ask'` ou `'all'`. Voir les détails.
`lang`	character. Permet de choisir la langue des messages utilisateur. `fr` pour le français, `en` pour l'anglais.

var.cities	character vector. Liste spécifiant, dans l'ordre, le nom des variables de cities correspondant à la longitude, la latitude et le nom de chaque ville.
var.clust	character vector. Liste spécifiant, dans l'ordre, le nom des variables de clust correspondant à la longitude, la latitude et le milieu de résidence. Le milieu de résidence doit être de type factor ou character.
urban.code	character. Valeur texte spécifiant comment est codée la valeur *urbain* de la variable *milieu de résidence* dans clust.
dist.func	character. Valeur texte spécifiant le nom de la fonction utilisée pour calculer les distances. Utilisez rdist.earth si vous utilisez des coordonnées longitude/latitude en degrés décimaux. Utilisez rdist pour des distances euclidiennes.
miles	logical. TRUE ou FALSE. Variable transmise à rdist.earth si cette fonction est appelée. Calcul des distances en kilomètres si FALSE et en miles si TRUE.

Details

Si type = 'all', les distances de chaque cluster à l'ensemble des villes spécifiées dans cities sont calculées, puis la ville la plus proche de chaque cluster est sélectionnée. Si type = 'ask', une fenêtre présentant l'ensemble des villes contenues dans cities apparaîtra et vous serez invité à sélectioner les villes que vous souhaitez retenir pour la définition des agglomérations urbaines.

Si vous utilisez la fonction rdist, dist doit être exprimé dans la même unité que les coordonnées des clusters et des villes. Si vos coordonnées sont exprimées en degrés décimaux, utilisez la fonction rdist.earth et précisez avec le paramètre miles si vous souhaitez que les résultats soient exprimés en kilomètres ou en miles. dist devra être exprimé en kilomètres ou en miles selon le cas.

Value

La fonction renvoie clust en lui ajoutant trois variables nommées *dist.city*, *city.name* et *urban.area* (voir les détails ci-dessous). Si clust comporte déjà ces trois variables, les anciennes valeurs sont remplacées par les nouvelles valeurs calculées.

dist.city	numeric. Distance à la ville la plus proche, exprimée avec la même unité que dist.
city.name	character. Nom de la ville la plus proche.
urban.area	factor with 2 levels : *in urban area*, *outside urban area*. Indique si le cluster appartient ou non à l'agglomération urbaine de city.name. Est considéré comme appartenant à une agglomération urbaine tout cluster à la fois urbain et situé à moins de dist de la ville considérée.

Note

Pour un pays de taille moyenne, 15 kilomètres est souvent un bon compromis pour dist. Pour le choix des agglomérations urbaines, voir le tutoriel de prevR ('tutoriel.prevR.pdf').

See Also

verif.urb et infos.prev pour obtenir des statistiques sur les fichiers de données.
map.clust et map.cities pour réaliser des cartes.

Examples

```
## Chargement des données
data(alicante)

## Carte des villes de plus de 100.000 habitants
map.cities(alicante.cities, alicante.bounds, min.population=100000,
   new.window=FALSE, main='Main cities of Alicante')

## Carte des clusters par milieu de résidence
x11()
map.clust(alicante.clust, alicante.bounds, type='urb', new.window=FALSE,
   main='Clusters by type of residence')

## Sélection des villes de plus de 100.000 habitants
main.cities <- alicante.cities[alicante.cities$population > 100000, ]

## Calcul des distances à la ville de plus de 100.000 habitants
## la plus proche et vérification des résulats
alicante.clust <- calcul.dist.cities(alicante.clust, main.cities,
   type='all', dist=25)
verif.urb(alicante.clust)

## Sélection des villes de plus de 250.000 habitants
main.cities <- alicante.cities[alicante.cities$population > 250000, ]

## Calcul des distances à la ville de plus de 250.000 habitants
## la plus proche et vérification des résulats
alicante.clust <- calcul.dist.cities(alicante.clust, main.cities,
   type='all', dist=25)
verif.urb(alicante.clust)
infos.prev(alicante.clust)

## Carte des clusters selon leur appartenance à une agglomération urbaine
x11()
map.clust(alicante.clust, alicante.bounds, type='urb', new.window=FALSE,
   var.urb='urban.area', inverse=TRUE,
        main='Clusters in urban agglomeration')
```

`check.names` *Permet de renommer et de supprimer les colonnes d'un tableau de données.*

Description

Cette fonction vérifie si les noms de colonnes d'un tableau de données dépassent une certaine longueur. Si c'est le cas, l'utilisateur est invité à renommer les noms de colonnes du tableau de données. Il peut également supprimer certaines colonnes par la même occasion. Cette fonction est particulièrement indiquée avant d'exporter des données, notamment aux formatx *dbf* et *shapefile* dans la mesure où les noms de colonnes ne doivent pas dépasser les 10 caractères pour ce type de fichier.

Usage

```
check.names(data, lang='en', size.max = 10)
```

Arguments

`data`	data.frame. Dont on veut vérifier les noms de colonnes.
`lang`	character. Permet de choisir la langue des messages utilisateur. `fr` pour le français, `en` pour l'anglais.
`size.max`	character vector. Nombre maximum de caractères autorisés pour le nom des variables.

Value

Renvoie `data` après avoir renommé les différentes colonnes. Si une colonne a été renommée NULL, alors elle sera supprimée.

Si tous les noms de colonnes de `data` ont une longueur inférieure ou égale à `size.max`, alors `data` est directement renvoyé tel quel par la fonction.

Note

Si, après une saisie par l'utilisateur de nouveaux noms, certains ont toujours une longueur supérieure à `size.max`, alors l'utilisateur sera contraint de modifier à nouveau les noms de colonnes concernés. Afin de repérer facilement les noms de colonnes trop longs, la longueur (en nombre de caractères) de chaque nom est indiquée dans une colonne *size*.

Cette fonction est notamment appelée par `write.prev.shp`.

Examples

```
## Not run:
data(alicante)

alicante.clust.check <- check.names(alicante.clust)
write.dbf(alicante.clust.check, 'alicante_clust.dbf')
## End(Not run)
```

`estimate.prev` *Estime la prévalence de chaque cluster par la méthode des cercles.*

Description

Estime la prévalence de chaque cluster par la méthode des cercles selon trois paramètre *N* (effectif minimum), *R* (rayon maximum) et *U* (prise en compte de l'appartenance à une agglomération urbaine). Plusieurs estimations peuvent être réalisées simultanément, une pour chaque combinaison des paramètres *N*, *R* et *U*.

Usage

```
estimate.prev(
   clust,
   N = seq(100, 500, 50),
   R = Inf,
   U = FALSE,
   dist.func = "rdist.earth",
   miles = FALSE,
```

```
  lang = "en",
  progression = TRUE,
  merge.result = FALSE,
  var.clust = c("x", "y", "n", "nweight", "obs.prevalence",
                                    "urban.area", "city.name"),
  urban.area.code = "in urban area")
```

Arguments

clust	data.frame. Une observation (cluster) par ligne. Typiquement le résultat de la fonction make.clust.dbf. Si le paramètre *U* est utilisé, ce fichier doit comprendre également la distance à la ville, le nom de la ville la plus proche et l'appartenance ou non à une agglomération urbaine, typiquement le résultat de la fonction calcul.dist.cities.
N	integer vector. Entier ou liste d'entiers représentant l'effectif minimum des cercles. Voir les fonctions c et seq pour la génération de listes d'entiers. Si l'on ne souhaite pas utiliser le paramètre *N*, attribuez à N la valeur Inf.
R	integer vector. Entier ou liste d'entiers représentant le rayon maximum des cercles. Voir les fonctions c et seq pour la génération de listes d'entiers. Si l'on ne souhaite pas utiliser le paramètre *R*, attribuez à R la valeur Inf.
U	logical or integer. Spécifie si l'appartenance à une agglomération urbaine doit être prise en compte. TRUE pour la prise en compte, FALSE pour la non prise en compte, 2 si l'on souhaite réaliser les estimations avec et sans prise en compte de ce paramètre.
dist.func	character. Nom de la fonction utilisée pour calculer les distances. Utilisez rdist.earth si vous utilisez des coordonnées longitude/latitude en degrés décimaux. Utilisez rdist pour des distances euclidiennes.
miles	logical. Transmise à rdist.earth si cette fonction est appelée. Calcul des distances en kilomètres si FALSE et en miles si TRUE.
lang	character. Permet de choisir la langue des messages utilisateur. fr pour le français, en pour l'anglais.
progression	logical. Si TRUE, affiche la progression du calcul.
merge.result	logical. Si TRUE, applique la fonction merge.prev aux résultats. Voir *Value*.
var.clust	character vector. Liste correspondant aux noms des différentes variables de clust. Dans l'ordre : longitude, latitude, effectif, effectif pondéré, prévalence observée, appartenance à une agglomération urbaine et nom de la ville la plus proche. Les deux derniers noms n'ont pas besoin d'être précisés si U = FALSE. Les noms par défaut correspondent aux noms de variables produits par make.clust.dbf et calcul.dist.cities.
urban.area.code	character. Valeur texte spécifiant comment est codée la valeur *appartient à une agglomération urbaine* de la variable *appartenance à une agglomération urbaine* dans clust. La valeur par défaut correspond à une sortie de calcul.dist.cities.

Details

Estime la prévalence de chaque cluster pour chaque combinaison des paramètres *N*, *R* et *U*. Pour une combinaison de paramètres, la fonction calcule la distance du cluster à l'ensemble des autres clusters, puis trie les clusters par distance croissante. Si *U* est requis et que le cluster appartient à

une agglomération urbaine, seuls les clusters de la même agglomération urbaine sont sélectionnés. Si *U* est requis et que le cluster n'appartient pas à une agglomération urbaine, seuls les clusters hors agglomération urbaine sont sélectionnés. Si *N* est spécifié (différent de `Inf`), seuls les clusters les plus proches sont sélectionnés de manière à ce que le nombre total de personnes enquêtées (variable *n*) soit au moins égal à *N*. Si *R* est spécifié (différent de `Inf`), seuls les clusters situés à une distance inférieure à *R* sont retenus. La prévalence du cluster est alors estimée sur l'ensemble des clusters sélectionnés en tenant compte de la pondération de chaque cluster (variable *nweight*).

Pour plus de détails, voir le tutoriel de prevR ('tutoriel.prevR.pdf').

Value

Renvoie `clust` en lui ajoutant 8 nouvelles variables, décrites ci-dessous. S'il y a plusieurs combinaisons des trois paramètres *N*, *R* et *U*, `clust` est répété autant de fois qu'il y a de combinaisons.

`est.prevalence`
: Prévalence estimée.

`circle.count`
: Effectif total sur lequel l'estimation a été effectuée.

`circle.radius`
: Rayon du cercle de lissage dans lequel sont contenus les clusters retenus pour l'estimation.

`circle.nb.clusters`
: Nombre de clusters retenus pour l'estimation.

`quality.indicator`
: Indicateur de qualité de l'estimation. Il est obtenu pour chaque cluster selon l'équation suivante :

$$quality.indicator = \frac{circle.radius^2}{\sqrt{circle.count}}$$

Plus la valeur de cet indicateur est élevée, plus l'estimation est incertaine.

`N.parameter`
: Valeur du paramètre *N* pour cette estimation. `Inf` si le paramètre n'a pas été appliqué.

`R.parameter`
: Valeur du paramètre *R* pour cette estimation. `Inf` si le paramètre n'a pas été appliqué.

`U.parameter`
: 0 si le paramètre *U* n'a pas été appliqué. Le nombre d'agglomérations urbaines retenues sinon.

Si `merge.result = TRUE`, `clust` est renvoyé en lui ajoutant autant de fois qu'il y a de combinaisons des trois paramètres, les variables *est.prevalence*, *circle.count*, *circle.radius*, *circle.nb.clusters* et *quality.indicator*. Les noms de ces variables prennent alors comme suffixe :.*Nvaleur-de-n.Rvaleur-de-r.Uvaleur-de-u* (voir `merge.prev`).

Warning

Le temps de calcul de cette fonction peut prendre plusieurs minutes selon la puissance de votre machine. Soyez donc patient.

Note

Pour plus d'informations sur le choix des paramètres, voir le tutoriel de prevR ('tutoriel.prevR.pdf'). La valeur optimale de *N* peut être déterminée à partir du nombre total de personnes testées, du nombre total de clusters et de la prévalence globale. Voir `infos.prev` et `N.optim`. Un bon

compromis pour choisir R consiste à retenir la valeur du neuvième décile de *circle.radius* lorsque seul le paramètre *N* est appliqué avec sa valeur optimale. Voir `infos.prev`.

Il est possible, si `merge.result = FALSE`, d'extraire du data.frame renvoyé les éléments correspondant à une seule combinaison des trois paramètres à l'aide de la fonction `extract.data`.

Les résultats peuvent être exportés au format *texte tabulé* à l'aide de `write.txt`, au format *dbf* à l'aide de `write.dbf` ou encore au format *shape* pour importation dans un logiciel de cartographie à l'aide de `write.prev.shp`.

References

- Joseph Larmarange et al., 2006, 'Cartographier les données des enquêtes démographiques et de santé à partir des coordonnées des zones d'enquête', Chaire Quételet, 29 novembre au 1er décembre 2006, Université Catholique de Louvain, Louvain-la-Neuve, Belgique.
Disponible en ligne à (http://www.uclouvain.be/13881.html).

- Joseph LARMARANGE, *Prévalences du VIH en Afrique : validité d'une mesure*, thèse de doctorat en démographie, sous la direction de Benoît FERRY, université Paris Descartes, 2007.
Disponible en ligne sur (http://joseph.larmarange.net/).

See Also

`infos.prev` pour obtenir des infos sur le résultat produit, `extract.data` et `merge.prev` pour manipuler les résultats, `write.prev.shp`, `write.txt` et `write.dbf` pour exporter les résultats, `krige.prev` pour réaliser une interpolation spatiale.

Examples

```
data(alicante)

#  Premier lissage - juste selon N,
## pour les valeurs 100, 150, 200, 250, 300, 350 et 400

alicante.prev <- estimate.prev(
        alicante.clust,
        N=seq(100, 400, 50),
        R=Inf, U=FALSE)
str(alicante.prev)

alicante.prev.n250 <- extract.data(
        alicante.prev,
        value=c(250, Inf, 0))
str(alicante.prev.n250)
infos.prev(alicante.prev.n250)

str(merge.prev(alicante.prev))

# Second lissage - En utilisant les trois paramètres

alicante.prev <- estimate.prev(
        alicante.clust,
        N=c(250),
        R=c(128,Inf),
        U=2,
        merge.result=TRUE)
str(alicante.prev)
```

`extract.col` *Extrait les colonnes d'un data.frame selon une condition sur leur nom.*

Description

Cette fonction permet d'extraire d'un data.frame les colonnes dont le nom contient une certaine chaîne de caractère.
Typiquement, cette fonction est pratique pour afficher les résultats de `krige.prev` avec `spplot` (voir exemple).

Usage

```
extract.col(
   x,
   value='.pred'
```

Arguments

`x`	data.frame ou SpatialPixelsDataFrame. Dont on veut extraire des colonnes.
`value`	character. Chaîne de caractères recherchées dans le nom des colonnes.

Value

Renvoie `data` avec uniquement les colonnes dont le nom contient `value`.

Examples

```
data(alicante)

spplot(extract.col(alicante.krige))
```

`extract.data` *Extrait les observations d'un data.frame selon une ou plusieurs conditions.*

Description

Cette fonction permet d'extraire d'un data.frame les observations dont les valeurs à une ou plusieurs variables données sont égales à des valeurs passées en paramètre de la fonction.
Typiquement, cette fonction permet de sélectionner les observations correspondant à une combinaison précise des paramètres *N*, *R* et *U* d'un résultat de la fonction `estimate.prev` appelée avec `merge.result = FALSE`.

Usage

```
extract.data(
   data,
   value = "ask",
   lang = "en",
   var.data = c("N.parameter", "R.parameter", "U.parameter"))
```

Arguments

data	data.frame. Dont on veut extraire des observations.
value	character or vector. Si value="ask", la fonction liste les différentes valeurs présentes dans le fichier et propose à l'utilisateur de choisir celles qui lui conviennent. Sinon, liste des valeurs pour lesquelles on sélectionne les observations. Doit alors être de même longueur que var.data
lang	character. Permet de choisir la langue des messages utilisateur. fr pour le français, en pour l'anglais.
var.data	character vector. Liste des variables sur lesquelles porte la sélection.

Details

Seules sont sélectionnées les observations dont les valeurs aux variables var.data correspondent à value. Voir le tutoriel de prevR ('tutoriel.prevR.pdf') pour une démonstration avec value="ask".

Value

Renvoie data avec uniquement les observations sélectionnées.

Note

Cette fonction sert typiquement à sélectionner, dans les résultats de estimate.prev avec merge.result = FALSE, les observations correspondant à une combinaison donnée des trois paramètres.

Elle peut également servir à sélectionner, par exemple, les clusters ruraux ou bien les clusters appartenant à une agglomération donnée. Voir les exemples ci-dessous.

See Also

[.data.frame.

Examples

```
data(alicante)

## Extraction à partir d'un résultat de estimate.prev

alicante.prev <- estimate.prev(
        alicante.clust,
        N=seq(200, 300, 50),
        R=Inf,
        U=FALSE)
str(alicante.prev)
alicante.prev.n250 <- extract.data(
        alicante.prev,
```

```
        value=c(250, Inf, 0))
str(alicante.prev.n250)

## Sélection des clusters ruraux

rur <- extract.data(
        alicante.clust,
        value='Rural',
        var.data='residence')
str(rur)

## Sélection des clusters appartenant à l'agglomération urbaine de la ville A

urb.A <- extract.data(
        alicante.clust,
        value=c('in urban area','A'),
        var.data=c('urban.area','city.name'))
str(urb.A)
```

`infos.prev` *Fournit des informations à propos d'un data.frame de type clust ou prev.*

Description

Affiche le nombre de clusters, d'observations, la prévalence globale, une valeur optimale pour N, le nombre d'agglomérations urbaines et leurs noms, ainsi que les quantiles des rayons des cercles de lissage.

Usage

```
infos.prev(
   prev,
   lang = "en",
   var.n = "n",
   var.nweight = "nweight",
   var.obs.prevalence = "obs.prevalence",
   var.city.name = "city.name",
   var.circle.radius = "circle.radius")
```

Arguments

`prev`	data.frame. Chaque ligne doit correspondre à un cluster.
`lang`	character. Permet de choisir la langue des messages utilisateur. `fr` pour le français, `en` pour l'anglais.
`var.n`	character. Nom de la variable correspondanta aux effectifs observés dans chaque cluster.
`var.nweight`	character. Nom de la variable correspondants aux effectifs pondérés de chaque cluster.

`var.obs.prevalence`
: character. Nom de la variable correspondant à la prévalance observée dans chaque cluster.

`var.city.name`
: character. Optionnel. Nom de la variable correspondant au nom de la ville la plus proche.

`var.circle.radius`
: character. Optionnel. Nom de la variable correspondant aux rayons des cercles de lissage.

Details

Cette fonction affiche les renseignements suivants :

- **nombre de clusters** correspond au nombre de lignes dans `data`.
- **nombre d'observations valides** correspond à la somme de la colonne définie par `var.n`.
- **prévalence globale** correspondant à la moyenne de la colonne définie par `var.obs.prevalence`, pondérée par la colonne définie par `var.nweight`.
- **valeur optimale proposée pour *N*** calculée à partir de la fonction `N.optim`.
- **nombre et nom des villes** si une colonne valide est entrée pour `var.city.name`, le nombre de villes trouvées ainsi que leurs noms.
- **quantile du rayon des cercles** si une colonne valide est entrée pour `var.circle.radius`, les quantiles à 50%, 75%, 80%, 85%, 90%, 95% et 99% de cette colonne.

Value

Cette fonction renvoie la valeur `NULL`.

Warning

Si vous souhaitez appliquer cette fonction à un résultat obtenu avec `estimate.prev` appelée avec plusieurs combinaisons des trois paramètres et `merge.result = FALSE`, utilisez `extract.data` ou `merge.prev` avant d'utiliser `infos.prev`.

See Also

`N.optim` pour les détails concernant le calcul d'une valeur optimale de N.

Examples

```
data(alicante)

infos.prev(alicante.clust)

infos.prev(alicante.prev,var.circle.radius='circle.radius.N250.RInf.U6')
```

`krige.prev` *Réalise des interpolations spatiales par krigeage et/ou selon une pondération inverse à la distance.*

Description

Cette fonction met en forme les données contenues dans un data.frame et permet de réaliser des interpolations spatiales sur une grille régulière en ayant recours aux fonctions du package **gstat**. Elle permet également de guider l'utilisateur pas à pas pour choisir le modèle de variogramme utilisé.

Usage

```
krige.prev(
   data,
   formula = est.prevalence ~ 1,
   locations = ~x + y,
   type = "ask",
   boundary = NULL,
   cell.size = 0.05,
   ask.cell.size = TRUE,
   lang = "en",
   model = vgm(1, "Exp", 1),
   idp = 2,
   show.variogram = TRUE, ...)
```

Arguments

`data`	data.frame. Chaque ligne doit correspondre à une observation ponctuelle. ATTENTION : deux observations ne peuvent avoir les mêmes coordonnées.
`formula`	formula or formulae list. Formule ou liste de formules spécifiant les interpolations à réaliser. Chaque formule définit la variable dépendante comme un modèle linéaire de variables indépendantes. Pour un krigeage ordinaire ou une interpolation selon l'inverse de la distance, utilisez une formule du type `z~1`. Voir `krige` pour plus de détails.
`locations`	formula. Spécifie sous la forme d'une formule les colonnes de `data` correspondant aux coordonnées géographiques des observations. Si, par exemple, la longitude correspond à la colonne `x` et la latitude à la colonne `y`, entrez `~x+y`.
`type`	character. Peut prendre les valeurs *ask*, *auto*, *model* ou *idw*. Voir les détails.
`boundary`	data.frame. Optionnel. Frontières de la zone géographique étudiée, sous la forme d'une série de points dessinant un polygone fermé. Les coordonnées des points doivent être inscrites dans des colonnes ayant les mêmes noms que ceux définis dans `locations` ou bien nommées *x* et *y*. Les unités doivent évidemment être identiques.
`cell.size`	numeric. Permet de définir la taille de chaque cellule de la grille sur laquelle l'interpolation spatiale sera réalisée. S'exprime dans la même unité que celle utilisée pour les coordonnées des points de `data`.
`ask.cell.size`	logical. Si `TRUE`, calcule et affiche la taille de la grille engendrée par la valeur de `cell.size` et permet à l'utilisateur de modifier cette dernière.

`lang` character. Permet de choisir la langue des messages utilisateur. `fr` pour le français, `en` pour l'anglais.

`model` object or list of objects of class `variogramModel`. Typiquement un résultat de la fonction `vgm`. S'il s'agit d'une liste de modèles de variogramme, elle doit avoir la même longueur que `formula`. Pour générer cette liste, ayez recours à la fonction `list`. Le recours à la fonction `c` provoquerait une erreur. Les différents types de modèles de variogrammes peuvent être visualisés à l'aide de la fonction `show.vgms`. Non requis si `type = 'idw'`. Voir les détails.

`idp` numeric or numeric vector. Requis uniquement si `type = 'idw'`. Spécifie la puissance de la pondération selon l'inverse de la distance. Voir `idw`. S'il s'agit d'un vecteur, il doit avoir la même longueur que `formula`.

`show.variogram` logical. Indique si les graphiques des variogrammes empiriques et des modèles utilisés doivent être affichés à la fin du calcul. Voir les exemples. Sans effet si `type = 'idw'`.

`...` Autres arguments passés à chaque appel de `krige` ou d'`idw` selon le cas. Permet de réaliser des interpolations locales et/ou par blocs. Voir la documentation de ces fonctions.

Details

Si `type='model'`, une interpolation spatiale par krigeage est réalisée pour chaque élément de `formula` en utilisant le même modèle de variogramme pour tous si un seul modèle est fourni dans `model`. Utilise chaque modèle respectivement s'il s'agit d'une liste de modèles de variogrammes.

Si `type='auto'`, les interpolations spatiales par krigeage sont réalisées pour chaque élément de `formula` en utilisant les modèles obtenus par la fonction `fit.variogram` qui permet d'ajuster un modèle de variogramme à un variogramme expérimental. Les valeurs de `model` sont passées à `fit.variogram` et servent de valeurs de départ pour l'ajustement. Pour plus de détails, voir l'aide de cette fonction.

Si `type='ask'`, pour chaque élément de `formula`, un modèle ajusté est calculé à partir du variogramme expérimental, `model` servant de valeurs de départ. Le variogramme expérimental et le modèle ajusté sont ensuite représentés graphiquement. Une invite utilisateur permet d'accepter le modèle ajusté ou bien d'effectuer un ajustement manuel du modèle. Voir le tutoriel de prevR ('tutoriel.prevR.pdf') pour des exemples.

Si `type='idw'`, une interpolation spatiale par pondération selon l'inverse de la distance est réalisée pour chaque élément de `formula`, les valeurs de `idp` étant passées à la fonction `idw`.

Value

Objet de la classe `SpatialPixelsDataFrame`. Pour chaque variable interpolée, deux variables sont produites : la prédiction de l'interpolation et la variance de cette prédiction (krigeage uniquement). Elles portent le nom de la variable suivi respectivement des suffixes *.pred* et *.var*. Si `boundary` a été spécifié, les limites de la grille des résultats correspondent aux limites de la frontière et les points situés en-dehors du polygone défini par `boundary` sont attribués d'une valeur manquante `NA`. Sinon, les limites de la grille correspondent aux coordonnées limites des points contenus dans `data`.

Warning

Le temps de calcul de cette fonction peut prendre plusieurs minutes selon la puissance de votre machine. Soyez donc patient.

Note

Les résultats de cette fonction peuvent être cartographiés à l'aide la fonction `spplot` et exportés au format *asc* pour importation dans un logiciel de cartographie à l'aide de la fonction `write.asciigrid`.Voir les exemples.

Pour plus de détails, notamment sur l'utilisation de la fonction lorsque `type='ask'` et le choix d'un modèle de variogramme, voir le tutoriel de prevR ('tutoriel.prevR.pdf').

See Also

`krige` pour les détails concernant le krigeage, `idw` pour l'interpolation selon l'inverse de la distance, `write.asciigrid` pour l'exportation des résultats, `spplot` pour représenter graphiquement les résultats, `vgm` pour spécifier un modèle de variogramme, `show.vgms` pour afficher les différents types de modèles de variogrammes et `extract.col` pour extraire certaines colonnes.

Examples

```
data(alicante)

## Krigeage

show.vgms()
alicante.krige <- krige.prev(alicante.prev, boundary=alicante.bounds,
        formula=c(est.prevalence.N100.RInf.U0~1,
                est.prevalence.N250.RInf.U0~1,
                est.prevalence.N250.R128.U0~1,
                est.prevalence.N250.R128.U6~1,
                circle.radius.N250.R128.U6~1,
                quality.indicator.N250.R128.U6~1),
        model=list(vgm(53.58,'Exp',0.56),
                vgm(48.97,'Exp',1.06),
                vgm(49.19,'Exp',1.07),
                vgm(42.47,'Exp',0.58),
                vgm(1056,'Exp',0.89),
                vgm(150000,'Exp',2.1)),
        type='model', ask.cell.size=FALSE,
                cell.size=0.075, show.variogram=FALSE)

## Représentation graphique

spplot(extract.col(alicante.krige))

spplot(alicante.krige,
        zcol = c('est.prevalence.N100.RInf.U0.pred',
                'est.prevalence.N250.RInf.U0.pred',
                'est.prevalence.N250.R128.U0.pred',
                'est.prevalence.N250.R128.U6.pred'),
        col.regions=prevR.colors.red(21),
        cuts=20,
        main='Estimated prevalence with several parameters')
x11()
spplot(alicante.krige,
        zcol = 'circle.radius.N250.R128.U6.pred',
        col.regions=prevR.colors.blue(21),
        cuts=20,
        main='Radius of circles - N=250 R=128 U=6')
```

```
x11()
spplot(alicante.krige,
        zcol = 'quality.indicator.N250.R128.U6.pred',
        col.regions=prevR.colors.green.inverse(21),
        cuts=20,
        main='Quality indicator - N=250 R=128 U=6')

## Exportation au format asc

## Not run:
write.asciigrid(alicante.krige, 'alicante-krige.asc')
## End(Not run)
```

`make.boundary.dcw` *Extrait les coordonnées des frontières d'un pays à partir d'un fichier de points du DCW.*

Description

Permet d'importer les coordonnées longitude/latitude des frontières d'un pays à partir d'un fichier de points fourni par le Digital Chart of the World (`http://www.maproom.psu.edu/dcw/`).

Usage

```
make.boundary.dcw(file, progression = TRUE, lang = "en")
```

Arguments

`file`	character. Nom du fichier texte récupéré par le Digital Chart of the World. Typiquement, ces fichiers sont nommés *country-name2pts.txt*.
`progression`	logical. Permet d'afficher des messages indiquant la progression de l'analyse du fichier.
`lang`	character. Permet de choisir la langue des messages utilisateur. `fr` pour le français, `en` pour l'anglais.

Details

Lit le fichier texte. Si celui-ci contient les coordonnées de plusieurs polygones, les affiche et demande à l'utilisateur de choisir le polygone principal.

Value

data.frame avec deux colonnes, *x* et *y*.

Note

La lecture ligne à ligne du fichier texte peut être longue. Soyez patient.

Pour un exemple concret, voir le tutoriel de prevR ('tutoriel.prevR.pdf').

Examples

```
## Not run:

burkina.bounds <- make.boundary.dcw('burkina_faso2pts.txt')

## End(Not run)
```

`make.cities.csv` *Génère un data.frame avec les coordonnées de villes à partir d'un fichier csv.*

Description

Importe un fichier csv, liste les différentes colonnes et demande à l'utilisateur de sélectionner les données correspondant à la longitude, la latitude, le nom et la population de chaque ville. Permet typiquement d'importer des données à partir du *Global Rural - Urban Mapping Project (GRUMP)*, disponibles gratuitement à http://sedac.ciesin.columbia.edu/gpw/.

Usage

```
make.cities.csv(file, lang = "en")
```

Arguments

`file`	character. Nom du fichier csv à importer. Les données doivent être séparées par des virgules, et le séparateur de décimales être le point.
`lang`	character. Permet de choisir la langue des messages utilisateur. `fr` pour le français, `en` pour l'anglais.

Details

Une interface textuelle guide l'utilisateur.

Value

Renvoie un data.frame avec 4 colonnes :

`city.name`	Nom de la ville
`x`	Longitude du centre de la ville
`y`	Latitude du centre de la ville
`population`	Population de la ville

Les données sont triées par population décroissante

Note

Pour un exemple concret, voir le tutoriel de prevR ('tutoriel.prevR.pdf').

Examples

```
## Not run:
burkina.cities <- make.cities.csv('bfapv1.csv')
## End(Not run)
```

`make.clust.dbf` *Lit et met en forme les données d'EDS.*

Description

Cette fonction lit les données GPS d'une Enquête Démographique et de Santé (EDS), fournies sous la forme d'un fichier *dbf* sur `http://www.measuredhs.com`, puis agrège les données individuelles préparées avec `make.ind.spss` pour produire un data.frame exploitable par prevR.

Usage

```
make.clust.dbf(file, ind, lang = "en")
```

Arguments

`file`	character. Nom du fichier *dbf* comportant les coordonnées de chaque cluster.
`ind`	data.frame. Données individuelles préparées avec `make.ind.spss`.
`lang`	character. Permet de choisir la langue de l'interface utilisateur. `fr` pour le français, `en` pour l'anglais.

Details

Lit le fichier `file` et demande à l'utilisateur les variables suivantes : *numéro du cluster*, *longitude*, *latitude*, *milieu de résidence*, *code de la région* et *nom des régions*. La fonction prépare les données sous forme d'un data.frame et agrège les données de `ind` pour chaque cluster. Si les variables *sex* et *age* sont présentes dans `ind`, il est possible de restreindre l'analyse à une tranche d'âges et/ou à un sexe donné.

Value

Renvoie un data.frame, où chaque ligne correspond à un cluster, avec les colonnes suivantes :

`cluster`	Identifiant du cluster.
`x`	Longitude.
`y`	Latitude.
`residence`	Milieu de résidece.
`region`	Code de la région.
`region.name`	Nom de la région.
`n`	Effectif, soit le nombre de personnes observées dans le cluster (résultats indéterminés et personnes ayant une pondération nulle exclue).
`nweight`	Effectif pondéré, soit la somme des poids de chaque individu du cluster.
`obs.prevalence`	Prévalence observée dans le cluster, en tenant compte des pondérations. Est exprimée en pourcents.

Note

Voir le tutoriel de prevR ('tutoriel.prevR.pdf') pour plus de renseignements.

See Also

`make.ind.spss`.

Examples

```
## Not run:
burkina.ind <- make.ind.spss('burkina_hiv.sav', lang='fr')
burkina.clust <- make.clust.dbf('burkina_gps.dbf', burkina.ind, lang='fr')
## End(Not run)
```

`make.ind.spss` *Prépare les données individuelles d'une EDS.*

Description

Cette fonction lit les données individuelles d'une Enquête Démographique et de Santé (EDS), fournies sous la forme d'un fichier *SPSS* sur `http://www.measuredhs.com`, pour produire un data.frame exploitable par *make.clust.dbf*.

Usage

```
make.ind.spss(file, lang = "en")
```

Arguments

`file`	character. Nom du fichier de données individuelles. Doit être au format *sav* de SPSS.
`lang`	character. Permet de choisir la langue de l'interface utilisateur. `fr` pour le français, `en` pour l'anglais.

Details

Lit le fichier `file` et demande à l'utilisateur les variables suivantes : *identifiant des individus*, *numéro du cluster de l'individu*, *âge*, *sexe*, *variable analysée* et *poids statistique de chaque individu*. Si *numéro du cluster de l'individu* n'est pas renseigné, il sera calculé à partir de l'*identifiant*.

L'utilisateur sera invité à spécifier les codes de la *variable analysée* correspondant à un résultat positif, un résultat négatif et à un résultat indéterminé.

âge et *sexe* sont optionnels. Il s'agit de deux variables, respectivement numérique et qualitative, pouvant être utilisées par `make.clust.dbf` pour restreindre l'analyse à une sous-population.

Si la *pondération statistique* n'est pas précisée, chaque individu aura un poids égal à 1. Si cette variable est spécifiée, il est possible de lui appliquer un facteur multiplicatif. Souvent, dans les EDS, la pondération est indiquée dans une variable qui doit être divisée par 1 000 000.

Value

Renvoie un data.frame, où chaque ligne correspond à un individu, avec les colonnes suivantes :

`id`	Identifiant de l'individu.
`cluster`	Numéro du cluster.
`age`	Variable de type *integer*.

`sex` Variable de type *factor*.

`original.result`
Valeurs de la variable analysée telles qu'enregistrées dans `file`.

`weight` Poids statistique de chaque individu.

`result` Variable analysée recodée en *Negative*, *Positive* ou *NA*.

Si certaines variables n'ont pas été spécifiées par l'utilisateur, les colonnes correspondantes seront inexistantes dans le data.frame retourné.

Note

Voir le tutoriel de prevR ('tutoriel.prevR.pdf') pour plus de renseignements.

See Also

`make.clust.dbf`.

Examples

```
## Not run:
burkina.ind <- make.ind.spss('burkina_hiv.sav', lang='fr')
burkina.clust <- make.clust.dbf('burkina_gps.dbf', burkina.ind, lang='fr')
## End(Not run)
```

---

map.cities *Représente les villes contenues dans un data.frame sur une carte.*

Description

Permet de représenter sur une carte les points contenus dans un data.frame ainsi que le nom de chaque point.

Usage

```
map.cities(
   cities,
   boundary,
   var.cities = c("x", "y", "city.name", "population"),
   min.population = NULL,
   new.window = TRUE,
   ...)
```

Arguments

`cities` data.frame. Contient les villes/points à représenter.

`boundary` data.frame. Contient deux colonnes, *x* et *y*. Série de points définisant un polygone correspondant aux limites de la zone étudiée.

`var.cities` character vector. Liste des noms des variabales de `cities` correspondant, dans l'ordre, à la longitude, la latitude, le nom et la population associés à chaque ville/point. Si `min.population = NULL`, la quatrième valeur de `var.cities` n'est pas requise.

`min.population`
: numeric. Si précisé, seules les villes ayant une population au moins égale à cette valeur seront représentées.

`new.window`
: logical. Indique si la carte doit être effectuée dans une nouvelle fenêtre.

`...`
: Paramètres supplémentaires passés à la fonction `title`. Voir l'aide de cette fonction.

Details

Dessine le polygone défini par `boundary`, puis les différents points sélectionnés de `cities` et ajoute leur nom à côté de chacun d'eux.

Examples

```
data(alicante)

map.cities(
        alicante.cities,
        alicante.bounds,
        main='Cities of Alicante')

map.cities(
        alicante.cities,
        alicante.bounds,
        min.population=100000,
        main='Main cities of Alicante')
```

---

map.clust *Permet de cartographier les zones d'enquêtes ou clusters.*

Description

Affiche sur une carte les clusters enquêtés selon leur milieu de résidence, le nombre de personnes enquêtées ou selon le nombre de personnes présentant la caractéristique étudiée (nombre de cas positifs).

Usage

```
map.clust(
   clust,
   boundary,
   type = "urb",
   lang = "en",
   var.urb = "residence",
   var.n = "n",
   var.obs.prevalence = "obs.prevalence",
   var.coords = c("x", "y"),
   inverse = FALSE,
   add.legend = TRUE,
   legend.location = "bottomright",
   factor.size = 0.2,
   new.window = TRUE,
   ...)
```

Arguments

clust	data.frame. Il doit comporter une ligne par cluster. Typiquement le résultat de make.clust.dbf.
boundary	data.frame. Doit contenir deux colonnes, *x* et *y*. Série de points définissant un polygone correspondant aux limites de la zone étudiée.
type	character. Peut prendre les valeurs *urb*, *flower* ou *count*. Voir les détails.
lang	character. Permet de choisir la langue de l'interface utilisateur. fr pour le français, en pour l'anglais.
var.urb	character. Nom de la colonne de clust utilisée pour représenter le type des clusters si type= 'urb'. Non requise sinon.
var.n	character. Nom de la colonne de clust utilisée pour représenter la taille des clusters si type= 'count'. Requise également si type= 'flower'.
var.obs.prevalence	character. Nom de la colonne de clust correspondant à la prévalence observée de chaque cluster. Requise uniquement si type= 'flower'.
var.coords	character vector. Liste des noms des deux variables de clust correspondant à la longitude et la latitude de chaque cluster.
inverse	logical. Utilisé si type= 'urb'. Si TRUE, inverse l'ordre des modalités. Voir détails.
add.legend	logical. Indique si une légende doit être affichée.
legend.location	Position de la légende. Variable transmise à legend. Voir la section *Details* de l'aide de cette fonction. Le plus simple consiste à utiliser les termes *bottomright*, *bottom*, *bottomleft*, *left*, *topleft*, *top*, *topright*, *right* and *center*.
factor.size	numeric. Utilisé si type= 'count'. Facteur d'agrandissement ou de réduction de la taille des cercles.
new.window	logical. Indique si la carte doit être dessinnée dans une nouvelle fenêtre.
...	Paramètres supplémentaires passés à la fonction title. Voir l'aide de cette fonction.

Details

Dessine tout d'abord le polygone défini par boundary. Si type = 'urb', représente les clusters, selon leur milieu de résidence, par des points verts et rouges. Par défaut, le vert est utilisé pour coder la première modalité, d'un point de vue alphabétique, de la colonne définie par var.urb. Si inverse = TRUE, le vert sera utilisé pour la dernière modalité, d'un point de vue alphabétique.

Si type = 'flower', la carte générée représentera le nombre d'obervations positives par cluster. Un cluster sans cas positif sera représenté par un point vert, un cluster avec un seul cas positif par un point mauve, un cluster avec plusieurs cas positifs par un point mauve et autant de 'rayons' rouges que de cas positifs. Voir sunflowerplot pour plus de détails sur ce type de graphiques.

Si type = 'count', la carte représente le nombre d'observations de chaque cluster par un cercle proportionnel au nombre d'observatuibs du cluster. La taille des cercles peut être contrôlée par factor.size. Pour un exemple d'exportation des résultats vers un logiciel de dessin afin d'appliquer une transparence aux cercles, voir le tutoriel de prevR ('tutoriel.prevR.pdf').

Examples

```
data(alicante)

map.clust(alicante.clust, alicante.bounds,
        type='count',
        new.window=FALSE,
        main='Number of tested persons by cluster',
        factor.size=0.15)

map.clust(alicante.clust, alicante.bounds,
        type='flower',
        main='Number of HIV positive persons by cluster')

map.clust(alicante.clust, alicante.bounds,
        type='urb',
        main='Clusters by type of residence')

map.clust(alicante.clust, alicante.bounds,
        type='urb',
        var.urb='urban.area',
        inverse=TRUE,
        main='Clusters in urban agglomeration')
```

---

merge.prev *Réorganise les résultats de estimate.prev.*

Description

Réorganise un data.frame produit par la fonction *estimate.prev* en un data.frame comportant un seul cluster par ligne, afin de permettre, par exemple, de réaliser plusieurs interpolations spatiales simultanément avec *krige.prev*.

Usage

```
merge.prev(data)
rename.variables.parameters(data)
```

Arguments

`data`	data.frame. L'appelation des colonnes doit correspondre à celui produit par `estimate.prev`.

Details

La fonction `rename.variables.parameters` supprime les colonnes *N.parameter*, *R.parameter* et *U.parameter* de `data` ; renomme les colonnes *est.prevalence*, *circle.count*, *circle.radius*, *circle.nb.clusters* et *quality.indicator* en *xxx.Naa.Rbb.Ucc* où *xxx* correspond à l'ancien nom de la colonne, *aa* à la valeur de *N.parameter*, *bb* à la valeur de *R.parameter* et *cc* à la valeur de *U.parameter*.

La fonction `merge.prev` extrait les données de `data` pour chaque combinaison des paramètres *N*, *R* et *U*, leur applique `rename.variables.parameters` et fusionne les données en un seul data.frame.

Value

data.frame avec un seul cluster par ligne. Voir les exemples.

Note

`merge.prev` est appliquée directement au résultat de `estimate.prev` si `merge.result = TRUE`.

See Also

`estimate.prev`.

Examples

```
data(alicante)

alicante.prev <- estimate.prev(
        alicante.clust,
        N=c(100,250),
        R=c(128,Inf),
        U=2)
str(alicante.prev)

alicante.prev.n250 <- extract.data(
        alicante.prev,
        value=c(250, Inf, 0))
str(alicante.prev.n250)

alicante.prev.n250 <- rename.variables.parameters(alicante.prev.n250)
str(alicante.prev.n250)

alicante.prev <- merge.prev(alicante.prev)
str(alicante.prev)
```

---

prevR-package *Prevalence estimation with DHS data - Estimation des prévalences à partir des données EDS.*

Description

Ce package permet d'importer des données de type EDS (Enquête Démographique et de Santé), de les formater, puis de cartographier la prévalence d'un phénomène par estimation de la prévalence de chaque point enquêté et interpolation spatiale. Les résultats peuvent ensuite être exportés vers d'autres logiciels de statistiques ou de cartographie (SIG). La documentation n'est disponible pour le moment qu'en français.

Details

Package:	prevR
Type:	Package
Version:	1.1
Date:	2007-10-10
License:	CeCILL-C - http://www.cecill.info/
Web:	http://ceped.cirad.fr/prevR/, http://joseph.larmarange.net/prevR/
License:	CeCILL-C - http://www.cecill.info/

Pour plus d'informations sur la manière d'utiliser ce package, nous vous conseillons la lecture du tutoriel de prevR ('tutoriel.prevR.pdf').

Author(s)

Joseph LARMARANGE <joseph@larmarange.net> IRD - Centre Muraz with supports of ANRS

Maintainer: Joseph LARMARANGE <joseph@larmarange.net>

References

- Joseph Larmarange et al., 2006, 'Cartographier les données des enquêtes démographiques et de santé à partir des coordonnées des zones d'enquête', Chaire Quételet, 29 novembre au 1er décembre 2006, Université Catholique de Louvain, Louvain-la-Neuve, Belgique.
Disponible en ligne à (http://www.uclouvain.be/13881.html).

- Joseph LARMARANGE, *Prévalences du VIH en Afrique : validité d'une mesure*, thèse de doctorat en démographie, sous la direction de Benoît FERRY, université Paris Descartes, 2007.
Disponible en ligne sur (http://joseph.larmarange.net/).

See Also

Packages nécessaires pour l'exécution de prevR : fields sp gstat maptools

Examples

```
## Pour une démonstration des possibilités de prevR

demo(prevR)
```

`prevR.colors` *Palettes de couleurs continues.*

Description

Fonctions générant des palettes de couleurs utilisables par les fonctions graphiques de R, en particulier spplot. Elles créent des palettes de couleurs continues, les contrastes étant renforcés par l'éclaircissement ou l'assombrissement des valeurs extrêmes.

Usage

```
prevR.demo.pal(n)
prevR.demo.pal(n, border, main, ch.col)
prevR.colors.red(n)
prevR.colors.red.inverse(n)
prevR.colors.blue(n)
prevR.colors.blue.inverse(n)
prevR.colors.green(n)
prevR.colors.green.inverse(n)
prevR.colors.gray(n)
prevR.colors.gray.inverse(n)
```

Arguments

n	integer. Nombre de couleurs constituant la palette.
border	color. Couleur de la bordure des cases.
main	character. Titre du graphique.
ch.col	character vector. Liste des fonctions à représenter.

Details

Le code de prevR.demo.pal a été repris sur celui de la fonction demo.pal décrite dans les exemples de la documentation de rainbow.

prevR.colors.red réalise un gradient allant du blanc/jaune au rouge/rouge foncé.
prevR.colors.blue réalise un gradient allant du bleu pâle au bleu foncé.
prevR.colors.green réalise un gradient allant du vert pâle au vert foncé.
prevR.colors.gray réalise un gradient allant du blanc/gris clair au gris foncé/noir.

Les fonctions avec le suffixe *.inverse* réalisent les mêmes gradients mais en partant des couleurs foncées vers les couleurs claires.

Value

prevR.demo.pal affiche les différentes palettes. Les autres fonctions renvoient une liste de couleurs codées de manière hexadécimale. Pour récupérer la liste des couleurs au format RGB (pour Red Green Blue), utilisez la fonction col2rgb.

See Also

D'autres palettes de couleurs existent sous R. Voir rainbow ainsi que le package **RColorBrewer**.

Examples

```
## Affiche les différentes palettes
prevR.demo.pal(25)

## Exemples d'utilisation avec spplot()

data(alicante)

# Représentation graphique
```

```
x11()
spplot(alicante.krige,
        c('est.prevalence.N100.RInf.U0.pred',
                'est.prevalence.N250.RInf.U0.pred',
                'est.prevalence.N250.R128.U0.pred',
                'est.prevalence.N250.R128.U6.pred'),
        col.regions=prevR.colors.red(21),
        cuts=20,
        main='Estimated prevalence with several parameters')

x11()
spplot(alicante.krige,
        'circle.radius.N250.R128.U6.pred',
        col.regions=prevR.colors.blue(21),
        cuts=20,
        main='Radius of circles - N=250 R=128 U=6')

x11()
spplot(alicante.krige,
        'quality.indicator.N250.R128.U6.pred',
        col.regions=prevR.colors.green.inverse(21),
        cuts=20,
        main='Quality indicator - N=250 R=128 U=6')
```

`read.spss2` *Lire un fichier SPSS au format sav.*

Description

Cette fonction est identique à `read.spss` exceptée sur un point. En effet, si dans le fichier SPSS, une variable comporte des étiquettes de valeurs non présentes dans les données, `read.spss` renvoie une colonne de type *atomic* tandis que `read.spss2` renvoie une colonne de type factor.

Usage

```
read.spss2(
   file,
   use.value.labels = TRUE,
   to.data.frame = FALSE,
   max.value.labels = Inf,
   trim.factor.names = FALSE)
```

Arguments

`file`	character. Nom du fichier *sav* à lire et transformer en data.frame.
`use.value.labels`	logical. Convertir les variables avec des étiquettes de valeurs en colonnes de type *factor* ?
`to.data.frame`	logical. Renvoyer un data.frame ? Renvoie une liste sinon.

`max.value.labels`
: logical. Seules les variables ayant au maximum ce nombre de modalités seront converties en *factor*.

`trim.factor.names`
: logical. Supprimer les espaces des *factor levels* ?

Details

Cette fonction est identique à read.spss exceptée sur un point. Si, dans le fichier SPSS, une variable comporte des étiquettes de valeurs non présentes dans les données, read.spss renvoie une colonne de type *atomic* tandis que `read.spss2` renvoie une colonne de type factor.

See Also

`read.spss` pour plus de détails sur cette fonction.

Examples

```
## Not run:
read.spss('file.sav')
## End(Not run)
```

---

verif.urb *Calcule des statistiques sur les agglomérations urbaines.*

Description

Calcule, pour chaque agglomération urbaine présente dans le fichier de donnée, la prévalence observée dans l'agglomération considérée et permet de comparer ces données avec celles provenant d'autres sources.

Usage

```
verif.urb(
   clust,
   conf.level = 0.9,
   add = FALSE,
   lang = "en",
   var.clust = c("n", "nweight", "obs.prevalence",
                                    "urban.area", "city.name"),
   urban.area.code = "in urban area")
```

Arguments

`clust`	data.frame. Chaque ligne doit correspondre à un cluster.
`conf.level`	numeric. Niveau de confiance pour les tests statistiques.
`add`	logical. Si `TRUE`, invite l'utilisateur à ajouter manuellement des données provenant d'une autre source.
`lang`	character. Permet de choisir la langue des messages utilisateur. `fr` pour le français, `en` pour l'anglais.

`var.clust` character vector. Noms des variables de `clust` correspondant dans l'ordre au nombre d'observations valides, à l'effectif pondéré, à la prévalence observée, à l'appartenance à une agglomération urbaine et au nom de la ville la plus proche de chaque cluster.

`urban.area.code` character. Valeur du facteur indiquant l'appartenance à une agglomération urbaine.

Details

Retient pour chaque agglomération urbaine les clusters appartenant à la dite agglomération et calcule une prévalence observée sur ces clusters.

Les intervalles de confiance sont calculés à l'aide de la fonction `prop.test`. Les comparaisons de deux proportions sont calculées à l'aide de la fonction `fisher.test`. Les effectifs pondérés de chaque cluster sont pris en compte.

Value

data.frame avec les colonnes suivantes :

`city` factor. Nom de l'agglomération.

`city.nb.cluster` numeric. Nombre de clusters appartenant à l'agglomération.

`city.n` numeric. Nombre d'observations valides de l'agglomération.

`city.prevalence` numeric. Prévalence observée dans l'agglomération, en pourcents.

`city.low` numeric. Valeur basse de l'intervalle de confiance de la prévalence observée.

`city.high` numeric. Valeur haute de l'intervalle de confiance de la prévalence observée.

`conf.level` numeric. Niveau de confiance pour les intervalles.

`add.prevalence` numeric. Prévalence de la source additionnelle saisie, en pourcents.

`add.n` numeric. Effectif de la source additionnelle.

`add.low` numeric. Valeur basse de l'intervalle de confiance de la prévalence additionnelle.

`add.high` numeric. Valeur haute de l'intervalle de confiance de la prévalence additionnelle.

`p.value.comparaison` numeric. Résultat du test de comparaison des deux proportions.

Note

Pour des exemples d'utilisation de cette fonction, en particulier avec `add = TRUE`, voir le tutoriel de prevR ('tutoriel.prevR.pdf').

Examples

```
data(alicante)

main.cities <- alicante.cities[alicante.cities$population > 100000, ]
alicante.clust <- calcul.dist.cities(
        alicante.clust,
```

```
        main.cities,
        type='all',
        dist=25)
verif.urb(alicante.clust)

main.cities <- alicante.cities[alicante.cities$population > 250000, ]
alicante.clust <- calcul.dist.cities(
        alicante.clust,
        main.cities,
        type='all',
        dist=25)
verif.urb(alicante.clust)
```

`write.boundary.shp` *Exporter les frontières au format shape.*

Description

Exporte un polygone fermé, défini dans un data.frame à deux colonnes, au format shape.

Usage

```
write.boundary.shp(boundary, file, country = file)
```

Arguments

`boundary`	data.frame. Typiquement un résultat de `make.boundary.dcw`.
`file`	character. Nom du fichier à créer.
`country`	character. Nom de la zone. Par défaut, prend le nom du fichier.

Note

Ce type de fichier peut être facilement importé dans un logiciel de cartographie.

See Also

`writePolyShape`.

Examples

```
## Not run:
data(alicante)
write.boundary.shp(alicante.bounds,'alicante-bounds','alicante')
## End(Not run)
```

`write.prev.shp` *Exporter des points au format shape.*

Description

Exporte les points contenus dans un data.frame ainsi que les données qui leur sont associées au format shape file.

Usage

```
write.prev.shp(x, file, coords = ~x+y, check=TRUE, lang='en')
```

Arguments

`x`	data.frame. Données à exporter, un point par ligne.
`file`	character. Nom du fichier à créer.
`coords`	formula. Variables définissant la longitude et la latitude des points, exprimées en tant que formule.
`check`	logical. Si vrai, alors la fonction `check.names` sera appliquée à `x` avant l'exportation, afin de renommer les noms de colonnes de plus de 10 caractères.
`lang`	character. Permet de choisir la langue des messages utilisateur. `fr` pour le français, `en` pour l'anglais.

Note

Ce type de fichier peut être facilement importé dans un logiciel de cartographie.

Le contenu du tableau de données est exporté au format *dbf*. Or ce type de format n'accepte pas les noms de variables de plus de 10 caractères. Au moment de l'export, les noms de colonnes de plus de 10 caractères seront donc tronqués, ce qui peut poser problème lorsque les 10 premiers caractères de deux noms de colonnes sont identiques. Si `check = TRUE`, vous serez invité à renommer les noms de variable de plus de 10 caractères.

See Also

`writePointsShape`, `check.names`.

Examples

```
## Not run:
data(alicante)
write.prev.shp(alicante.prev,'alicante-prev', lang='fr')
## End(Not run)
```

`write.txt` *Exporter un data.frame au format texte tabulé.*

Description

Permet d'exporter un data.frame au format texte, les valeurs étant séparées par des tabulations.

Usage

```
write.txt(x, file, dec = ".")
```

Arguments

`x`	data.frame. Données à exporter.
`file`	character. Nom du fichier à créer.
`dec`	character. Charactère de séparation des décimales.

Details

Est équivalent à
`write.table(x, file = file, sep="\t", row.names = FALSE, quote = FALSE, dec = dec)`.

See Also

`write.table` pour plus d'options d'export.

Examples

```
## Not run:
data(alicante)

write.txt(alicante.clust,'alicante-clusters.txt')

#Pour importer dans Excel version française :
write.txt(alicante.clust,'alicante-clusters.txt', dec=',')
## End(Not run)
```

Index

Annexe 6
Tutoriel de prise en main de prevR

Tutoriel prevR

Version du 10 octobre 2007 pour prevR 1.1

prevR a été développé par Joseph LARMARANGE dans le cadre d'un projet de recherche de l'IRD et du Centre Muraz financé par l'ANRS (Agence Nationale de Recherche sur le VIH/Sida). Ce projet, numéroté ANRS 12114, porte sur la mesure et les estimations des prévalences nationales du VIH en Afrique subsaharienne.

prevR a été conçu initialement pour représenter les variations spatiales de la prévalence du VIH à partir des données des Enquêtes Démographiques et de Santé (EDS ou DHS). Les exemples présentés ici portent donc sur cette problématique. Cependant, prevR peut être utilisé pour représenter tout type d'indicateur correspondant à une proportion dans le cadre d'enquêtes présentant un échantillonnage comparable à celui des EDS, c'est-à-dire un sondage en grappes.

prevR permet d'importer des données de type EDS, de les formater, puis de cartographier la prévalence d'un phénomène par estimation de la prévalence de chaque zone enquêtée (méthode des cercles) et interpolation spatiale (krigeage ordinaire). Les résultats peuvent être ensuite exportés vers d'autres logiciels de statistiques ou de cartographie (SIG). Des exemples d'exportation seront fournis à la fin de ce document.

Le site officiel de distribution de prevR est http://ceped.cirad.fr/prevR/. prevR est également disponible sur http://joseph.larmarange.net/prevR/. Ce second site propose également un forum de discussion utilisateur ainsi qu'une mailing liste pour être tenu informé des mises à jour de prevR.

Plan d'ensemble du tutoriel

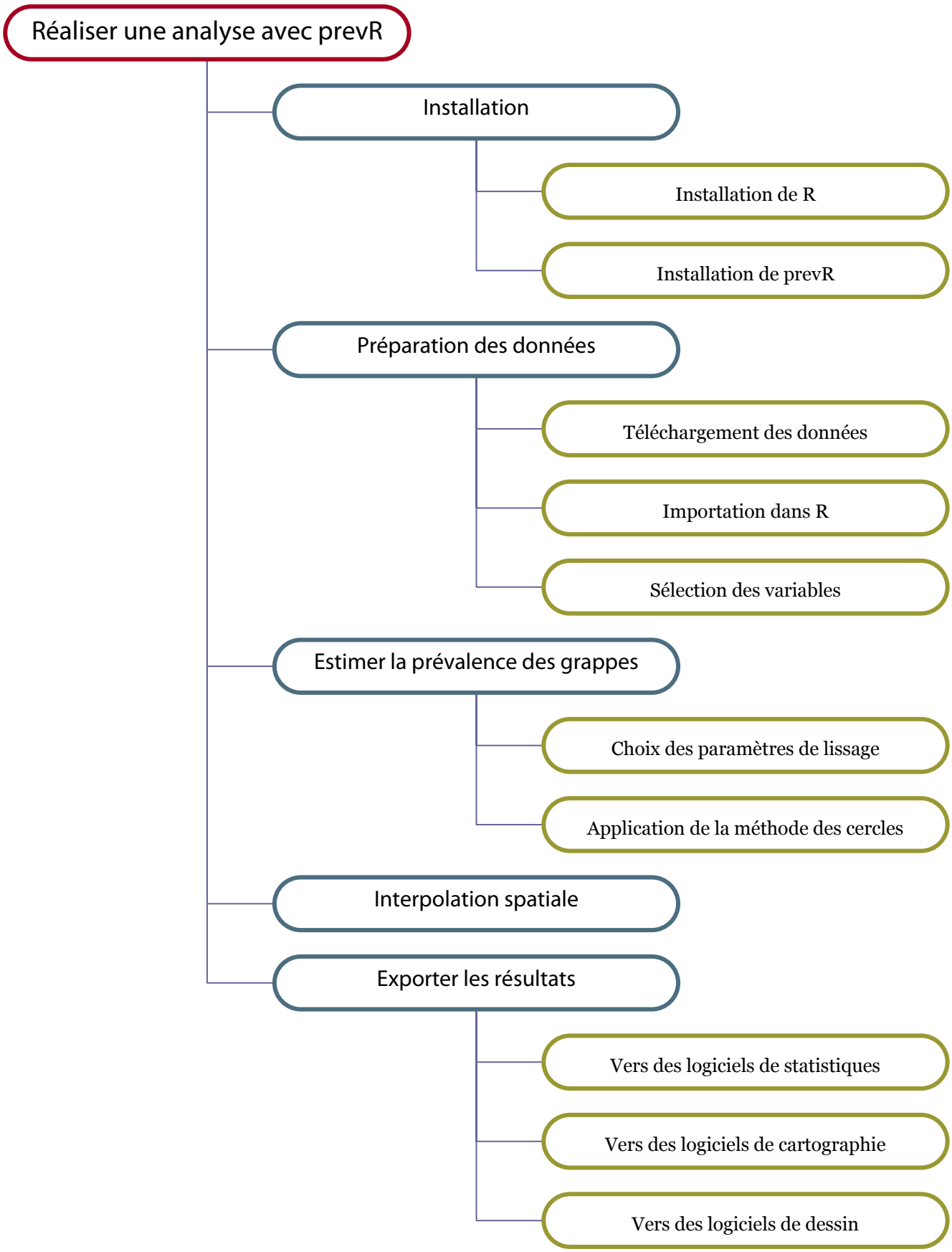


Table des Matières

Note :

Ce tutoriel a été conçu en priorité pour des utilisateurs ayant une connaissance minimum de R. La lecture du document **R pour les débutants** d'Emmanuel Paradis est donc fortement conseillée. Ce dernier est téléchargeable sur http://cran.r-project.org/other-docs.html.

Pour des utilisateurs expérimentés de R, nous leur conseillons de lire la documentation des fonctions de prevR.

Pour plus de détails techniques sur les méthodologies employées, nous vous renvoyons aux documents suivants :

> J. Larmarange, S. Yaro, R. Vallo, P. Msellati, N. Méda et B. Ferry, « Cartographier les données des enquêtes démographiques et de santé à partir des coordonnées des zones d'enquêtes », *Chaire Quételet 2006*, 29 novembre au 1er décembre 2006, Université Catholique de Louvain, Louvain-la-Neuve, Belgique[1].
>
> J. Larmarange, *Prévalences du VIH : validité d'une mesure*, chapitre 4, thèse de doctorat en démographie sous la direction de Benoît Ferry, Université Paris Descartes, 2007, disponible en ligne sur http://joseph.larmarange.net/.

1. Installation de R

1.1 Installation sous Windows

Il vous faut d'abord télécharger la dernière version de l'installateur pour Windows sur CRAN (Comprehensive R Archive Network) à cette adresse : http://cran.r-project.org/.

Cliquez sur *Windows (95 or later)*, puis sur *base*, et télécharger le fichier d'installation *R-2.6.0-win32.exe* (le numéro de version peut évoluer).

Lors de l'installation :

- Choisissez le dossier de destination (par défaut R sera installé dans *C:\Program Files\R*.
- Composants à installer : choisissez *Installation utilisateur complète*.
- Dans les options de démarrage, nous vous conseillons de choisir *Démarrage personnalisé*.
- Mode d'affichage : sélectionnez le mode *SDI* (ce dernier mode est nécessaire si vous vous voulez exploiter pleinement Tinn-R, voir 1.3).
- Style d'aide : si vous ne savez pas quoi choisir, sélectionnez le mode *CHM*.
- Accès internet : si vous avez un doute, sélectionnez *standard*.

[1] Ce document est disponible en ligne sur le site de l'UCL à http://www.uclouvain.be/13881.html ou sur le site du premier auteur à http://joseph.larmarange.net.

- Choisissez où vous souhaitez faire apparaître des raccourcis.

Pour information, il est possible de faire fonctionner R sur une clé USB. Pour cela, copier le répertoire *c:\Program Files\R* sur votre clé USB. Pour démarrer R, cliquez sur *Clé:\R\R-2.6.0\bin\Rgui.exe*.

1.2 Installation sous MacOS ou Linux

Pour une installation sur une autre plateforme que Windows, nous vous renvoyons à la documentation de R, en particulier au document intitulé **R Installation and Administration** disponible à cette adresse http://cran.r-project.org/manuals.html.

1.3 Optionnel : installer quelques projets d'amélioration de l'interface de R : Tinn-R et SciViews R Gui

Pour éditer du code R, nous vous recommandons le logiciel Tinn-R disponible gratuitement en ligne à cette adresse http://www.sciviews.org/Tinn-R/.

Vous pouvez également améliorer l'interface de R à l'aide du projet SciViews R Gui disponible en ligne à cette adresse http://www.sciviews.org/SciViews-R/index.html.

Il existe plusieurs projets d'amélioration de l'interface de R. Tous les renseignements sont disponibles sur http://www.sciviews.org/_rgui/.

2. Installation de prevR

2.1 Installation sous Windows.

Il faut d'abord installer les packages suivants qui sont nécessaires au fonctionnement de prevR :

- *sp*,
- *gstat*,
- *fields* et
- *maptools*.

Par ailleurs, nous vous recommandons vivement d'installer également les packages :

- *lattice* (fonctions graphiques supplémentaires) et
- *RSvgDevice* (permet d'exporter au format *SVG*).

L'ensemble de ces packages sont disponible sur CRAN. Ils peuvent être installés facilement dans R si vous disposez d'une connexion internet :

- Lancez R.
- Cliquez sur *Packages* > *Installer le(s) package(s)*.
- Sélectionnez un site miroir proche de vous.
- À l'aide de la touche *CTRL*, sélectionnez les packages suivants : *sp, gstat, fields, maptools, lattice* et *RSvgDevice*.

R téléchargera automatiquement et installera automatiquement ces différents packages.

Si vous ne disposez pas d'une connexion internet, vous pouvez télécharger manuellement ces différents package sur CRAN (choisissez le format *zip*) et les installer à partir de la commande *Packages* > *Installer le(s) package(s) des fichiers zip* disponible dans les menus de la console R (voir ci-dessous).

Il reste à installer prevR à partir du fichier *zip* disponible sur les sites du CEPED (http://ceped.cirad.fr/prevR/) et de Joseph Larmarange (http://joseph.larmarange.net/prevR/) :

- Téléchargez prevR au format *zip* et enregistrez le sur votre disque dur.
- Dans R, cliquez sur *Packages* > *Installer le(s) package(s) des fichiers zip*.
- Sélectionnez le fichier adéquat.

2.2 Installation sous MacOS ou Linux

Nous vous renvoyons à la section 6 du document **R Installation and Administration**, disponible à cette adresse http://cran.r-project.org/manuals.html, qui détaille la procédure à suivre pour installer des packages à partir des fichiers sources (distribués sous la forme d'archives *tar.gz*).

Les fichiers sources de prevR sont disponibles sur les site du CEPED (http://ceped.cirad.fr/prevR/) et de Joseph Larmarange (http://joseph.larmarange.net/prevR/). Téléchargez l'archive au format *tar.gz*.

Les packages nécessaires au fonctionnement de prevR sont disponibles sur CRAN (http://cran.r-project.org). Avant d'installer prevR, vous devrez installer les packages suivants : *sp*, *gstat*, *fields* et *maptools*. Nous vous recommandons d'installer également *lattice*, *RcolorBrewer*, *Rarcinfo* et *RSvgDevice*.

3. Remarques générales sur l'utilisation de R et de prevR

R est un langage, orienté objets, de programmation statistique basé sur le langage S. R est sensible à la casse des caractères. Ainsi, l'objet AbcD sera différent de ABCD ou encore de abcd. Nous vous déconseillons fortement d'utiliser des caractères accentués dans le nom des objets que vous manipulez avec R.

> De nombreuses ressources sont disponibles sur internet pour vous initier à R. Outre une recherche avec Google, nous vous recommandons de consulter cette page http://cran.r-project.org/other-docs.html et en particulier le document **R pour les débutants** d'Emmanuel Paradis.
>
> **La suite de ce tutoriel présuppose que vous ayez lu au minimum ce document.**

Lorsqu'un package est installé, les fichiers du package sont copiés dans le répertoire d'installation de R. Cependant, il n'est pas directement utilisable. En effet, lorsque que vous démarrez R, seules les fonctionnalités de base de R sont chargées en mémoire. Pour pouvoir utiliser les fonctions d'un package particulier, il faut au préalable le charger en mémoire, et ce à chaque fois que vous démarrez R.

3.1 Charger prevR, obtenir de l'aide, citer prevR et démonstration

Pour utiliser un package, vous devez le charger en mémoire. Deux solutions possibles :

- Utilisez le menu *Packages* > *Charger le package*.
- Utilisez la commande library ou la commande require.

Ainsi, pour charger prevR, il suffit de taper :

```
> library(prevR)
Le chargement a nécessité le package : fields
fields is loaded use help(fields) for an overview of this library
Le chargement a nécessité le package : sp
Le chargement a nécessité le package : gstat
Le chargement a nécessité le package : maptools
Le chargement a nécessité le package : foreign
```

Les autres packages nécessaires à l'utilisation de prevR sont chargés automatiquement.

Pour obtenir de l'aide sur une fonction dans R, il suffit de taper ? suivi du nom de la fonction ou du package, ou bien la fonction help. Ainsi, pour obtenir une aide générale sur prevR ou sur la fonction krige.prev, il suffit de taper :

```
> ?prevR
> help('prevR')
> ?krige.prev
```

Chaque fonction documentée fournit des exemples d'utilisation. Il est possible d'exécuter les exemples d'une fonction à l'aide de example(). La fonction mean() de R permet de calculer une moyenne. Pour voir des exemples d'utilisation de cette fonction il suffit de taper :

```
> example(mean)

mean> x <- c(0:10, 50)

mean> xm <- mean(x)

mean> c(xm, mean(x, trim = 0.1))
[1] 8.75 5.50

mean> mean(USArrests, trim = 0.2)
  Murder  Assault UrbanPop     Rape
    7.42   167.60    66.20    20.16
```

Pour rechercher un texte dans la documentation de R et de ses packages, il suffit d'avoir recours à la fonction help.search :

```
> help.search('proportion') # Pour chercher les fonctions travaillant sur des
proportions
```

NB : On notera que des commentaires peuvent être mis dans du code R à l'aide du caractère #. Le texte qui suit ce caractère n'est alors pas interprété.

prevR est livré avec un jeu de données permettant d'essayer ses différentes fonctions. Ce jeu de données est appelé alicante. Ces données sont issues de la simulation d'une Enquête Démographique et de Santé sur un pays fictif présentant une prévalence nationale de 10 pour cent. 8 000 personnes ont été enquêtées, réparties en 401 clusters.

Pour charger les données en mémoire, il suffit d'avoir recours à la fonction data. La fonction ls permet à tout moment de savoir quels objets sont présents en mémoire. Pour plus de détails sur les données fournies par alicante, il suffit de consulter l'aide associée.

```
> data(alicante)

> ls()
[1] "alicante.bounds" "alicante.cities" "alicante.clust"  "alicante.krige"  "alicante.prev"

> ?alicante
```

Pour savoir comment citer prevR dans un article, vous pouvez avoir recours à la fonction citation().

```
> citation('prevR')
To cite package prevR in publications use:
 Joseph Larmarange et al., 2006, 'Cartographier les données des enquêtes démographiques et de santé
à partir des coordonnées des zones  d'enquête', Chaire Quételet, 29 novembre au 1er décembre 2006,
Université Catholique de Louvain, Louvain-la-Neuve, Belgique (http://www.uclouvain.be/13881.html).
```

Pour une démonstration des possibilités du package prevR, utilisez la fonction demo().

```
> demo(prevR)
```

ATTENTION : la démonstration peut prendre plusieurs minutes, selon la puissance de votre machine, certaines fonctions ayant des temps de calcul relativement longs.

3.2 Quelques fonctions de base

Voici une liste, non exhaustive, de quelques fonctions de base particulièrement utiles. Pour plus de détails, voir l'aide de chaque fonction.

Nom	Description
as.character	Pour passer un objet en mode texte.
as.factor	Pour passer un objet en mode facteurs.
as.numeric	Pour passer un objet en mode numérique.
c	Permet de créer un vecteur.
data	Charge des données contenues dans un package.
demo	Permet d'exécuter une démonstration.
dev.copy	Copie une sortie graphique vers une autre sortie graphique.
dev.off	Ferme la sortie graphique courante.
dev.set	Active une sortie graphique.
edit	Édite un objet.
example	Exécute les exemples fournis dans la documentation d'une fonction.
function	Pour écrire ses propres fonctions.
getwd	Affiche le répertoire de travail courant.
graphics.off	Ferme toutes les fenêtres graphiques.
help	Fournit de l'aide sur une fonction.
help.search	Effectue une recherche dans l'aide.
levels	Affiche les étiquettes de valeur d'un objet de type facteurs.
library	Charge un package en mémoire.
list	Pour créer des listes.
load	Charge un fichier de données.
ls	Liste l'ensemble des objets en mémoire.
order	Pour trier des données.
rm	Supprime un objet.
save	Sauve un ou plusieurs objets dans un fichier de données.
save.image	Sauve l'ensemble des objets en mémoire.
seq	Permet de générer une liste de nombres.
setwd	Définit le répertoire de travail.
str	Détaille la structure d'un objet.
summary	Fournit un résumé détaillé du contenu d'un objet.
write.table	Exporte des données sous la forme d'un fichier texte.
x11	Ouvre une nouvelle fenêtre graphique.

Pour des manipulations avancées des tableaux de données, nous vous conseillons la lecture de l'aide de l'opérateur [:

```
> help('[.data.frame')
```

4. Préparation des données

Nous illustrerons l'utilisation de prevR à travers l'estimation des variations spatiales de la prévalence du VIH au Cameroun à partir de l'Enquête Démographique et de Santé de 2004.

4.1 Téléchargement des données

Afin de réaliser cette analyse, il est nécessaire de récupérer les données suivantes :

- localisation des zones d'enquêtes de l'EDS ou clusters,
- résultats au test VIH des personnes enquêtées,
- frontières du pays (sous la forme d'un polygone géoréférencé),
- localisation des principales villes du pays (coordonnées longitude/latitude).

4.1.1 Bases de données des EDS

Les données des enquêtes EDS peuvent être obtenues gratuitement en ligne sur le site de Measure DHS : http://www.measuredhs.com/. Si vous n'êtes pas inscrit, il vous faudra créer un compte, décrire votre projet de recherche et demander l'accès aux données du pays intéressé. Il faut réaliser une demande spécifique pour l'accès aux données VIH et une autre pour l'accès aux données GPS. Pour l'obtention des données GPS, un formulaire d'engagement éthique devra être imprimé, signé et à retourné à Measure DHS. Il faut compter parfois quelques jours avant que l'accès aux données ne vous soit notifié.

Les données d'enquêtes ou les résultats au test VIH doivent être téléchargées au format SPSS (fichiers avec le suffixe *sv.zip*) ou au format rectangulaire (fichiers avec le suffixe *rt.zip*). Décompressez les archives et copiez le fichier portant l'extension *.sav* dans un répertoire de travail. Dans notre exemple, nous avons renommé le fichier des résultats du test VIH de l'EDS 2004 du Cameroun en *cm.hiv.sav* pour plus de commodités.

Les données GPS sont téléchargeables directement au format *dbf*. Par commodité, nous avons renommé ce fichier en *cm.gps.dbf*.

4.1.2 Frontières du pays via le DCW

Il existe plusieurs bases de données cartographiques fournissant les frontières nationales des différents pays du monde. L'une des plus connues est le *Digital Chart of the World*, dont les données sont téléchargeables gratuitement en ligne sur http://www.maproom.psu.edu/dcw/. prevR fournit une fonction permettant d'importer facilement un fichier de points téléchargé depuis ce site.

Cependant, il faut noter que les données du DCW datent de 1992 et n'ont pas été actualisées. Elles ne seront donc plus valables si les frontières du pays étudié ont subi des modifications depuis cette date.

Vous pouvez néanmoins utiliser toute autre source pour définir les limites de votre zone d'études. Il vous suffit de les importer et de les mettre en forme de manière à obtenir un *data.frame* (tableau de données) avec deux colonnes nommées *x* et *y* dont chaque ligne correspond à un point d'un polygone fermé[2].

Dans le présent exemple, nous utiliserons les données du DCW pour le Cameroun. Après avoir sélectionné votre continent et votre pays, choisissez l'option *Download Points*. Faites un clic droit sur *download data* et choisissez *Enregistrez la cible du lien sous*. Vous obtiendrez alors un fichier de la forme *pays2pts.txt* et soit *cameroon2pts.txt* dans notre exemple.

4.1.3 Position des principales villes via le GRUMP

Pour prendre en compte les principales agglomérations urbaines dans l'analyse, il est nécessaire de connaître les coordonnées des principales villes du pays. Celles-ci peuvent être obtenues gratuitement en ligne à partir du projet Global Rural-Urban Mapping Project (GRUMP). Les données de ce projet sont accessibles à cette adresse : http://sedac.ciesin.org/gpw/.

Arrivé sur leur site, choisissez *downloadable data*. Sélectionner le pays qui vous intéresse dans la liste située en fin de page. Dans la partie *Select a product*, choisissez *Get GRUMP > Settlement Points*. Dans *Select Options*, sélectionnez le format *csv* et l'année *circa 2000*.

Vous devrez vous enregistrer pour pouvoir télécharger le fichier désiré. Certains utilisateurs peuvent rencontrer des soucis de téléchargement avec le logiciel *Internet Explorer*. Préférez dans ce cas là le navigateur libre *Firefox*, téléchargeable gratuitement sur http://www.mozilla-europe.org/fr/products/firefox/.

Après décompression de l'archive, placez le fichier *csv* dans votre répertoire de travail. Par commodité, nous avons renommé le fichier obtenu dans notre exemple en *cm.cities.csv*.

NB : si vous comptez exporter ultérieurement vos résultats vers un logiciel de cartographie, vous pouvez également télécharger la position des villes au format *shapefile* pour pouvoir habiller vos cartes.

4.2 Importation et mise en forme des données

Nous vous recommandons de placer les différents fichiers obtenus dans un même répertoire que nous appellerons répertoire de travail.

Au lancement de R, utilisez la commande *Fichier > Changer le répertoire courant…* pour spécifier votre répertoire de travail.

Puis, utilisez la commande *Packages > Charger le package…* pour charger prevR en mémoire.

[2] Une aire géographique est définie d'un point de vue informatique par une série de points, chacun défini par une latitude et une longitude, formant un polygone. Voir 4.2.4 pour plus de détails.

4.2.1 Création du fichier individu

La fonction make.ind.spss a été spécialement conçue pour lire et mettre en forme des données individuelles, au format SPSS, fournies par Measure DHS. Elle prend deux paramètres : le nom du fichier SPSS à importer et le code de langue pour l'interface utilisateur ('*en*' pour l'anglais, '*fr*' pour le français).

```
> cm.ind <- make.ind.spss('cm.hiv.sav',lang='fr')
```

La fonction commence par lire les différentes variables contenues dans le fichier SPSS, puis vous demande de spécifier un certain nombre d'entre elles :

```
Veuillez indiquer les variables suivantes :

Identifiant (0 s'il n'existe pas) :

 1: ACASEID - space dummy                          2: HIVCLUST - Cluster number
 3: HIVHHN - HH structure                          4: HIVMENA - HH menage
 5: HIV60 - Line number of respondent              6: HIV62 - Sex of household member
 7: HIV63 - Age of household members               8: HIV64 - Age 15-17, 18+
 9: HIV65 - Line number of parent/responsible     10: HIV66 - Consent statement to parent
11: HIV67 - Consent to woman/man                  12: HIV68 - Sample result
13: HIVREG - Province                             14: HIVLOC - Locality
15: HIVTYPE - Urban/rural                         16: HDEFACTO - Slept last night
17: HIVEDUC - Level of education attending        18: HIVGRADE - Highest grade of education
19: HIVQNL - Sequence order in section            20: INDINT - Result of individual interview
21: TESTED - Found and tested in LAB file         22: HHDUP - Duplicate ID in HH
23: LABDUP - Duplicate ID in LAB file             24: HIVWT - Weight for HIV sample
25: RESULT.1 - Result of testing                  26: RESULT.2 - Result of testing
27: RESULT.3 - Result of testing                  28: FRESULT - Final result of testing
29: HIVCHILD - Had a child in last 5 years        30: HIVANC - Received ANC in last 5 years
31: HIVPET - Cluster in petrol line

Sélection : 1

Numéro du cluster (0 s'il n'existe pas. Le numéro de cluster sera alors calculé à partir des
identifiants.) :
Sélection : 0

Age (0 s'il n'existe pas) :
Sélection : 7

Sexe (0 s'il n'existe pas) :
Sélection : 6

Variable analysée (par exemple, résultat du test VIH) :
Sélection : 28

Poids statistique (0 s'il n'existe pas. Tous les individus auront un poids égal à 1.) :
Sélection : 24
```

Dans le cas présent, bien que la variable numéro du cluster était présente (variable 2), nous ne l'avons pas entrée de manière à montrer comment calculer le numéro de cluster à partir des identifiants.

Une fois les variables saisies, la fonction vous demande de confirmer votre choix. En cas d'erreur, sélectionnez *Non* et recommencez la saisie :

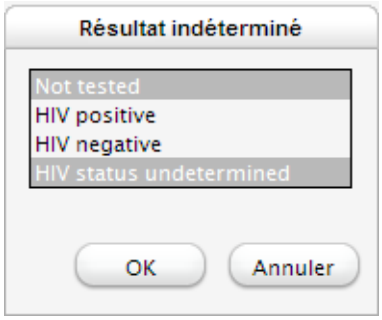
```
ATTENTION : veuillez vérifier les informations suivantes :

* Identifiant des individus :
ACASEID - space dummy
* Numéro de cluster :
Non renseigné - Il sera calculé à partir du numéro d'identifiants
* Age :
HIV63 - Age of household members
* Sexe :
HIV62 - Sex of household member
* Variable analysée :
FRESULT - Final result of testing
* Poids statistique :
HIVWT - Weight for HIV sample
-------------------------------------------------
Ces données sont-elles correctes ?
1: Oui
2: Non
```

Sélection : 1

Vous êtes ensuite invité à spécifier comment la variable analysée a été codée.

```
Trois fenêtres vont s'ouvrir pour recoder la variable analysée.
Vous devrez spécifier les modalités correspondant à un résultat positif (le phénomène étudié a eu
lieu),
négatif (n'a pas eu lieu) ou indéterminé (considéré alors comme valeur manquante).
Vous pouvez sélectionner plusieurs modalités à l'aide de la touche CTRL.
Êtes-vous-prêt ?

1: Oui
```

Sélection : 1



Exemple de fenêtre, pour la modalité *indéterminé :*

Les fenêtres pour les modalités *Résultat positif* et *Résultat négatif* sont identiques.

Par sécurité, il vous est demandé de confirmer à nouveau votre saisie :

```
ATTENTION : veuillez vérifier les informations suivantes :

* Résultat positif :
- HIV positive

* Résultat négatif :
- HIV negative

* Résultat indéterminé :
- Not tested
- HIV status undetermined

Ces données sont-elles correctes ?

1: Oui
2: Non
```

Sélection : 1

Si une variable de pondération a été saisie, il vous est demandé s'il est nécessaire de la diviser par un facteur donné. Cela dépend bien entendu de votre enquête. Dans les EDS, il faut en général les diviser par 1 000 000.

```
Souvent, dans les EDS, la variable poids doit être divisée par un facteur, usuellement 1 000 000.
Sa valeur moyenne est de :
999999.41

Si cette valeur est proche de 1, a priori la variable n'a pas à être modifiée. Si elle est proche
de 1 000 000, alors elle doit être divisée par ce facteur, sinon consultez la documentation de
l'enquête.
La variable doit-elle être gardée telle quelle ou divisée par un facteur ?

1: Pas de modification
2: Division par 1 000 000
3: Division par un autre facteur
```

Sélection : 2

Lorsque le numéro de cluster n'est pas renseigné, il peut être calculé à partir de l'identifiant des individus si celui-ci contient le numéro de cluster (ce qui est le cas dans les EDS). Typiquement, le numéro de cluster correspond aux premiers chiffres de l'identifiant.

```
La variable cluster n'est pas renseignée et doit donc être calculée à partir du numéro
d'identifiant. Usuellement, dans les EDS, le numéro de cluster correspond aux trois premiers
chiffres du numéro d'identifiant. Voici trois numéros d'identifiants, pris au début, au milieu et à
la fin du fichier :
Numéro d'identification 1 :       1 26 3  1
Numéro d'identification 2 :     236149 6  2
Numéro d'identification 3 :     466312 3  4
```

Dans le cas présent, des trois identifiants prélevés dans la base, on voit clairement que les numéros de cluster correspondant sont 1, 236 et 466. La découpe adéquate des identifiants est donc la découpe numéro 5.

```
Repérez pour chacun d'eux le numéro de cluster. Parmi les différentes propositions ci-dessous,
laquelle extrait les bons numéros de cluster ?
  1: Cluster 1:              Cluster 2:                  Cluster 3:
  2: Cluster 1:              Cluster 2:                  Cluster 3:
  3: Cluster 1:              Cluster 2:    2             Cluster 3:    4
  4: Cluster 1:              Cluster 2:   23             Cluster 3:   46
  5: Cluster 1:    1         Cluster 2:  236             Cluster 3:  466
  6: Cluster 1:   1          Cluster 2:  361             Cluster 3:  663          .
  7: Cluster 1:  1 2         Cluster 2:  614             Cluster 3:  631
  8: Cluster 1:   26         Cluster 2:  149             Cluster 3:  312
  9: Cluster 1:  26          Cluster 2:  49              Cluster 3:  12
 10: Cluster 1:  6 3         Cluster 2:  9 6             Cluster 3:  2 3      .
 11: Cluster 1:   3          Cluster 2:   6              Cluster 3:   3
 12: Cluster 1:  3           Cluster 2:  6               Cluster 3:  3
 13: Cluster 1:    1         Cluster 2:    2             Cluster 3:    4
```

Sélection : 5

Une fois la fonction exécutée, on peut vérifier son résultat à l'aide de str().

```
> str(cm.ind)
'data.frame':   12065 obs. of  7 variables:
 $ id             : chr  "     1 26 3  1" "     1 26 3  2" "     1137 1  1" "     1137 1  2" ...
 $ age            : int  24 25 36 29 47 48 44 18 16 15 ...
 $ sex            : Factor w/ 2 levels "Male","Female": 2 2 1 2 2 1 2 2 2 1 ...
 $ original.result: Factor w/ 4 levels "Not tested","HIV positive",..: 1 1 3 3 3 3 3 3 3 3 ...
 $ weight         : num  0.00 0.00 1.25 1.16 1.16 ...
 $ result         : Factor w/ 2 levels "Negative","Positive": NA NA 1 1 1 1 1 1 1 1 ...
 $ cluster        : int  1 1 1 1 1 1 1 1 1 1 ...
```

Pour le détail des différentes variables, voir l'aide de make.ind.spss() :

```
> ?make.ind.spss
```

4.2.2 Création du fichier cluster

Les analyses effectuées par prevR portent essentiellement sur un tableau de données (*data.frame*). Pour voir la structure de celui-ci, vous pouvez vous référer au fichier *alicante.clust* fourni avec prevR.

```
> data(alicante)
> str(alicante.clust)
'data.frame':   401 obs. of  11 variables:[3]
 $ cluster       : int  1 2 3 4 5 6 7 8 9 10 ...
 $ x             : num  -1.21 -1.87 -1.04 -1.37 -2.21 ...
 $ y             : num  7.29 7.54 7.96 6.46 6.88 ...
 $ residence     : Factor w/ 2 levels "Rural","Urban": 1 1 1 1 1 1 1 1 1 1 ...
 $ region        : num  1 1 1 1 1 1 1 1 1 1 ...
 $ n             : num  23 21 24 16 26 22 21 22 17 22 ...
 $ nweight       : num  19.8 19.8 19.8 19.8 19.8 ...
 $ obs.prevalence: num  0.00 4.76 0.00 0.00 3.85 ...
 $ dist.city     : num   72.9  93.2 146.9  39.8  70.5 ...
 $ city.name     : chr  "D" "D" "D" "D" ...
 $ urban.area    : Factor w/ 2 levels "in urban area",..: 2 2 2 2 2 2 2 2 2 2 ...
```

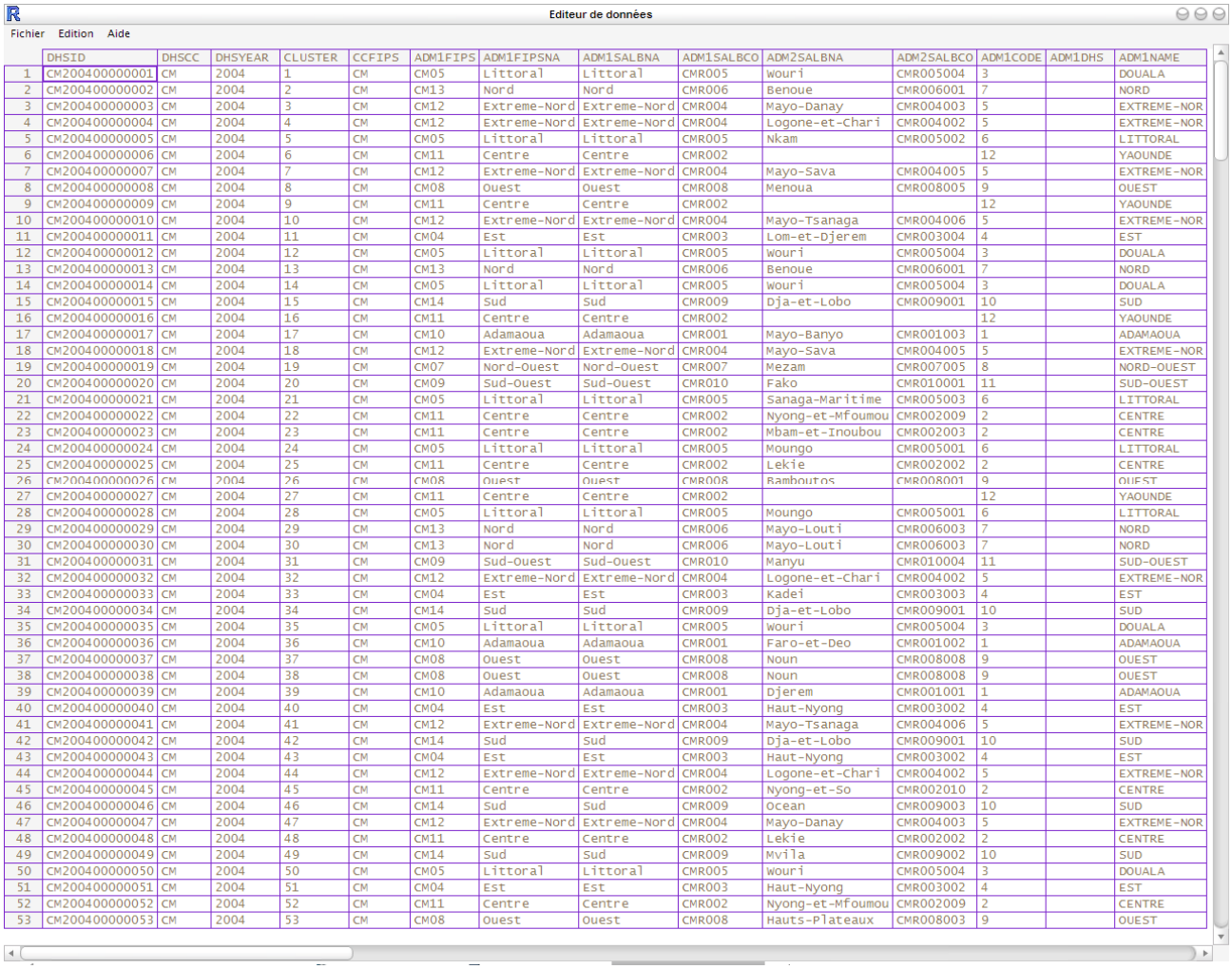
La fonction make.clust.dbf() permet de construire ce tableau de données à partir du fichier *dbf* fourni par Measure DHS et du fichier individu créé à l'étape précédente.

```
> cm.clust <- make.clust.dbf('cm.gps.dbf',cm.ind,lang='fr')
```

Dans un fichier *dbf*, on dispose seulement du nom des variables mais il n'y a pas d'étiquette spécifiant leur contenu comme dans un fichier SPSS. Une fenêtre présentant le contenu du fichier *dbf* s'affiche alors afin que vous puissiez identifier un certain nombre de variables. Une fois cela fait, fermez cette fenêtre pour pouvoir effectuer votre saisie.

```
Une fenêtre va s'ouvrir présentant les données contenues dans le fichier. Merci de repérer les
variables suivantes :
- Numéro du cluster (nécessaire)
- Longitude (en degrés au format décimal, nécessaire)
- Latitude (en degrés au format décimal, nécessaire)
- Milieu de résidence (urbain/rural, nécessaire si utilisation du paramètre U)
- Code numérique des régions (optionnel)
- Nom des régions (optionnel)
 Une fois les noms de ces variables identifiés, fermez la fenêtre pour que le programme puisse
continuer.

 Êtes-vous prêt ?
1: Oui

Sélection : 1
```

[3] Les variables *dist.city, city.name* et *urban.area* ne sont pas nécessaires au début de l'analyse. Elles seront rajoutées ultérieurement au tableau de données par la fonction calcul.dist.cities().



Editeur de données

Fichier Edition Aide

	DHSID	DHSCC	DHSYEAR	CLUSTER	CCFIPS	ADM1FIPS	ADM1FIPSNA	ADM1SALBNA	ADM1SALBCO	ADM2SALBNA	ADM2SALBCO	ADM1CODE	ADM1DHS	ADM1NAME
1	CM200400000001	CM	2004	1	CM	CM05	Littoral	Littoral	CMR005	Wouri	CMR005004	3		DOUALA
2	CM200400000002	CM	2004	2	CM	CM13	Nord	Nord	CMR006	Benoue	CMR006001	7		NORD
3	CM200400000003	CM	2004	3	CM	CM12	Extreme-Nord	Extreme-Nord	CMR004	Mayo-Danay	CMR004003	5		EXTREME-NOR
4	CM200400000004	CM	2004	4	CM	CM12	Extreme-Nord	Extreme-Nord	CMR004	Logone-et-Chari	CMR004002	5		EXTREME-NOR
5	CM200400000005	CM	2004	5	CM	CM05	Littoral	Littoral	CMR005	Nkam	CMR005002	6		LITTORAL
6	CM200400000006	CM	2004	6	CM	CM11	Centre	Centre	CMR002			12		YAOUNDE
7	CM200400000007	CM	2004	7	CM	CM12	Extreme-Nord	Extreme-Nord	CMR004	Mayo-Sava	CMR004005	5		EXTREME-NOR
8	CM200400000008	CM	2004	8	CM	CM08	Ouest	Ouest	CMR008	Menoua	CMR008005	9		OUEST
9	CM200400000009	CM	2004	9	CM	CM11	Centre	Centre	CMR002			12		YAOUNDE
10	CM200400000010	CM	2004	10	CM	CM12	Extreme-Nord	Extreme-Nord	CMR004	Mayo-Tsanaga	CMR004006	5		EXTREME-NOR
11	CM200400000011	CM	2004	11	CM	CM04	Est	Est	CMR003	Lom-et-Djerem	CMR003004	4		EST
12	CM200400000012	CM	2004	12	CM	CM05	Littoral	Littoral	CMR005	Wouri	CMR005004	3		DOUALA
13	CM200400000013	CM	2004	13	CM	CM13	Nord	Nord	CMR006	Benoue	CMR006001	7		NORD
14	CM200400000014	CM	2004	14	CM	CM05	Littoral	Littoral	CMR005	Wouri	CMR005004	3		DOUALA
15	CM200400000015	CM	2004	15	CM	CM14	Sud	Sud	CMR009	Dja-et-Lobo	CMR009001	10		SUD
16	CM200400000016	CM	2004	16	CM	CM11	Centre	Centre	CMR002			12		YAOUNDE
17	CM200400000017	CM	2004	17	CM	CM10	Adamaoua	Adamaoua	CMR001	Mayo-Banyo	CMR001003	1		ADAMAOUA
18	CM200400000018	CM	2004	18	CM	CM12	Extreme-Nord	Extreme-Nord	CMR004	Mayo-Sava	CMR004005	5		EXTREME-NOR
19	CM200400000019	CM	2004	19	CM	CM07	Nord-Ouest	Nord-Ouest	CMR007	Mezam	CMR007005	8		NORD-OUEST
20	CM200400000020	CM	2004	20	CM	CM09	Sud-Ouest	Sud-Ouest	CMR010	Fako	CMR010001	11		SUD-OUEST
21	CM200400000021	CM	2004	21	CM	CM05	Littoral	Littoral	CMR005	Sanaga-Maritime	CMR005003	6		LITTORAL
22	CM200400000022	CM	2004	22	CM	CM11	Centre	Centre	CMR002	Nyong-et-Mfoumou	CMR002009	2		CENTRE
23	CM200400000023	CM	2004	23	CM	CM11	Centre	Centre	CMR002	Mbam-et-Inoubou	CMR002003	2		CENTRE
24	CM200400000024	CM	2004	24	CM	CM05	Littoral	Littoral	CMR005	Moungo	CMR005001	6		LITTORAL
25	CM200400000025	CM	2004	25	CM	CM11	Centre	Centre	CMR002	Lekie	CMR002002	2		CENTRE
26	CM200400000026	CM	2004	26	CM	CM08	Ouest	Ouest	CMR008	Bamboutos	CMR008001	9		OUEST
27	CM200400000027	CM	2004	27	CM	CM11	Centre	Centre	CMR002			12		YAOUNDE
28	CM200400000028	CM	2004	28	CM	CM05	Littoral	Littoral	CMR005	Moungo	CMR005001	6		LITTORAL
29	CM200400000029	CM	2004	29	CM	CM13	Nord	Nord	CMR006	Mayo-Louti	CMR006003	7		NORD
30	CM200400000030	CM	2004	30	CM	CM13	Nord	Nord	CMR006	Mayo-Louti	CMR006003	7		NORD
31	CM200400000031	CM	2004	31	CM	CM09	Sud-Ouest	Sud-Ouest	CMR010	Manyu	CMR010004	11		SUD-OUEST
32	CM200400000032	CM	2004	32	CM	CM12	Extreme-Nord	Extreme-Nord	CMR004	Logone-et-Chari	CMR004002	5		EXTREME-NOR
33	CM200400000033	CM	2004	33	CM	CM04	Est	Est	CMR003	Kadei	CMR003003	4		EST
34	CM200400000034	CM	2004	34	CM	CM14	Sud	Sud	CMR009	Dja-et-Lobo	CMR009001	10		SUD
35	CM200400000035	CM	2004	35	CM	CM05	Littoral	Littoral	CMR005	Wouri	CMR005004	3		DOUALA
36	CM200400000036	CM	2004	36	CM	CM10	Adamaoua	Adamaoua	CMR001	Faro-et-Deo	CMR001002	1		ADAMAOUA
37	CM200400000037	CM	2004	37	CM	CM08	Ouest	Ouest	CMR008	Noun	CMR008008	9		OUEST
38	CM200400000038	CM	2004	38	CM	CM08	Ouest	Ouest	CMR008	Noun	CMR008008	9		OUEST
39	CM200400000039	CM	2004	39	CM	CM10	Adamaoua	Adamaoua	CMR001	Djerem	CMR001001	1		ADAMAOUA
40	CM200400000040	CM	2004	40	CM	CM04	Est	Est	CMR003	Haut-Nyong	CMR003002	4		EST
41	CM200400000041	CM	2004	41	CM	CM12	Extreme-Nord	Extreme-Nord	CMR004	Mayo-Tsanaga	CMR004006	5		EXTREME-NOR
42	CM200400000042	CM	2004	42	CM	CM14	Sud	Sud	CMR009	Dja-et-Lobo	CMR009001	10		SUD
43	CM200400000043	CM	2004	43	CM	CM04	Est	Est	CMR003	Haut-Nyong	CMR003002	4		EST
44	CM200400000044	CM	2004	44	CM	CM12	Extreme-Nord	Extreme-Nord	CMR004	Logone-et-Chari	CMR004002	5		EXTREME-NOR
45	CM200400000045	CM	2004	45	CM	CM11	Centre	Centre	CMR002	Nyong-et-So	CMR002010	2		CENTRE
46	CM200400000046	CM	2004	46	CM	CM14	Sud	Sud	CMR009	Ocean	CMR009003	10		SUD
47	CM200400000047	CM	2004	47	CM	CM12	Extreme-Nord	Extreme-Nord	CMR004	Mayo-Danay	CMR004003	5		EXTREME-NOR
48	CM200400000048	CM	2004	48	CM	CM11	Centre	Centre	CMR002	Lekie	CMR002002	2		CENTRE
49	CM200400000049	CM	2004	49	CM	CM14	Sud	Sud	CMR009	Mvila	CMR009002	10		SUD
50	CM200400000050	CM	2004	50	CM	CM05	Littoral	Littoral	CMR005	Wouri	CMR005004	3		DOUALA
51	CM200400000051	CM	2004	51	CM	CM04	Est	Est	CMR003	Haut-Nyong	CMR003002	4		EST
52	CM200400000052	CM	2004	52	CM	CM11	Centre	Centre	CMR002	Nyong-et-Mfoumou	CMR002009	2		CENTRE
53	CM200400000053	CM	2004	53	CM	CM08	Ouest	Ouest	CMR008	Hauts-Plateaux	CMR008003	9		OUEST

Une fois les variables identifiées, fermez l'éditeur pour que R puisse reprendre la main et vous inviter à saisir les variables correspondantes. Faites attention, les noms des variables peuvent différer d'un pays à l'autre.

```
Veuillez indiquez les variables suivantes :

Numéro des clusters :

 1: DHSID          2: DHSCC          3: DHSYEAR        4: CLUSTER        5: CCFIPS         6: ADM1FIPS
 7: ADM1FIPSNA     8: ADM1SALBNA     9: ADM1SALBCO    10: ADM2SALBNA    11: ADM2SALBCO    12: ADM1CODE
13: ADM1DHS       14: ADM1NAME      15: ADM2CODE      16: ADM2DHS       17: ADM2NAME      18: ADM3CODE
19: ADM3DHS       20: ADM3NAME      21: ADM4CODE      22: ADM4DHS       23: ADM4NAME      24: PPLNAME
25: PPLCODE       26: REPAR1DHS     27: REPAR1NAME    28: REPAR2DHS     29: REPAR2NAME    30: SOURCE
31: LOCAL_LVL     32: U.R           33: LATNUM        34: LATDEG        35: LATMIN        36: LATSEC
37: LATTHOU       38: LATHEMI       39: LONGNUM       40: LONGDEG       41: LONGMIN       42: LONGSEC
43: LONGTHOU      44: LONGHEMI      45: UTMLAT        46: UTMLONG       47: UTMZONE       48: ALT_GPS
49: ALT_DEM       50: DATUM         51: SYMBOL        52: WAF_ID

Sélection : 4

Longitude (valeur décimale) :
Sélection : 39

Latitude (valeur décimale) :
Sélection : 33

Milieu de résidence :
Sélection : 32

Code numérique des régions (0 si non renseigné) :
Sélection : 12

Nom des régions (0 si non renseigné) :
```

```
Sélection : 14
```

Vérification des informations saisies :

```
ATTENTION : veuillez vérifier les informations suivantes :

* Numéro de cluster :
CLUSTER
* Longitude (valeur decimale) :
LONGNUM
* Latitude (valeur décimale) :
LATNUM
* Milieu de résidence :
U.R
* Code numérique des régions :
ADM1CODE
* Nom des régions :
ADM1NAME

--------------------------------------------------

Ces données sont-elles correctes ?

1: Oui
2: Non
```

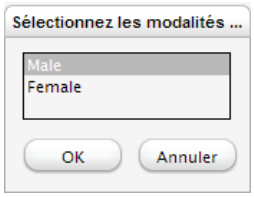
```
Sélection : 1
```

Si la variable *sexe* est présente dans le fichier individu, il est possible de restreindre l'analyse à l'une des modalités de cette variable.

```
La variable sexe a été détectée dans le fichier ind.
Voici ses modalités :
- Male
- Female

Voulez-vous restreindre l'analyse à l'une ou plusieurs de ces modalités ?

1: Oui
2: Non
```

```
Sélection : 2
```

Si vous choisissez de restreindre l'analyse à une ou plusieurs modalités de cette variable, une fenêtre telle que celle ci-dessous s'ouvrira vous permettant de restreindre l'analyse à une ou plusieurs modalités (utilisez la touche CTRL pour sélectionner plusieurs modalités.

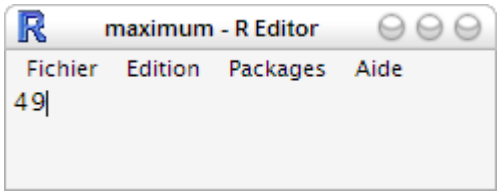


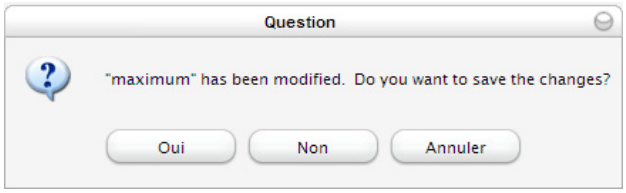
De même, si la variable *age* est présente dans le fichier individu, il est possible de restreindre l'analyse à une classe d'âges donnée. Dans le cas présent, comme de 50 à 59 ans seuls des hommes ont été testés, nous allons restreindre notre analyse aux 15-49 ans.

```
La variable âge a été détectée dans le fichier ind.
Voici ses caractéristiques :
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  15.00   20.00   26.00   28.78   36.00   59.00

Voulez-vous restreindre l'analyse sur un intervalle de cette variable ?

1: Oui
2: Non
```

Sélection : 1

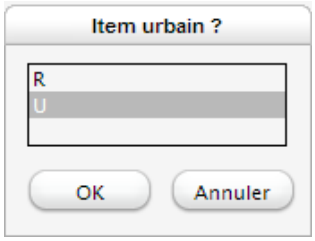
```
Deux fenêtre vont s'ouvrir. Modifiez la valeur minimum à garder, fermez la fenêtre, modifiez la
valeur maximum, fermez la fenêtre.
```





Astuce : si l'on souhaite restreindre l'analyse à d'autres variables que les variables sexe et âge, il suffit, au moment de la création du fichier individu, de désigner d'autres variables à la place du sexe et/ou de l'âge.

À l'étape suivante, vous serez invité à spécifier les items urbain et rural de la variable milieu de résidence.




```
Une fenêtre va s'ouvrir. Veuillez spécifier l'item urbain et l'item rural.
```

Quelques informations récapitulatives sont affichées à la fin de la création du tableau de données.

```
L'analyse a été restreinte aux personnes âgées de  15 - 49  ans.
Statistiques du fichier :
* 466 clusters.
* 9900 observations valides.
* Prévalence globale de 5.51%.
* Valeur de Noptimal proposée : 363
```

Vous pouvez avoir un aperçu du contenu du tableau de données obtenu à l'aide de str() et de summary().

```
> str(cm.clust)
'data.frame':   466 obs. of  9 variables:
 $ cluster       : int  1 2 3 4 5 6 7 8 9 10 ...
 $ x             : num   9.72 13.53 15.23 14.58 10.30 ...
 $ y             : num   4.04  9.10 10.33 12.77  4.52 ...
 $ residence     : Factor w/ 2 levels "Rural","Urban": 2 1 2 1 2 2 2 1 2 1 ...
 $ region        : num  3 7 5 5 6 12 5 9 12 5 ...
 $ region.name   : Factor w/ 12 levels "ADAMAOUA","CENTRE",..: 3 7 5 5 6 12 5 9 12 5 ...
 $ n             : num  9 26 31 22 10 4 9 17 16 22 ...
 $ nweight       : num  10.79 24.15 62.14 34.87  6.59 ...
 $ obs.prevalence: num  0.00 0.00 3.13 0.00 0.00 ...
```

```
> summary(cm.clust)
    cluster           x               y          residence       region
 Min.   :  1.0   Min.   : 9.025   Min.   : 2.270   Rural:222   Min.   : 1.000
 1st Qu.:117.3   1st Qu.: 9.988   1st Qu.: 4.009   Urban:244   1st Qu.: 3.000
 Median :233.5   Median :11.333   Median : 4.734               Median : 6.000
 Mean   :233.5   Mean   :11.595   Mean   : 5.581               Mean   : 6.534
 3rd Qu.:349.8   3rd Qu.:13.367   3rd Qu.: 6.425               3rd Qu.: 9.750
 Max.   :466.0   Max.   :15.446   Max.   :12.771               Max.   :12.000

      region.name        n           nweight      obs.prevalence
 DOUALA      : 46   Min.   : 2.00   Min.   : 2.335   Min.   : 0.00
 YAOUNDE     : 45   1st Qu.:14.00   1st Qu.:11.710   1st Qu.: 0.00
 EXTREME-NOR: 41   Median :20.00   Median :19.074   Median : 4.26
 OUEST       : 40   Mean   :21.24   Mean   :21.234   Mean   : 5.79
 NORD-OUEST : 39   3rd Qu.:27.00   3rd Qu.:28.353   3rd Qu.: 8.87
 CENTRE      : 38   Max.   :57.00   Max.   :94.920   Max.   :45.53
 (Other)     :217
```

Pour information, les individus ayant un résultat indéterminé ou une pondération nulle sont considérés comme manquant et ne sont donc pas comptabilisés pour le calcul de *n*, *nweight* et *obs.prevalence*.[4]

Pour le détail des différentes variables, voir l'aide de make.clust.dbf :

```
> ?make.clust.dbf
```

4.2.3 Création du fichier villes

La fonction make.cities.csv() permet de générer un tableau de données des principales villes du pays à partir d'un fichier *csv* tel que celui récupéré sur le GRUMP.

Cette fonction agit de la même manière que la précédente. Tout d'abord le contenu du fichier *csv* est affiché dans un éditeur, puis vous êtes invité à saisir les variables demandées.

```
cm.cities <- make.cities.csv('cm.cities.csv',lang='fr')
```

```
Une fenêtre va s'ouvrir présentant les données contenues dans le fichier. Merci de repérer les
variables suivantes (toutes nécessaires) :
- Nom des villes
- Longitude (en degrés au format décimal)
- Latitude (en degrés au format décimal)
- Population (effectif)
 Une fois les noms de ces variables identifiés, fermez la fenêtre pour que le programme puisse
continuer.

 Êtes-vous prêt ?

1: Oui
```

```
Sélection : 1
```

```
-------------------------------------------------

Veuillez indiquez les variables suivantes :

Nom des villes :

 1: CONTINENT    2: UNREGION     3: COUNTRY      4: UNSD         5: ISO3
 6: UQID         7: SCHNM        8: SCHADMNM     9: LATITUDE    10: LONGITUDE
11: TYPE        12: POP         13: YEAR        14: URBORRUR    15: ES90POP
16: ES95POP     17: ES00POP     18: POPSRC      19: SRCTYP      20: LOCNDATSRC
21: COORDSRCE
```

[4] Dans la suite de ce tutoriel nous appellerons observations valides les individus dont le résultat est déterminé (positif ou négatif) et dont la pondération est non nulle. Le nombre d'observations valides d'un cluster correspond donc à la variable *n*.

```
Sélection : 7

Longitude (valeur décimale) :
Sélection : 10

Latitude (valeur décimale) :
Sélection : 9

Population des villes :
Sélection : 17

ATTENTION : veuillez vérifier les informations suivantes :

* Nom des villes :
SCHNM
* Longitude (valeur decimale) :
LONGITUDE
* Latitude (valeur décimale) :
LATITUDE
* Population of cities :
ES00POP

-----------------------------------------------------

Ces données sont-elles correctes ?

1: Oui
2: Non

Sélection : 1
```

La structure du fichier créé est obtenue à l'aide de str().

```
> str(cm.cities)
'data.frame':   166 obs. of  4 variables:
 $ city.name : Factor w/ 166 levels "ABONGMBANG","AKO",..: 48 162 65 95 11 14 135 127 83 61 ...
 $ x         : num   9.7 11.5 13.4 14.3 10.4 ...
 $ y         : num   4.05  3.87  9.30 10.60  5.47 ...
 $ population: int  1512379 1213902 265294 230353 210707 206552 159663 145942 131145 107075 ...
```

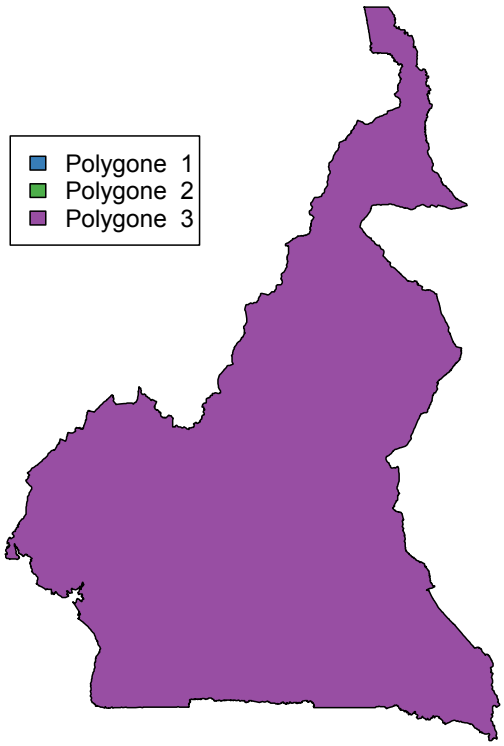
4.2.4 Création du fichier frontières du pays

Vous pouvez créer votre propre tableau de données à partir des fonctions d'importation et de manipulation des données de R. prevR utilise les frontières de la zone d'enquête sous la forme d'un *data.frame* (tableau de données) à deux colonnes (x et y) représentant une liste de points formant un polygone fermé.

Il est possible d'importer un fichier de points au format texte fourni par le Digital Chart of the World (DCW) à l'aide de la fonction make.boundary.dcw(). Le temps d'exécution de cette fonction peut prendre quelques minutes selon la puissance de votre ordinateur. Un indicateur de progression s'affiche sous forme de messages dans R.

Certains pays sont décrits par plusieurs polygones. Par exemple, la France métropolitaine sera définie par un polygone dessinant son territoire continental, un polygone formant la Corse et plusieurs petits polygones correspondant aux petites îles situées le long de la côte. prevR n'est capable de tenir compte que d'un seul polygone. Si plusieurs polygones sont détectés dans le fichier du DCW, vous êtes invité à choisir quel polygone sera utilisé (les polygones détectés étant affichés sous forme graphique).

```
> cm.bounds <- make.boundary.dcw('cameroon2pts.txt',lang='fr')
100 sur 4411 points traités.
200 sur 4411 points traités.
....
4400 sur 4411 points traités.

-------------------------------------------------------
Longitude maximale observée dans le fichier : 16.192116
Longitude minimale observée dans le fichier : 8.494763
Latitude maximale observée dans le fichier : 13.078056
Latitude minimale observée dans le fichier : 1.652548
-------------------------------------------------------

Ce fichier contient 3 polygones.
Une nouvelle fenêtre vous montre les différents polygones
contenus dans le fichier et leur numéro.
ATTENTION : certains polygones sont invisibles à l'oeil nu
(petites îles par exemple).
Veuillez sélection le polygone principal qui sera utilisé
comme limites du pays.

1: Polygone  1
2: Polygone  2
3: Polygone  3

Sélection : 3


> str(cm.bounds)
'data.frame':   4380 obs. of  2 variables:
 $ x: num  14.2 14.2 14.2 14.2 14.2 ...
 $ y: num  12.5 12.5 12.5 12.5 12.5 ...
```



NB : nous vous conseillons de sauver vos données à l'aide de la commande *Fichier* > *Sauver environnement de travail*... ou bien à partir des fonctions save() et save.image().

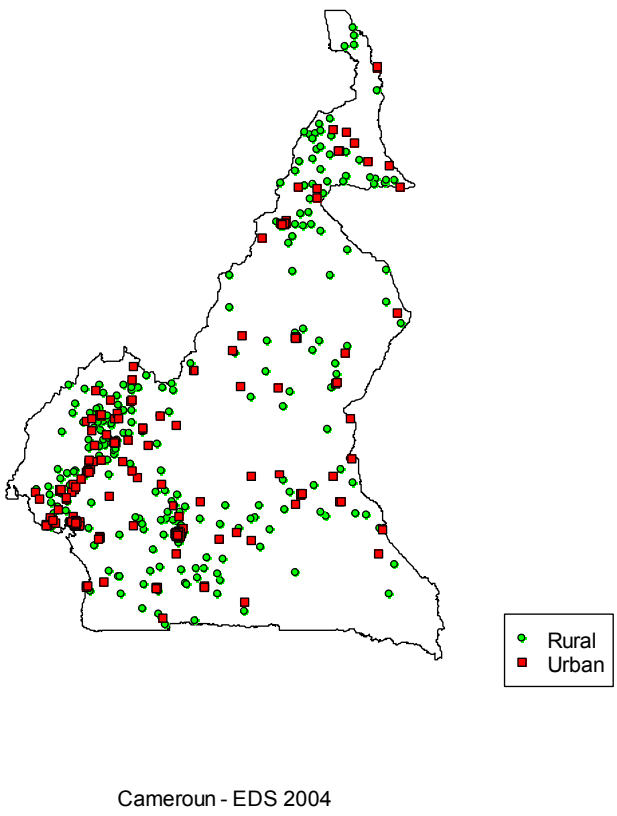
5. Cartographier le contenu du fichier cluster

Plusieurs représentations graphiques du fichier cluster peuvent être effectuées à l'aide de la fonction map.clust().

5.1 Afficher les clusters par milieu de résidence

Il suffit de préciser à la fonction map.clust() le nom du *data.frame* des clusters et celui correspondant aux frontières du pays. Un titre peut être précisé avec *main* et un sous-titre avec *sub*. D'autres options sont possibles (voir l'aide de map.clust() et celle de title()).

```
> map.clust(cm.clust,cm.bounds, lang='fr', main='Clusters par milieu de
résidence',sub='Cameroun - EDS 2004')
```



Le graphique obtenu peut être facilement exporter aux formats *emf, postscript, pdf, png, bmp* et *jpg* à l'aide du menu *Fichier* > *Sauver sous...* de la fenêtre graphique.

Pour un export au format *SVG* pour importation ultérieure dans un logiciel de dessin vectoriel (Inkscape ou Illustrator par exemple), voir l'annexe 1.

5.2 Afficher le nombre d'observations valides par cluster

Le nombre d'observations valides d'un cluster (variable *n*) correspond aux nombres d'observations avec un résultat déterminé (positif ou négatif) et une pondération non nulle (voir note 4 page 20).

Pour cela il faut modifier la variable *type* et lui attribuer la valeur *'count'*. Au passage, notez l'ajout du caractère \ devant le caractère ' dans le titre du graphique pour lui indiquer que cette apostrophe ne marque pas la fin du texte mais doit être affichée.

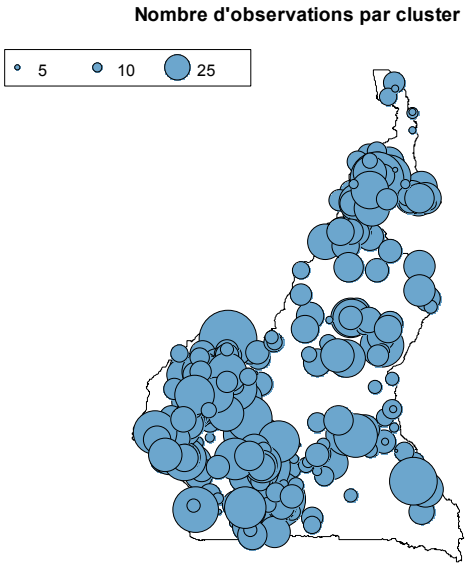
```
> map.clust(
      cm.clust,
      cm.bounds,
      type='count',
      lang='fr',
      main='Nombre d\'observations par cluster',
      sub='Cameroun - EDS 2004 - factor.size=0.2'
)
```

Cameroun - EDS 2004 - factor.size=0.2

Afin d'améliorer la lisibilité de la carte, il est possible de faire varier la taille des cercles à l'aide du paramètre *factor.size*. Sa valeur par défaut est de 0,2. Pour réduire la taille des cercles, on peut le passer à 0,15. La taille des cercles de la légende est également modifiée en conséquence.

La position de la légende peut être changée à l'aide du paramètre *legend.location*.

```
> map.clust(
      cm.clust,
      cm.bounds,
      type='count',
      lang='fr',
      factor.size=0.15,
      main='Nombre d\'observations par cluster',
      sub='Cameroun - EDS 2004 - factor.size=0.15',
      legend.location='topleft'
)
```



Cameroun - EDS 2004 - factor.size=0.15

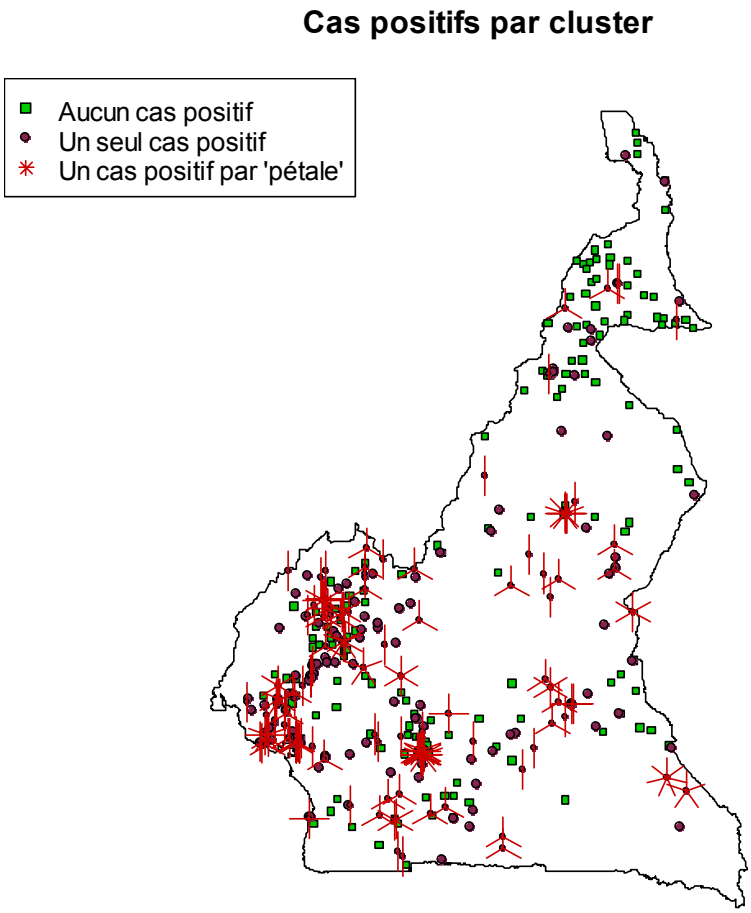
Pour exporter cette carte au format SVG puis lui appliquer une transparence avec le logiciel Inkscape, voir les annexes 1 et 2. Le fait d'appliquer une transparence aux cercles rend la lecture de la carte plus aisée et intuitive.

5.3 Représenter le nombre de cas positifs par cluster

Cette représentation graphique repose sur la fonction sunflowerplot(). Un cluster sans cas positif sera représenté par un point vert, un cluster avec un seul cas positif par un point mauve et un cluster avec plusieurs cas positifs par un point mauve et autant de « pétales » rouges que de cas positifs.

Pour cela, il suffit d'appeler map.clust() avec *type = 'flower'*.

```
> map.clust(
      cm.clust,
      cm.bounds,
      type='flower',
      lang='fr',
      main='Cas positifs par cluster',
      sub='Cameroun - EDS 2004',
      legend.location='topleft'
)
```



6. La méthode des cercles

Dans les enquêtes de type EDS, le nombre de personnes enquêtées par cluster est faible (entre 10 et 40 le plus souvent). Il en résulte que les prévalences observées dans chaque cluster (calculées à partir des personnes enquêtées de chaque cluster) varient fortement et reflètent les aléas de l'échantillonnage[5]. Pour être porteuse de sens, une prévalence nécessite d'être calculée sur un nombre suffisants d'individus, condition non remplie concernant les prévalences observées dans chaque cluster. Une interpolation spatiale réalisée à partir des prévalences observées n'apporte que peu ou pas d'information, les techniques d'interpolation spatiale classique présupposant une mesure relativement fine du phénomène étudié en chacun des points connus.

Note sur les Interpolations spatiales

Les méthodes d'interpolation spatiale permettent d'estimer les valeurs d'une variable en des points où elle n'est pas connue à partir des valeurs de cette variable aux points observés. Plus précisément, l'interpolation est la procédure qui consiste à estimer la valeur d'attribut pour des sites non échantillonnés situés à l'intérieur des limites définies par les positions de sites échantillonnés. L'extrapolation est la procédure qui consiste à estimer la valeur d'attribut pour des sites non échantillonnés situés à l'extérieur des limites définies par les positions des sites échantillonnés

Pondération selon l'inverse de la distance : il s'agit d'une méthode de moyenne pondérée ou chaque valeur de la grille à interpoler est calculée comme une moyenne pondérée des observations. Les facteurs de pondérations sont calculés proportionnellement à l'inverse de la distance élevée à une puissance. Cette méthode permet d'obtenir des grilles très rapidement mais crée des zones circulaires autour des valeurs observées (bull'eyes). Cet aspect peut être lissé en jouant sur la puissance et le voisinage. C'est un interpolateur exact (il passe par les valeurs observées).
Le **Krigeage** est une interpolation qui estime les valeurs aux points non échantillonnés par une combinaison des données. Les poids des échantillons sont pondérés par une fonction de structure qui est issue des données. On tient ainsi compte des distances, des valeurs et des corrélations. La fonction n'est pas fixée à priori mais suite à l'analyse du variogramme. On considère que la valeur estimée en un point est le produit d'un processus sous-jacent, il fournit une variance d'estimation contrairement aux autres approches. Elle permet d'appréhender la structure spatiale du phénomène étudié. Le Krigeage s'inscrit donc dans une démarche d'analyse des données géostatistique.

Les textes ci-dessus sont extraits des documents ci-dessous dont nous vous conseillons la lecture pour plus de détails :

- Cours sur l'*interpolation spatiale* de l'Université de Montréal (www.geog.umontreal.ca/donnees/geo2512/geo2512cours10.ppt).
- *Statistiques et Interpolations dans les SIG*, Laurent DRAPEAU, Centre I.RD Montpellier, Laboratoire HEA (www.faocopemed.org/vldocs/0000028/publi10.pdf).
- *Le Krigeage : la méthode optimale d'interpolation spatiale*, Yves Gratton, Institut d'Analyse Géographique (www.iag.asso.fr/pdf/krigeage_juillet2002.pdf).
- *Spatial analysis*, Wikipédia (http://en.wikipedia.org/wiki/Spatial_analysis).

Les techniques d'analyse spatiale décomposent les variations d'une variable *Z* en l'addition d'une tendance régionale *TR* et de résidus locaux *RL*. Dans le cadre des EDS, nous devons rajouter un troisième terme, l'erreur aléatoire *EA* due à l'échantillonnage.

[5] Dans la majorité des cas, les intervalles de confiance des prévalences observées sont tellement larges qu'il devient impossible de tirer la moindre conclusion.

La prévalence observée n'étant pas utilisable, l'approche développée dans prevR consiste donc à estimer une prévalence pour chaque cluster en ayant recours à une méthode dite *des cercles*. Cette dernière permet d'estimer une tendance régionale pour chaque grappe ou cluster. Dans un second temps, une fois la prévalence de chaque cluster correctement estimée, il devient possible d'appliquer des techniques classiques d'interpolation spatiale, notamment le krigeage ordinaire.

Outre une présentation succincte de la méthodologie des cercles dans le présent document, nous vous renvoyons au document suivant pour une présentation détaillée :

> Joseph Larmarange, *Prévalences du VIH en Afrique : validité d'une mesure*, chapitre 4, thèse de doctorat en démographie, sous la direction de Benoît Ferry, Université Paris Descartes, 2007, disponible sur http://joseph.larmarange.net.

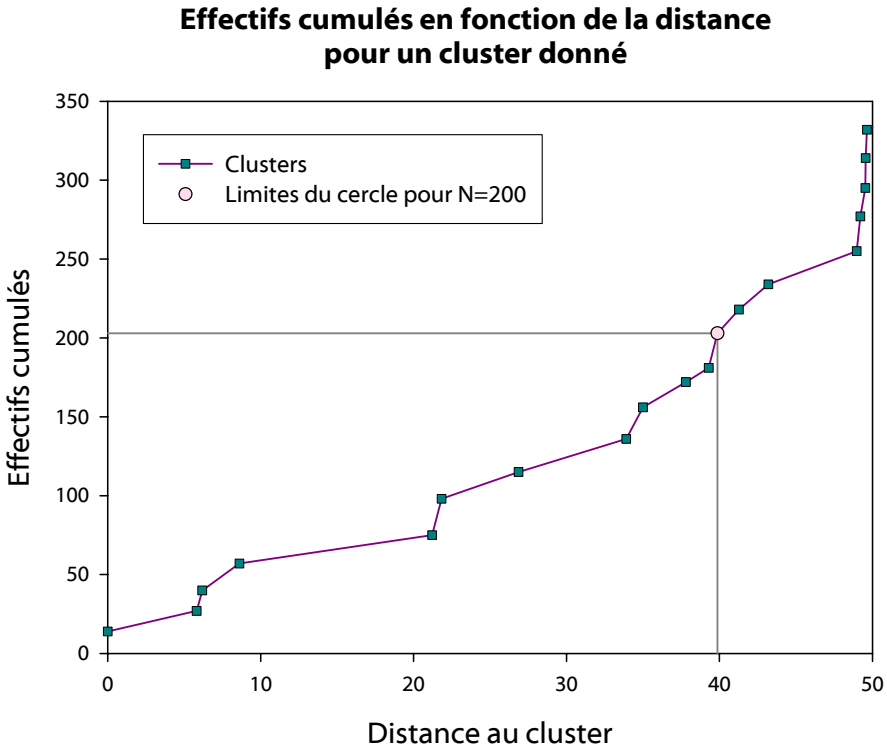
6.1 Recours à des cercles de même effectif : le paramètre *N*

S'inspirant de techniques de lissage utilisées pour le calcul de tendances régionales, l'approche de prevR consiste à tracer autour de chaque zone d'enquête un cercle puis à estimer la prévalence de cette zone à partir de l'ensemble des clusters situés à l'intérieur dudit cercle.

Plusieurs techniques de lissage utilisent des cercles de même rayon. Cependant, dans le cadre d'enquêtes de type EDS, il s'avère que les zones d'enquêtes ne sont pas uniformément réparties sur le territoire. Au contraire, leur maillage est dépendant de la densité de la population étudiée sur le territoire. Ainsi, le recours à des cercles de même rayon induirait que les prévalences seraient estimées sur un tout petit nombre d'observations dans les zones les moins peuplées et sur un très grand nombre dans les zones plus denses. Les prévalences estimées resteraient alors fortement aléatoires dans les zones peu denses, tandis que, dans les zones peuplées, il y aurait un effet d'uniformisation et une perte d'information.

Dans la mesure où c'est le nombre d'observations qui donne sens aux prévalences, l'approche de prevR privilégie le recours à des cercles de même effectif. Une fois posé un effectif *N* donné, le rayon des cercles pour chaque cluster est calculé de manière à ce que le nombre d'observations situées à l'intérieur du dit cercle soit au moins égal à *N*.

Le graphique ci-après montre, pour un cluster *x* donné, la répartition des clusters en fonction de leur distance au cluster *x* et du nombre cumulé d'observations. Si l'on a choisi une valeur de 200 pour le paramètre *N*, la prévalence du cluster *x* sera calculée sur les observations de l'ensemble des clusters situés à moins de 40 kilomètres de *x* (*x* inclus) soit sur 203 observations. Cette distance de 40 kilomètres correspond à la distance du premier cluster tel que l'effectif cumulé des observations soit supérieur à 200.



Le recours à un effectif minimum comme paramètre pour déterminer la taille des cercles permet à la fois de s'assurer que la prévalence de chaque cluster sera estimée sur un nombre d'observations suffisant et d'appliquer un niveau de lissage différent selon les zones d'enquêtes, le niveau de lissage étant déterminé par la superficie du cercle.

Lorsque l'on augmente progressivement la valeur du paramètre *N*, les prévalences estimées sont calculées sur un nombre d'observations plus important et sont lissées, atténuant ainsi les variations aléatoires de l'échantillon. Dans le même temps, elles tendent progressivement vers une valeur unique (effet d'uniformisation). Il s'agit alors de déterminer un compromis minimisant suffisamment les aléas de l'échantillonnage tout en conservant une précision locale.

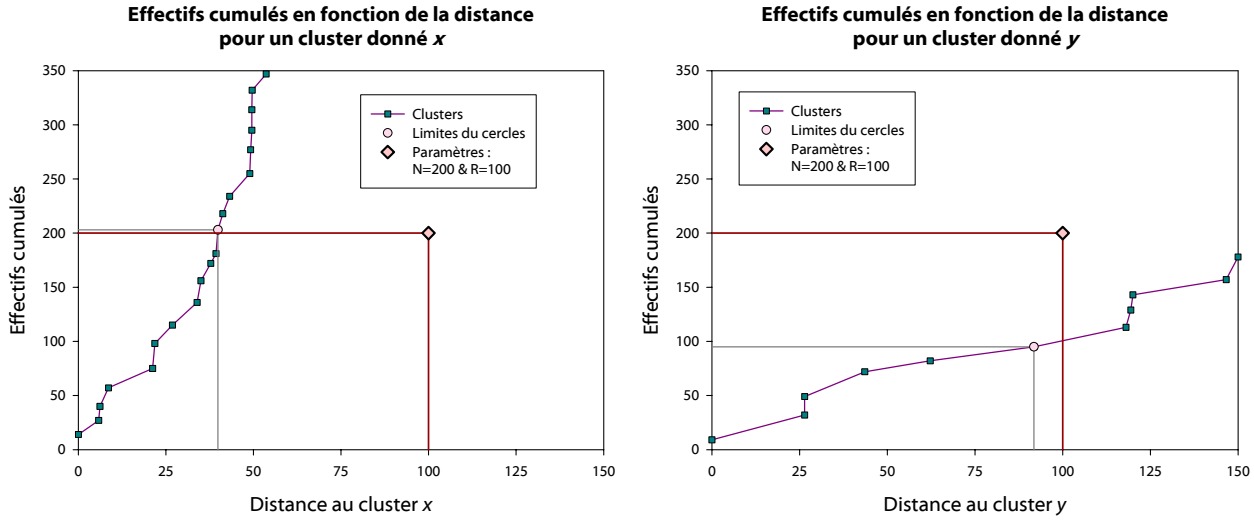
À cette fin, des modélisations et simulations d'enquêtes ont été réalisées en grand nombre. À partir de 24 500 simulations, nous avons montré qu'il était possible de déterminer une valeur optimale pour le paramètre *N* qui variait en fonction de la prévalence globale, du nombre total de personnes enquêtées et, dans une moindre mesure, du nombre de clusters.

Cette valeur peut être calculée à partir des fonctions infos.prev() et N.optim(). Elle est valable pour des enquêtes portant sur au moins 5 000 individus, 7 000 si la prévalence globale est inférieure à 2 %, 8 500 si la prévalence globale est inférieure à 1 %.

6.2 Ajout d'un rayon maximum : le paramètre *R*

Dans les zones faiblement enquêtées, notamment le long des frontières, le calcul des prévalences fait intervenir des clusters très éloignés les uns des autres. Il peut être alors préférable de limiter le lissage à des cercles plus petits, quitte à estimer, pour ces clusters là, les prévalences sur un nombre d'observations moindre.

prevR permet de rajouter au paramètre *N* un paramètre noté *R* correspondant à un rayon maximum des cercles de lissage. Deux situations peuvent alors se présenter, illustrées par les deux exemples ci-après correspondant à des valeurs de *N* et de *R* respectivement de 200 observations et de 100 kilomètres.



Pour le cluster *x*, l'effectif minimum de 200 observations est atteint pour une distance inférieure à 100 kilomètres. L'estimation de la prévalence restera donc inchangée par rapport à la situation sans le paramètre *R*. Pour le cluster *y* en revanche, très isolé, l'estimation ne sera réalisée que sur les 95 observations situées à moins de 100 kilomètres du cluster.

L'ajout du paramètre *R* servant le plus souvent à limiter la taille des cercles pour les clusters les plus isolés, une valeur adéquate de ce paramètre peut être obtenue en prenant le 9e décile des rayons des cercles de lissage lorsque seul le paramètre *N* est appliqué. Cela peut être calculé directement par la fonction infos.prev().

La technique des cercles de même effectif (paramètre *N* seulement) et celle des cercles de même rayon (paramètre *R* seulement) constituent deux cas particuliers de l'application conjointe des paramètres *N* et *R*, respectivement en attribuant la valeur *Infinie* à *R* ou à *N*.

Conseil : nous vous conseillons de produire une carte avec seulement le paramètre *N* puis avec l'utilisation conjointe de *N* et de *R*. Ensuite, aidez vous de la carte de type *flower* (voir 5.3) pour décider de retenir ou non le paramètre R.

L'ajout éventuel du paramètre R permet simplement de tester deux hypothèses concernant les zones faiblement enquêtées. En l'absence de ce paramètre, la prévalence de ces régions est fortement lissée à partir des valeurs des régions voisines. Lorsque R est appliquée, elle et plus fortement dépendante des quelques observations effectuées localement. En raison de la faible quantité de données dans ces régions, nous ne pourrons interpréter les résultats estimées dans ces zones de manière précise. Lorsque R sera appliqué, nous ne pourrons déterminer si les variations locales que nous observerons correspondent à une réalité des variations locales de l'épidémie dans ces régions ou bien s'il s'agit de variations aléatoires dues à l'échantillonnage. Cependant, cela pourra constituer des pistes de recherche à investiguer.

6.3 Prise en compte des agglomérations urbaines : le paramètre *U*

De nombreux phénomènes présentent des différentiels marqués selon le milieu de résidence. Il est possible d'observer des diffusions progressives d'une ville vers son voisinage ou bien encore la concentration d'un phénomène sur une agglomération donnée.

prevR peut prendre en compte des agglomérations urbaines pour le calcul des prévalences. Si c'est le cas, la prévalence d'un cluster situé hors agglomération ne sera calculée qu'à partir de clusters situés également hors agglomération. De manière générale, la taille des cercles sera un peu plus grande dans la mesure où les clusters appartenant à une agglomération urbaine n'auront pas été pris en compte pour calculer la taille du cercle de lissage.

Pour les clusters appartenant à une agglomération urbaine, seuls des clusters appartenant à la **même** agglomération seront pris en compte pour l'estimation des prévalences. Il importe donc que les agglomérations urbaines retenues pour l'analyse aient été suffisamment enquêtées pour que leur nombre d'observations ne soit pas trop faible. Différents critères pour retenir une agglomération urbaine dans l'analyse seront présentés dans la section 7.

Afin de pouvoir distinguer deux estimations réalisées avec des sélections différentes d'agglomérations urbaines, nous appelons *U* le nombre d'agglomérations urbaines retenues dans une estimation. La non prise en compte des agglomérations urbaines correspondra donc au cas *U*=0.

7. Estimer la prévalence de chaque cluster

La fonction permettant d'estimer la prévalence de chaque cluster par la méthode des cercles est estimate.prev(). La lecture de la documentation de cette fonction est fortement conseillée.

```
> ?estimate.prev
```

7.1 Choisir les paramètres *N* et *R*

Une valeur optimale du paramètre *N* peut être obtenue à partir d'une modélisation effectuée sur 14.000 simulations d'enquêtes type EDS (voir 6.1). Cette valeur est fournie par la fonction infos.prev().

```
> infos.prev(cm.clust, lang='fr')
Statistiques du fichier :
* 466 clusters.
* 9900 observations valides.
* Prévalence globale de 5.51%.
* Valeur de Noptimal proposée : 363
```

Vous pouvez utiliser la valeur proposée, ou bien choisir votre propre valeur.

Approche complémentaire :

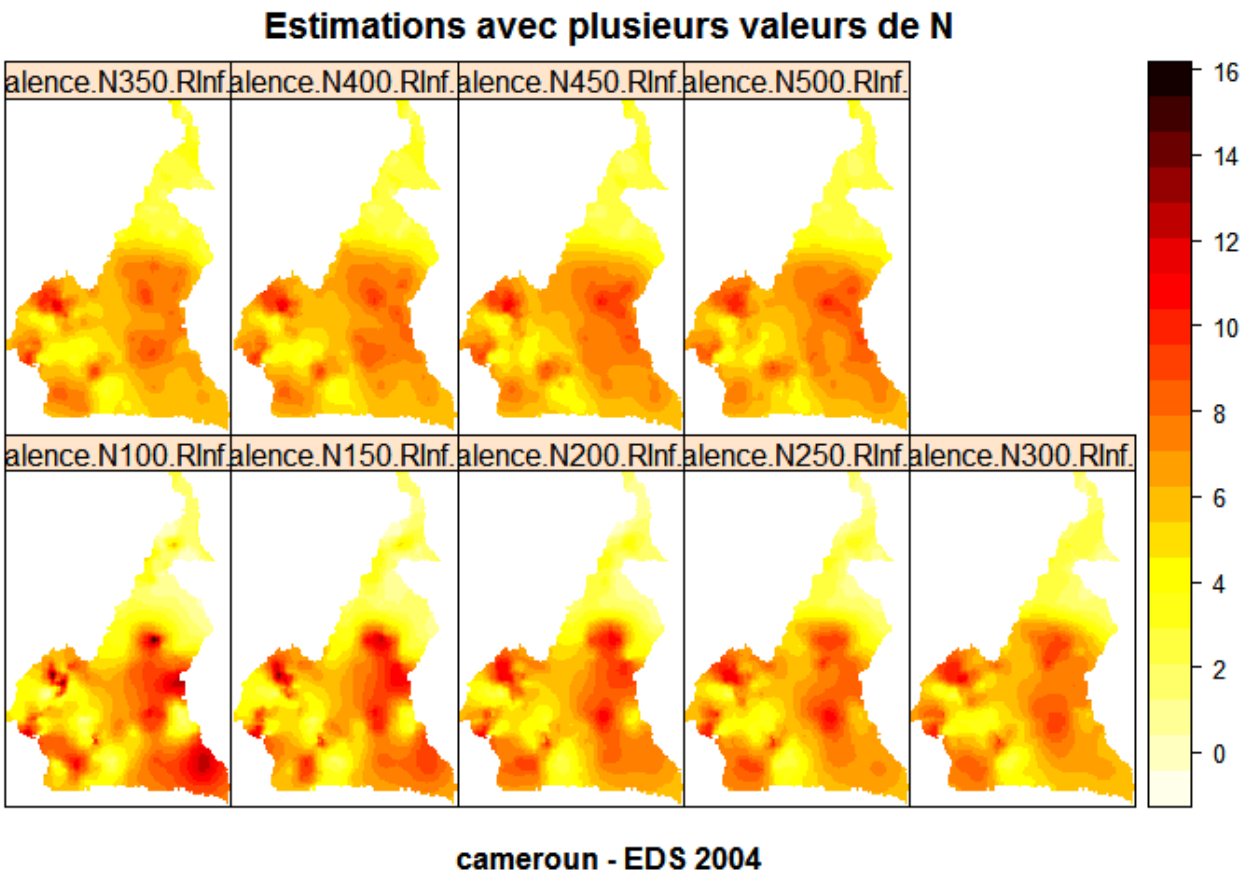
Avant de lire ce qui suit, nous vous conseillons de lire au préalable la fin de la présente section ainsi que la section suivante consacrée à l'interpolation spatiale de la prévalence.

Une possibilité consiste à réaliser plusieurs estimations avec des valeurs croissantes de *N* (par exemple faire varier *N* de 100 à 500 par pas de 50) puis à cartographier les différentes estimations réalisées pour faire son choix. La liste des valeurs de 100 à 500 par pas de 50 peut être obtenue à l'aide la fonction seq(). Par défaut, c'est-à-dire en l'absence de spécification par l'utilisateur, estimate.prev() ne tient pas compte des paramètres *R* et *U*.

```
> cm.prev.plusieursN <- estimate.prev(
      cm.clust,
      N=seq(100, 500, 50),
      lang='fr',
      merge.result = TRUE
)
> cm.krige.plusieursN <- krige.prev(
      cm.prev.plusieursN,
      c(
            est.prevalence.N100.RInf.U0~1,
            est.prevalence.N150.RInf.U0~1,
            est.prevalence.N200.RInf.U0~1,
            est.prevalence.N250.RInf.U0~1,
            est.prevalence.N300.RInf.U0~1,
            est.prevalence.N350.RInf.U0~1,
            est.prevalence.N400.RInf.U0~1,
            est.prevalence.N450.RInf.U0~1,
            est.prevalence.N500.RInf.U0~1),
      boundary = cm.bounds,
      lang = 'fr'
)
> spplot(
      cm.krige.plusieursN,
      zcol = c(
            'est.prevalence.N100.RInf.U0.pred',
            'est.prevalence.N150.RInf.U0.pred',
            'est.prevalence.N200.RInf.U0.pred',
            'est.prevalence.N250.RInf.U0.pred',
            'est.prevalence.N300.RInf.U0.pred',
            'est.prevalence.N350.RInf.U0.pred',
            'est.prevalence.N400.RInf.U0.pred',
            'est.prevalence.N450.RInf.U0.pred',
            'est.prevalence.N500.RInf.U0.pred'),
      col.regions = prevR.colors.red(21), cuts=20,
      main='Estimations avec plusieurs valeurs de N', sub='cameroun - EDS 2004'
)
```

Pour des informations sur le fonctionnement de krige.prev() et spplot(), veuillez consulter les sections 8 et 9.

On obtient ainsi la figure suivante :



Vous pouvez vous aidez de la carte de type *flower* (voir 5.3) pour vous aider à faire votre choix ainsi que des connaissances que vous avez sur l'épidémie du pays que vous étudié. Cette approche nécessite donc d'avoir une expertise préalable.

Une fois une valeur de *N* choisie (dans notre exemple nous avons choisi 363, la valeur optimale de N proposée par infos.prev()), la méthode des cercles s'applique de la manière suivante :

```
> cm.prev.N363 <- estimate.prev (cm.clust, N=363, lang='fr')

Début des calculs : 19:00:12
Cluster 25 sur 466 terminé. (19:00:13)
Cluster 50 sur 466 terminé. (19:00:14)
Cluster 75 sur 466 terminé. (19:00:14)
Cluster 100 sur 466 terminé. (19:00:14)
Cluster 125 sur 466 terminé. (19:00:15)
Cluster 150 sur 466 terminé. (19:00:15)
Cluster 175 sur 466 terminé. (19:00:16)
Cluster 200 sur 466 terminé. (19:00:16)
Cluster 225 sur 466 terminé. (19:00:17)
Cluster 250 sur 466 terminé. (19:00:17)
Cluster 275 sur 466 terminé. (19:00:18)
Cluster 300 sur 466 terminé. (19:00:18)
Cluster 325 sur 466 terminé. (19:00:19)
Cluster 350 sur 466 terminé. (19:00:19)
Cluster 375 sur 466 terminé. (19:00:20)
Cluster 400 sur 466 terminé. (19:00:20)
Cluster 425 sur 466 terminé. (19:00:21)
Cluster 450 sur 466 terminé. (19:00:21)

Fin des calculs : 19:00:22
Temp de calcul : 9.38 secs.
```

Lorsque plusieurs estimations sont réalisées simultanément, les temps de calcul peuvent devenir relativement longs. Un indicateur de progression est affiché à l'écran. Il peut être désactivé avec le paramètre progression = FALSE.

Le tableau de données renvoyé est de la forme :

```
> str(cm.prev.N363)
'data.frame':   466 obs. of  17 variables:
 $ cluster           : int  1 2 3 4 5 6 7 8 9 10 ...
 $ x                 : num   9.72 13.53 15.23 14.58 10.30 ...
 $ y                 : num   4.04  9.10 10.33 12.77  4.52 ...
 $ residence         : Factor w/ 2 levels "Rural","Urban": 2 1 2 1 2 2 2 1 2 1 ...
 $ region            : num  3 7 5 5 6 12 5 9 12 5 ...
 $ region.name       : Factor w/ 12 levels "ADAMAOUA","CENTRE",..: 3 7 5 5 6 12 5 9 12 5 ...
 $ n                 : num  9 26 31 22 10 4 9 17 16 22 ...
 $ nweight           : num  10.79 24.15 62.14 34.87  6.59 ...
 $ obs.prevalence    : num  0.00 0.00 3.13 0.00 0.00 ...
 $ est.prevalence    : num  5.46 1.90 2.00 1.83 3.81 ...
 $ circle.count      : num  380 380 380 363 370 374 391 365 371 369 ...
 $ circle.radius     : num    3.11  59.01 102.51 242.89  69.52 ...
 $ circle.nb.clusters: int  17 18 16 21 21 20 17 17 19 16 ...
 $ quality.indicator : num     0.498  178.620  539.037 3096.468  251.253 ...
 $ N.parameter       : num  363 363 363 363 363 363 363 363 363 363 ...
 $ R.parameter       : num  Inf Inf Inf Inf Inf ...
 $ U.parameter       : num  0 0 0 0 0 0 0 0 0 0 ...
```

8 variables ont été ajoutées à cm.clust :

- *est.prevalence* qui correspond à la prévalence estimée par la méthode des cercles ;
- *circle.count* qui correspond au nombre d'observations valides sur lesquelles la prévalence a été calculée ;
- *circle.radius* qui correspond au rayon du cercle de lissage de chaque cluster ;
- *circle.nb.clust* qui correspond au nombre de clusters inclus dans le cercle de lissage ;
- *quality indicator :* il s'agit d'un indicateur de qualité, calculé pour chaque cluster comme le carré de *circle.radius* divisé par la racine de *circle.count*. Une valeur élevée de cet indicateur indique que la prévalence estimée a été calculée en prenant en compte des clusters éloignés les uns des autres et que le nombre d'observations était faible, d'où une estimation incertaine et peu précise. Au contraire, une valeur faible de cet indicateur sera obtenue pour des estimations calculées à partir d'observations suffisantes et proches les unes des autres, d'où une information précise et locale.
- *N.parameter,*
- *R.parameter* et
- *U.parameter.*

La fonction info.prev() renvoie des informations supplémentaires, notamment les quantiles des rayon des cercles de lissage. Si l'on décide d'utiliser la valeur du 9e décile pour choisir le paramètre *R*, on pourra retenir dans cet exemple la valeur de 118 kilomètres.

```
> infos.prev(cm.prev.N363, lang='fr')
Statistiques du fichier :
* 466 clusters.
* 9900 observations valides.
* Prévalence globale de 5.51%.
* Valeur de Noptimal proposée : 363
* Quantiles des rayons des cercles de lissage :
  50%   75%   80%   85%   90%   95%   99%
 50.6  81.3  94.7 102.5 117.8 137.3 205.7
```

Par défaut, les distances sont exprimées en kilomètres. Cependant, il est possible d'obtenir des miles en utilisant le paramètre miles = TRUE. Si les coordonnées des clusters ne sont pas exprimées en degrés décimaux mais dans un autre référentiel, utilisez le paramètre dist.fonction = 'rdist' pour calculer des distances euclidiennes. Les distances renvoyées par estimate.prev() seront alors exprimées dans l'unité du système de coordonnées.

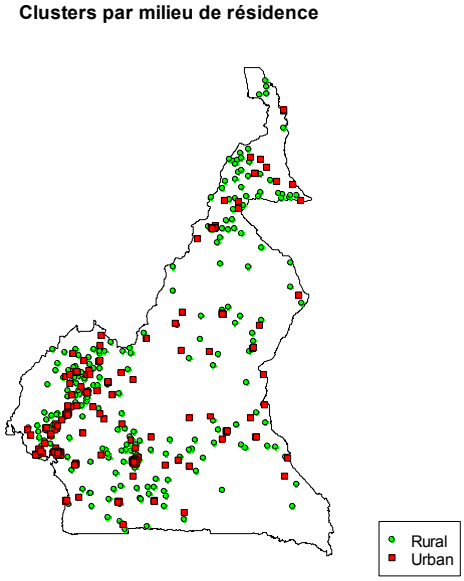
Si les noms des colonnes de votre tableau de données des clusters diffèrent des noms par défaut, vous devrez préciser les noms de vos variables avec var.clust et urban.area.code (voir la documentation de estimate.prev() pour plus de précisions).

7.2 Choix des agglomérations urbaines pour le paramètre *U*

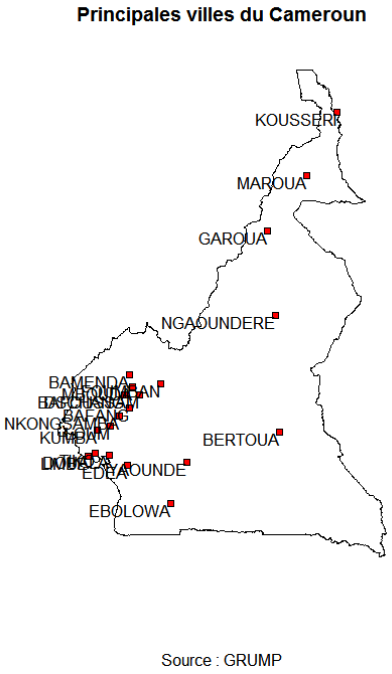
7.2.1 Carte des villes et carte des clusters par milieu de résidence

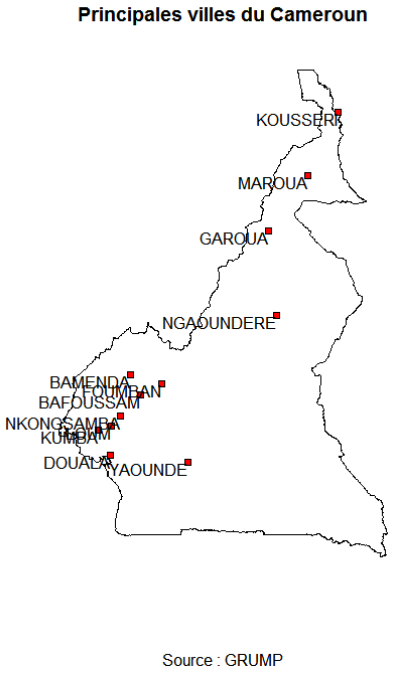
Les fonctions map.clust() et map.cities() permettent, dans un premier temps, de représenter la répartition des clusters par milieu de résidence et la position des principales villes du pays.

```
> map.clust(cm.clust,cm.bounds,main='Clusters par milieu de résidence',sub='Cameroun -
DHS 2004')
```



```
> map.cities(cm.cities, cm.bounds, min.population=65000,
      main='Principales villes du Cameroun', sub='Source : GRUMP')
> map.cities(cm.cities, cm.bounds, min.population=100000,
      main='Principales villes du Cameroun', sub='Source : GRUMP')
```





7.2.2 Recoder le milieu de résidence

Dans les variables disponibles dans les EDS, on ne peut savoir si un cluster appartient ou non à une ville donnée. Le milieu de résidence (urbain ou rural) est fourni mais ne peut être utilisé directement. En effet, cette dichotomie repose sur la définition en vigueur dans chaque pays et inclut le plus souvent des communes de taille moyenne et/ou chefs-lieux d'entité administrative.

Le projet GRUMP fournit les coordonnées longitude/latitude du centre des principales villes de chaque pays. prevR considère qu'une agglomération urbaine sera composée par les clusters urbains situés à une distance inférieure à 15 kilomètres du point central de la ville considérée. La recodification des clusters selon leur appartenance ou non à une agglomération urbaine est effectuée à l'aide de la fonction calcul.dist.cities(). La distance de 15 kilomètres ne représente pas la taille des agglomérations. Il s'agit plus précisément d'un critère de démarcation permettant de distinguer les clusters urbains appartenant à une agglomération des autres clusters urbains. Il peut bien sûr être modifié en fonction du contexte propre à chaque analyse, à travers le paramètre dist.

Dans un premier temps, nous conseillons de coder l'appartenance à une agglomération urbaine en prenant en compte un nombre important de villes. Lorsque la fonction calcul.dist.cities est appelée, une fenêtre s'ouvre et vous invite à choisir les villes pour lesquelles vous voulez calculer l'appartenance ou non à l'agglomération urbaine. Il est également possible de demander à la fonction de calculer cela pour l'ensemble des villes présentes dans le fichier villes avec type = 'all'.

```
> cm.clust <- calcul.dist.cities(cm.clust, cm.cities, lang='fr')
Une fenêtre va s'ouvrir. Veuillez sélectionner les villes retnues pour le paramètre U (utilisez les
touches CTRL et SHIFT).
```

```
> str(cm.clust)
'data.frame':   466 obs. of  12 variables:
 $ cluster       : int  1 2 3 4 5 6 7 8 9 10 ...
 $ x             : num   9.72 13.53 15.23 14.58 10.30 ...
 $ y             : num   4.04  9.10 10.33 12.77  4.52 ...
 $ residence     : Factor w/ 2 levels "Rural","Urban": 2 1 2 1 2 2 2 1 2 1 ...
 $ region        : num  3 7 5 5 6 12 5 9 12 5 ...
 $ region.name   : Factor w/ 12 levels "ADAMAOUA","CENTRE",..: 3 7 5 5 6 12 5 9 ...
 $ n             : num  9 26 31 22 10 4 9 17 16 22 ...
 $ nweight       : num  10.79 24.15 62.14 34.87  6.59 ...
 $ obs.prevalence: num  0.00 0.00 3.13 0.00 0.00 ...
 $ dist.city     : num    1.87  26.17 102.64  90.79  62.26 ...
 $ city.name     : chr  "DOUALA" "GAROUA" "MAROUA" "KOUSSERI" ...
 $ urban.area    : Factor w/ 2 levels "in urban area",..: 1 2 2 2 2 1 2 2 1 2 ...
```

```
> infos.prev(cm.clust, lang='fr')
Statistiques du fichier :
* 466 clusters.
* 9900 observations valides.
* Prévalence globale de 5.51%.
* Valeur de Noptimal proposée : 363
* Nombre d'agglomérations urbaines trouvées : 22
* Noms des agglomérations urbaines trouvées :
BAFANG, BAFOUSSAM, BAMENDA, BERTOUA, BUEA, DOUALA, DSCHANG, EBOLOWA, EDEA, FOUMBAN,
GAROUA, KOUSSERI, KUMBA, KUMBO, LIMBE, LOUM, MAROUA, MBOUDA, NGAOUNDERE, NKONGSAMBA,
TIKO, YAOUNDE.
```

calcul.dist.cities() ajoute trois colonnes au tableau de données cluster qui lui est fourni en entrée :

- *dist.city* : la distance à la ville la plus proche ;
- *city.name* : nom de la ville la plus proche ;
- *urban.area :* l'appartenance ou nom à l'agglomération urbaine de la ville la plus proche.

infos.prev() détecte la présence ou non de ces trois variables et affiche le cas échéant le nombre d'agglomérations urbaines retenues ainsi que leurs noms.

7.2.3 Choix des agglomérations urbaines

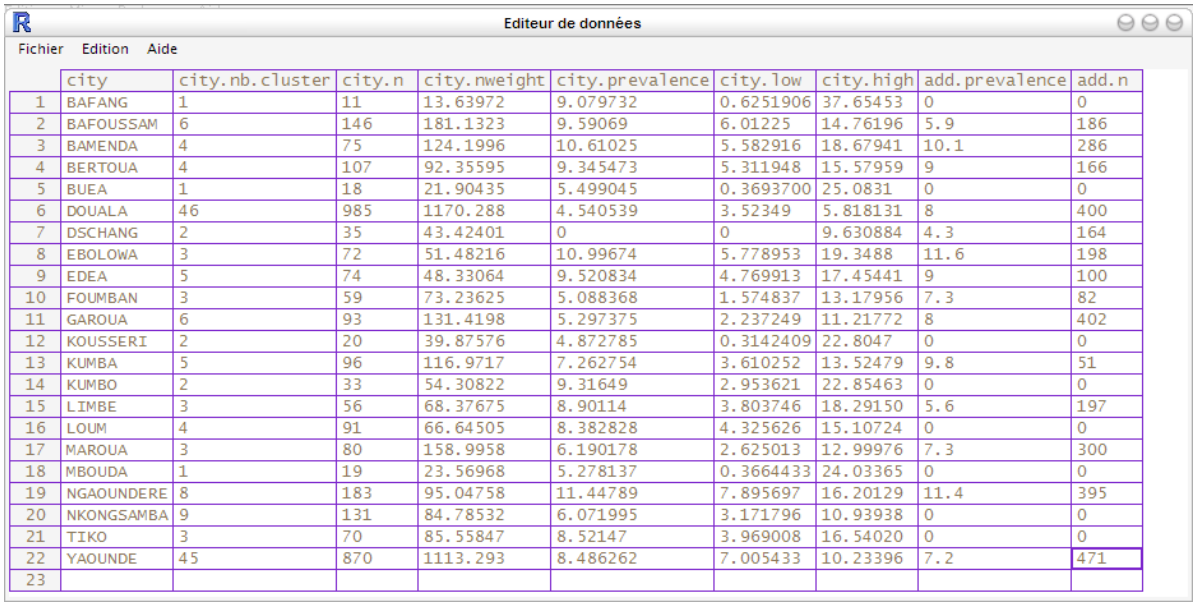
La fonction verif.urb() permet de calculer pour chaque agglomération urbaine le nombre de clusters concernés, le nombre d'observations valides, la prévalence observée sur l'agglomération et un intervalle de confiance de celle-ci. Le niveau de confiance de ce dernier peut être modifié avec le paramètre conf.level (0,9 soit 90 % par défaut).

```
> verif.urb(cm.clust)
         city city.nb.cluster city.n city.nweight city.prevalence  city.low city.high conf.level
1      BAFANG               1     11     13.63972        9.079732 0.6251906 37.654535        0.9
2   BAFOUSSAM               6    146    181.13233        9.590690 6.0122504 14.761959        0.9
3     BAMENDA               4     75    124.19958       10.610245 5.5829161 18.679413        0.9
4     BERTOUA               4    107     92.35595        9.345473 5.3119485 15.579592        0.9
5        BUEA               1     18     21.90435        5.499045 0.3693700 25.083099        0.9
6      DOUALA              46    985   1170.28768        4.540539 3.5234900  5.818131        0.9
7     DSCHANG               2     35     43.42401        0.000000 0.0000000  9.630884        0.9
8     EBOLOWA               3     72     51.48216       10.996741 5.7789535 19.348799        0.9
9        EDEA               5     74     48.33064        9.520834 4.7699133 17.454408        0.9
10    FOUMBAN               3     59     73.23625        5.088368 1.5748370 13.179564        0.9
11     GAROUA               6     93    131.41981        5.297375 2.2372492 11.217723        0.9
12   KOUSSERI               2     20     39.87576        4.872785 0.3142409 22.804702        0.9
13      KUMBA               5     96    116.97174        7.262754 3.6102521 13.524788        0.9
14      KUMBO               2     33     54.30822        9.316490 2.9536209 22.854626        0.9
```

```
15       LIMBE                3     56      68.37675         8.901140 3.8037463 18.291504         0.9
16        LOUM                4     91      66.64505         8.382828 4.3256258 15.107244         0.9
17      MAROUA                3     80     158.99576         6.190178 2.6250128 12.999756         0.9
18      MBOUDA                1     19      23.56968         5.278137 0.3664433 24.033647         0.9
19 NGAOUNDERE                8    183      95.04758        11.447887 7.8956974 16.201288         0.9
20 NKONGSAMBA                9    131      84.78532         6.071995 3.1717964 10.939380         0.9
21        TIKO                3     70      85.55847         8.521471 3.9690084 16.540203         0.9
22    YAOUNDE                45    870    1113.29302         8.486262 7.0054333 10.233962         0.9
```

Lorsque l'on dispose de données complémentaires provenant d'une autre source, il est possible de les ajouter afin de procéder à une comparaison, avec le paramètre add = TRUE. Une fenêtre s'ouvrira alors pour saisir la prévalence et l'effectif pour chaque agglomération de cette autre source. Dans notre exemple, nous avons ajouté, pour comparaison, des données provenant de la surveillance sentinelle des femmes enceintes[6].

```
> verif.urb(cm.clust, add=TRUE, lang='fr')
Une fenêtre va s'ouvrir. Compléter les colonnes 'add.prevalence' et 'add.n' pour chaque ville, puis
fermer. Pour 'add.prevalence', rentrer les données en %. Par exemple, pour 0,034=3,4%, saisir 3.4.
```



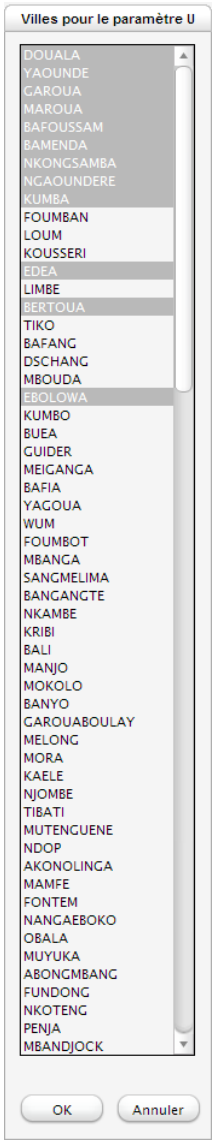
	city	city.nb.cluster	city.n	city.nweight	city.prevalence	city.low	city.high	add.prevalence	add.n
1	BAFANG	1	11	13.63972	9.079732	0.6251906	37.65453	0	0
2	BAFOUSSAM	6	146	181.1323	9.59069	6.01225	14.76196	5.9	186
3	BAMENDA	4	75	124.1996	10.61025	5.582916	18.67941	10.1	286
4	BERTOUA	4	107	92.35595	9.345473	5.311948	15.57959	9	166
5	BUEA	1	18	21.90435	5.499045	0.3693700	25.0831	0	0
6	DOUALA	46	985	1170.288	4.540539	3.52349	5.818131	8	400
7	DSCHANG	2	35	43.42401	0	0	9.630884	4.3	164
8	EBOLOWA	3	72	51.48216	10.99674	5.778953	19.3488	11.6	198
9	EDEA	5	74	48.33064	9.520834	4.769913	17.45441	9	100
10	FOUMBAN	3	59	73.23625	5.088368	1.574837	13.17956	7.3	82
11	GAROUA	6	93	131.4198	5.297375	2.237249	11.21772	8	402
12	KOUSSERI	2	20	39.87576	4.872785	0.3142409	22.8047	0	0
13	KUMBA	5	96	116.9717	7.262754	3.610252	13.52479	9.8	51
14	KUMBO	2	33	54.30822	9.31649	2.953621	22.85463	0	0
15	LIMBE	3	56	68.37675	8.90114	3.803746	18.29150	5.6	197
16	LOUM	4	91	66.64505	8.382828	4.325626	15.10724	0	0
17	MAROUA	3	80	158.9958	6.190178	2.625013	12.99976	7.3	300
18	MBOUDA	1	19	23.56968	5.278137	0.3664433	24.03365	0	0
19	NGAOUNDERE	8	183	95.04758	11.44789	7.895697	16.20129	11.4	395
20	NKONGSAMBA	9	131	84.78532	6.071995	3.171796	10.93938	0	0
21	TIKO	3	70	85.55847	8.52147	3.969008	16.54020	0	0
22	YAOUNDE	45	870	1113.293	8.486262	7.005433	10.23396	7.2	471
23									

```
         city city.nb.cluster city.n city.nweight city.prevalence  city.low city.high
1      BAFANG                1     11     13.63972        9.079732 0.6251906 37.654535
2   BAFOUSSAM                6    146    181.13233        9.590690 6.0122504 14.761959
3     BAMENDA                4     75    124.19958       10.610245 5.5829161 18.679413
4     BERTOUA                4    107     92.35595        9.345473 5.3119485 15.579592
5        BUEA                1     18     21.90435        5.499045 0.3693700 25.083099
6      DOUALA               46    985   1170.28768        4.540539 3.5234900  5.818131
7     DSCHANG                2     35     43.42401        0.000000 0.0000000  9.630884
8     EBOLOWA                3     72     51.48216       10.996741 5.7789535 19.348799
9        EDEA                5     74     48.33064        9.520834 4.7699133 17.454408
10    FOUMBAN                3     59     73.23625        5.088368 1.5748370 13.179564
11     GAROUA                6     93    131.41981        5.297375 2.2372492 11.217723
12   KOUSSERI                2     20     39.87576        4.872785 0.3142409 22.804702
13      KUMBA                5     96    116.97174        7.262754 3.6102521 13.524788
14      KUMBO                2     33     54.30822        9.316490 2.9536209 22.854626
15      LIMBE                3     56     68.37675        8.901140 3.8037463 18.291504
16       LOUM                4     91     66.64505        8.382828 4.3256258 15.107244
17     MAROUA                3     80    158.99576        6.190178 2.6250128 12.999756
18     MBOUDA                1     19     23.56968        5.278137 0.3664433 24.033647
19 NGAOUNDERE                8    183     95.04758       11.447887 7.8956974 16.201288
20 NKONGSAMBA                9    131     84.78532        6.071995 3.1717964 10.939380
21       TIKO                3     70     85.55847        8.521471 3.9690084 16.540203
22    YAOUNDE               45    870   1113.29302        8.486262 7.0054333 10.233962
```

[6] Suivant les agglomérations, des données datant de 2002 et lorsqu'elles n'étaient pas disponibles de 2000 : National AIDS Control Committe. **National HIV Sentinel surveillance Report (2002)**. Yaoundé. Ministry of Public Health, 2003, 42 pages. & National AIDS Control Committee. **Technical Report national Serosurvey on VIH/Syphilis (2000)**. Yaoundé. Ministry of Public Health, 2000, 219 pages.

```
   add.prevalence add.n  add.low  add.high p.value.comparison conf.level
1             0.0     0       NA        NA                NA        0.9
2             5.9   186 3.445394  9.762477         0.21670308        0.9
3            10.1   286 7.416399 13.654325         0.83359063        0.9
4             9.0   166 5.760419 13.731306         1.00000000        0.9
5             0.0     0       NA        NA                NA        0.9
6             8.0   400 5.932209 10.663398         0.01391705        0.9
7             4.3   164 2.115241  8.071971         0.60877976        0.9
8            11.6   198 8.163050 16.172756         1.00000000        0.9
9             9.0   100 4.936702 15.444009         1.00000000        0.9
10            7.3    82 3.410503 14.282425         0.73451911        0.9
11            8.0   402 5.902433 10.611184         0.51342936        0.9
12            0.0     0       NA        NA                NA        0.9
13            9.8    51 4.202260 19.996781         0.75275672        0.9
14            0.0     0       NA        NA                NA        0.9
15            5.6   197 3.251432  9.229574         0.35883764        0.9
16            0.0     0       NA        NA                NA        0.9
17            7.3   300 5.081109 10.398776         1.00000000        0.9
18            0.0     0       NA        NA                NA        0.9
19           11.4   395 8.908253 14.426918         1.00000000        0.9
20            0.0     0       NA        NA                NA        0.9
21            0.0     0       NA        NA                NA        0.9
22            7.2   471 5.399582  9.552642         0.46218264        0.9
```

Les informations renvoyées par verif.urb() peuvent guider le choix des agglomérations retenues pour le paramètre *U*. Il ne faut pas perdre de vue que, si une agglomération est retenue pour *U*, alors la prévalence d'un cluster de cette agglomération sera calculée uniquement à partir de clusters de cette même agglomération.



Plusieurs critères peuvent être pris en compte pour sélectionner les agglomérations urbaines :

- Tout d'abord, et c'est le plus évident, les villes pour lesquelles les nombres de clusters et d'observations valides sont importants. Dans le cas présent, nous pouvons décider de retenir les agglomérations comportant au moins 100 observations valides réparties sur au moins 6 clusters. Il s'agit de Douala, Yaoundé, Bafoussam, N'Gaoundéré et Nkongsamba.
- Il est également possible de retenir des agglomérations qui, bien que comportant un nombre d'observations valides plus faible, présentent une prévalence proche de celle calculée à partir d'une autre source de données. Dans notre exemple, nous pouvons retenir ainsi Bamenda, Bertoua, Kumba, Edéa et Ebolowa.
- Enfin, il est possible de retenir des agglomérations où la prévalence observée, bien qu'inférieure à celle mesurée par une autre source, reste supérieure à celle de son voisinage. La prise en compte de ces agglomérations permettra alors de les faire ressortir sur la carte produite. Nous retiendrons selon ce critère les agglomérations de Garoua et Maroua.

Pour les autres agglomérations, l'absence de données suffisantes dans l'EDS couplée à une absence de données complémentaires où à une contradiction avec la source complémentaire ne permet pas de se positionner. Il est alors préférable de ne pas les inclure pour le paramètre *U* et d'estimer leur prévalence à partir des clusters voisins. Au moment de l'interprétation des cartes produites, il conviendra de préciser que les tendances affichées sur la carte pour les zones

correspondant à ces agglomérations correspondront aux tendances régionales de la zone et non à la tendance limitée à l'agglomération elle-même.

La sélection des agglomérations urbaines retenues pour le paramètre U résulte d'un choix raisonné selon l'hypothèse que la prévalence observée au sein d'une agglomération rends compte de la réalité épidémique de cette agglomération, soit parce que la quantité d'information présente dans l'EDS est suffisante, soit parce que d'autres sources d'informations rendent cette hypothèse crédible. Il est possible de réaliser plusieurs cartes avec des sélections différentes pour le paramètre U afin de tester plusieurs hypothèses.

Une fois les agglomérations choisies pour la paramètre U, il reste à appliquer de nouveau la fonction calcul.dist.cities(), limitée cette fois-ci aux agglomérations retenues.

```
> cm.clust <- calcul.dist.cities(cm.clust, cm.cities, lang='fr')
Une fenêtre va s'ouvrir. Veuillez sélectionner les villes retnues pour le paramètre U (utilisez les
touches CTRL et SHIFT).

> infos.prev(cm.clust, lang='fr')
Statistiques du fichier :
* 466 clusters.
* 9900 observations valides.
* Prévalence globale de 5.51%.
* Valeur de Noptimal proposée : 363
* Nombre d'agglomérations urbaines trouvées : 12
* Noms des agglomérations urbaines trouvées :
BAFOUSSAM, BAMENDA, BERTOUA, DOUALA, EBOLOWA, EDEA, GAROUA, KUMBA, MAROUA, NGAOUNDERE, NKONGSAMBA,
YAOUNDE.
```

7.3 Réaliser plusieurs estimations simultanément

Ayant choisi les différents paramètres, il est possible d'appeler la fonction estimate.prev() avec plusieurs valeurs de ces derniers. Ainsi, il sera possible de cartographier par la suite les différentes estimations, lorsque seul le paramètre *N* est pris en compte, lorsque l'on tient compte de *N* et *R* et enfin lorsqu'on rajoute le paramètre *U*.

Les options N et R de estimate.prev peuvent prendre une valeur numérique ou une liste de valeurs numériques. L'estimation sera réalisée pour chaque combinaison des deux paramètres. Pour ne pas prendre en compte un de ces deux paramètres, il suffit de lui spécifier la valeur infinie Inf. Pour spécifier une liste, le plus simple consiste à utiliser la fonction c() (voir les exemples plus loin).

L'option U peut prendre trois valeurs : TRUE (le paramètre *U* est pris en compte), FALSE (il ne l'est pas), 2 (les estimations sont réalisées deux fois, une fois en tenant compte de *U*, une fois sans).

Lorsque plusieurs estimations sont réalisées simultanément, estimate.prev() renvoie un tableau de données avec une ligne par cluster **et** par estimation et fourni pour chaque cluster et chaque estimation les valeurs de la prévalence estimée (*est.prevalence*), le nombre d'observations retenues

(*circle.count*), le nombre de clusters inclus dans le cercle (*circle.nb.clusters*) et le rayon du cercle de lissage (*circle.radius*).

Avant de pouvoir réaliser une interpolation spatiale, il est nécessaire d'utiliser la fonction extract.data(), qui permet d'extraire une seule estimation du tableau de données, ou bien la fonction merge.prev(), qui réarrange les données de manière à obtenir un tableau de données avec une ligne par cluster. Les nouvelles colonnes créées par estimate.prev() sont alors dupliquées pour chaque estimation et leur nom prend alors un suffixe de la forme *.Nvaleur-de-n.Rvaleur-de-r.Uvaleur-de-U*. merge.prev() peut être directement appliquée aux résultats renvoyés par estimate.prev() en précisant merge.result = TRUE (voir les deux exemples ci-après).

```
> cm.prev <- estimate.prev(cm.clust,N=363, R=c(118, Inf), U=2, lang='fr')
> str(cm.prev)
'data.frame':   1864 obs. of  22 variables:
 $ cluster           : int  1 2 3 4 5 6 7 8 9 10 ...
 $ x                 : num   9.72 13.53 15.23 14.58 10.30 ...
 $ y                 : num   4.04  9.10 10.33 12.77  4.52 ...
 $ residence         : Factor w/ 2 levels "Rural","Urban": 2 1 2 1 2 2 2 1 2 1 ...
 $ region            : num  3 7 5 5 6 12 5 9 12 5 ...
 $ region.name       : Factor w/ 12 levels "ADAMAOUA","CENTRE",..: 3 7 5 5 6 12 5 9 12 5 ...
 $ longitude         : num  NA NA NA NA NA NA NA NA NA NA ...
 $ latitude          : num  NA NA NA NA NA NA NA NA NA NA ...
 $ n                 : num  9 26 31 22 10 4 9 17 16 22 ...
 $ nweight           : num  10.79 24.15 62.14 34.87  6.59 ...
 $ obs.prevalence    : num  0.00 0.00 3.13 0.00 0.00 ...
 $ dist.city         : num    1.87  26.17 102.64  90.79  62.26 ...
 $ city.name         : chr  "DOUALA" "GAROUA" "MAROUA" "KOUSSERI" ...
 $ urban.area        : Factor w/ 2 levels "in urban area",..: 1 2 2 2 2 1 2 2 1 2 ...
 $ est.prevalence    : num  5.46 1.90 2.00 3.01 3.81 ...
 $ circle.count      : num  380 380 380 69 370 374 391 365 371 369 ...
 $ circle.radius     : num    3.11  59.01 102.51  91.60  69.52 ...
 $ circle.nb.clusters: int  17 18 16 6 21 20 17 17 19 16 ...
 $ quality.indicator : num     0.498  178.620  539.037 1010.180  251.253 ...
 $ N.parameter       : num  363 363 363 363 363 363 363 363 363 363 ...
 $ R.parameter       : num  118 118 118 118 118 118 118 118 118 118 ...
 $ U.parameter       : num  0 0 0 0 0 0 0 0 0 0 ...

> cm.prev <- estimate.prev(cm.clust,N=363, R=c(118, Inf), U=2, lang='fr',
merge.result=TRUE)
> str(cm.prev)
'data.frame':   466 obs. of  32 variables:
 $ cluster                         : int  1 10 100 101 102 103 104 105 106 107 ...
 $ x                               : num   9.72 13.57 11.23 14.71 11.55 ...
 $ y                               : num   4.04 10.25  4.74 10.43  3.88 ...
 $ residence                       : Factor w/ 2 levels "Rural","Urban": 2 1 2 1 2 2 2 2 1 2 ...
 $ region                          : num  3 5 2 5 12 7 8 12 7 3 ...
 $ region.name                     : Factor w/ 12 levels "ADAMAOUA","CENTRE",..: 3 5 2 5 12 7 8 ...
 $ n                               : num  9 22 18 8 17 36 57 16 16 10 ...
 $ nweight                         : num  10.8 34.8 28.7 12.7 22.1 ...
 $ obs.prevalence                  : num   0.00 13.63  0.00  0.00  5.81 ...
 $ dist.city                       : num    1.87  91.90 102.20  45.20   3.59 ...
 $ city.name                       : chr  "DOUALA" "MAROUA" "YAOUNDE" "MAROUA" ...
 $ urban.area                      : Factor w/ 2 levels "in urban area",..: 1 2 2 2 1 2 2 1 2 1 ...
 $ est.prevalence.N363.R118.U0     : num  5.46 1.94 3.16 2.49 7.61 ...
 $ circle.count.N363.R118.U0       : num  380 369 375 379 367 376 387 370 380 386 ...
 $ circle.radius.N363.R118.U0      : num   3.11 70.93 69.52 61.67  3.73 ...
 $ circle.nb.clusters.N363.R118.U0 : int  17 16 15 15 17 18 18 16 18 18 ...
 $ quality.indicator.N363.R118.U0  : num    0.498 261.887 249.583 195.371   0.727 ...
 $ est.prevalence.N363.RInf.U0     : num  5.46 1.94 3.16 2.49 7.61 ...
 $ circle.count.N363.RInf.U0       : num  380 369 375 379 367 376 387 370 380 386 ...
 $ circle.radius.N363.RInf.U0      : num   3.11 70.93 69.52 61.67  3.73 ...
 $ circle.nb.clusters.N363.RInf.U0 : int  17 16 15 15 17 18 18 16 18 18 ...
 $ quality.indicator.N363.RInf.U0  : num    0.498 261.887 249.583 195.371   0.727 ...
 $ est.prevalence.N363.R118.U12    : num  5.46 1.94 3.16 1.54 7.61 ...
 $ circle.count.N363.R118.U12      : num  380 369 375 384 367 356 387 370 371 386 ...
 $ circle.radius.N363.R118.U12     : num   3.11 70.93 69.52 72.50  3.73 ...
```

```
 $ circle.nb.clusters.N363.R118.U12: int  17 16 15 15 17 15 18 16 15 18 ...
 $ quality.indicator.N363.R118.U12 : num    0.498 261.887 249.583 268.231   0.727 ...
 $ est.prevalence.N363.RInf.U12    : num  5.46 1.94 3.16 1.54 7.61 ...
 $ circle.count.N363.RInf.U12      : num  380 369 375 384 367 393 387 370 371 386 ...
 $ circle.radius.N363.RInf.U12     : num   3.11 70.93 69.52 72.50  3.73 ...
 $ circle.nb.clusters.N363.RInf.U12: int  17 16 15 15 17 16 18 16 15 18 ...
 $ quality.indicator.N363.RInf.U12 : num    0.498 261.887 249.583 268.231   0.727 ...
```

Dans ces exemples, quatre simulations ont été réalisées simultanément. Dans le premier cas, les données des 466 clusters ont été dupliquées 4 fois (une par estimation) et le tableau de données renvoyé comporte ainsi 1864 lignes.

Dans le second exemple, avec merge.result = TRUE, ce sont les variables *est.prevalence, circle.count, circle.radius* et *circle.nb.clusters* qui ont été dupliquées chacune 4 fois, et le nombre de lignes du tableau de données renvoyés correspond au nombre de clusters, soit 466. Les variables *N.parameter, R.parameter* et *U.parameter* ne sont plus présentes. Leur valeur a directement été intégrée dans le nom des nouvelles variables crées.

8. Interpolation spatiale de la prévalence

Une fois la prévalence de chaque cluster estimée, il est possible d'interpoler spatialement cette dernière pour obtenir une carte des variations spatiales du phénomène étudié. Deux techniques sont disponibles à partir de la fonction krige.prev() : l'interpolation linéaire selon l'inverse de la distance[7] et le krigeage (voir encadré page 26). Dans la suite de ce tutoriel nous parlerons essentiellement de la technique du krigeage ordinaire. Pour les personnes désirant utilisées la pondération inverse de la distance, les surfaces de tendances ou bien le krigeage universel, nous les renvoyons à la documentation des fonctions krige() et idw() du packages gstat puisque prevR a recours à ces deux fonctions pour la réalisation des interpolations spatiales.

Quelque soit la méthode utilisée, le principe de krige.prev() consiste à créer une grille plus ou moins fine composée de petites zones carrées ou pixels, puis à calculer pour chaque pixel la valeur correspondante de la variable interpolée, à partir de valeurs connues, en l'occurrence celles des zones enquêtées.

8.1 Principes généraux du krigeage ordinaire

Comme d'autres techniques d'interpolation, le krigeage calcule la valeur en un point de la variable interpolée à partir des valeurs connues en leur affectant à chacune une pondération. La technique de l'inverse de la distance utilise comme valeur de pondération l'inverse de la distance entre le point à estimée et le point enquêté. Le krigeage, quant à lui, calcule les pondérations à partir du degré de similarité de la variable. Statistiquement, ce degré de similarité correspond à la covariance entre les points enquêtés exprimée en fonction de la distance entre ces derniers.

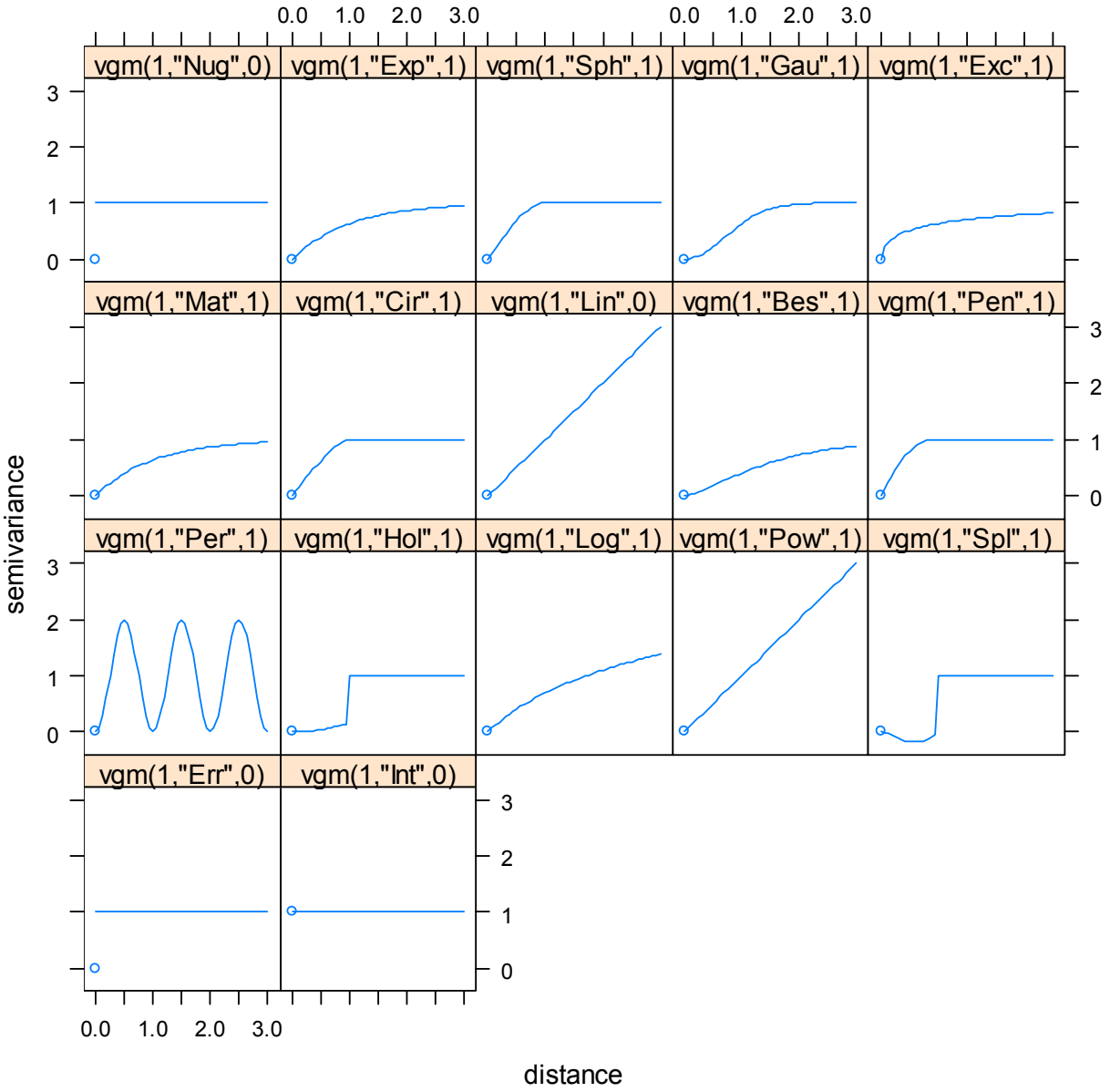
[7] En anglais, cette technique est appelée IDW pour Inverse Distance Weighting. Voir cet article en anglais de l'encyclopédie Wikipeia pour plus de renseignements : http://en.wikipedia.org/wiki/Inverse_distance_weighting.

Plus précisément, le krigeage n'utilise pas la covariance mais la moitié de celle-ci, encore appelée semi-variance. Est appelé semi-variogramme, l'expression de la semi-variance entre les points en fonction de leur distance. Il est représenté usuellement sous la forme d'un graphique avec la distance en abscisse et la semi-variance en ordonnée.

Pour réaliser une interpolation spatiale par krigeage, la première étape consiste à calculer sur l'échantillon des points enquêtés un semi-variogramme expérimental, à savoir les variations de la semi-variance en fonction de la distance entre les points.

Cependant, ce semi-variogramme expérimental n'est pas directement utilisable pour réaliser l'interpolation spatiale. Il faut le modéliser, c'est-à-dire trouver une fonction mathématique le décrivant au mieux. C'est cette fonction, ou modèle de semi-variogramme, qui sera utilisée pour le calcul effectif de l'interpolation.

Plusieurs types de courbes mathématiques existent pour modéliser le semi-variogramme. La fonction show.vgms() permet de visualiser celles qui sont prises en compte par le package gstat.



À l'usage, les modèles de semi-variogrammes de type *exponentiel* permettent d'obtenir des cartes lisibles et de qualité. La suite du tutoriel n'abordera que ce type de modèle de semi-variogrammes. Cependant, un utilisateur habitué à manier ce type d'outil géostatistique pourra avoir recours à d'autres types de semi-variogrammes.

L'ajustement d'un modèle de semi-variogramme aux données expérimentales sera abordé dans les parties suivantes au travers d'un exemple concret.

8.2 Les différents paramètres de krige.prev

Le tableau de données passé en paramètre de la fonction krige.prev() doit avoir été obtenu en spécifiant merge.result=TRUE à la fonction estimate.prev() ou bien en ayant eu recours à extract.data() de manière à ce qu'à chaque ligne du tableau de données correspondent à un cluster et à un seul. Si data comporte deux lignes ayant les mêmes coordonnées géographiques, alors vous obtiendrez une erreur dans krige.prev().

formula permet de spécifier la variable à interpoler sous la forme d'une formule. Pour interpoler, par exemple, la prévalence estimée pour *N=363*, *R=118* et *U=12*, nous appellerons krige.prev() avec formula = est.prevalence.N363.R118.U12 ~ 1. La formule employée est de la forme *variable ~ 1*. La partie *~1* indique qu'il s'agit d'un krigeage ordinaire. Des interpolations plus complexes peuvent être réalisés en spécifiant d'autres variables à la droite du ~. Pour cela nous vous renvoyons à la documentation de la fonction krige() du package gstat. Il est possible de réaliser plusieurs interpolations simultanément. Il suffit de passer une liste de formules au paramètre formula. Par exemple, pour interpoler à la fois la prévalence estimée et l'indicateur de qualité lorsque *N=363*, *R=118* et *U=12*, on entrera :

formula = c(est.prevalence.N363.R118.U12 ~ 1, quality.indicator.N363.R118.U12 ~ 1).

Le paramètre cell.size permet de définir la taille des pixels de la grille sur laquelle l'interpolation sera réalisée. Plus cell.size est petit, plus le nombre de pixels sera élevés, plus le temps de calcul sera long et plus la carte obtenue sera précise. Si ask.cell.size = TRUE (valeur par défaut), krige.prev calculera la taille de la grille obtenue avec la valeur de cell.size entrée et vous proposera de modifier cette valeur si besoin. L'unité de dimension de cell.size correspond à celle des coordonnées des clusters du tableau de données fourni en entrée.

Le paramètre type peut prendre plusieurs valeurs. Il détermine le type d'interpolation et la manière, en cas de krigeage, dont sera déterminé le modèle de semi-variogramme utilisé.

- Si type = 'idw', alors krige.prev réalisera une interpolation par inverse de la distance à l'aide la fonction idw() du package gstat. La puissance appliquée à l'inverse de la distance peut être précisée avec le paramètre idp.

- Si type = 'model', le modèle de semi-variogramme utilisé pour l'interpolation par krigeage sera celui passé au paramètre model. Si plusieurs interpolations sont réalisées simultanément, il est possible des spécifier plusieurs modèles à model. Voir la documentation de krige.prev() pour plus de détails.
- Si type = 'auto', krige.prev aura recours à la fonction fit.variogram() pour ajuster par la méthode des moindres carrés un modèle de semi-variogramme au semi-variogramme expérimental. C'est ce modèle qui sera ensuite utilisé pour l'interpolation spatiale. L'utilisateur doit rester vigilant dans la mesure où l'ajustement par la méthode des moindres carrés ne produit pas toujours un modèle de semi-variogramme adéquat.
- Par défaut, type = 'ask'. Dans le doute, utilisez de préférence ce mode de fonctionnement. Dans le cas présent, krige.prev() a toujours recours à fit.variogram() pour ajuster un modèle de semi-variogramme aux données expérimentales. Une fois celui-ci calculé, krige.prev() affiche le semi-variogramme expérimental ainsi que le modèle ajusté. L'utilisateur est alors invité à accepter le modèle ajusté par la méthode des moindres carrés ou à procéder à un ajustement manuel (voir l'exemple concret ci-après).

Le calcul d'interpolations spatiales sur des grilles fines est une opération longue et peut prendre plusieurs minutes. Soyez donc patient.

8.3 Exemple d'interpolation spatiale

Dans cet exemple, nous interpolerons la prévalence estimée au Cameroun avec les paramètres *N=363, R=118* et *U=12*. Le mode *ask* sera utilisé.

```
> cm.krige <- krige.prev(cm.prev, formula = est.prevalence.N363.R118.U12 ~ 1,
      boundary = cm.bounds, lang='fr')
```

La fonction commence par afficher la taille de la grille résultant de pixels carrés de 0,05 degrés de côté (valeur par défaut). Si l'on souhaite utiliser une grille plus fine (plus grand nombre de pixels), il faut réduire la longueur des côtés des pixels. (Une plus petite valeur de cell.size induit un nombre plus important de pixels sur une même zone géographique.)

```
Une taille de cellule de 0.05 induit une grille de 154x229 cellules.
Cela vous convient-il ?

1: Oui
2: Non

Sélection : 2

Entrez une nouvelle valeur de taille de cellule :
```

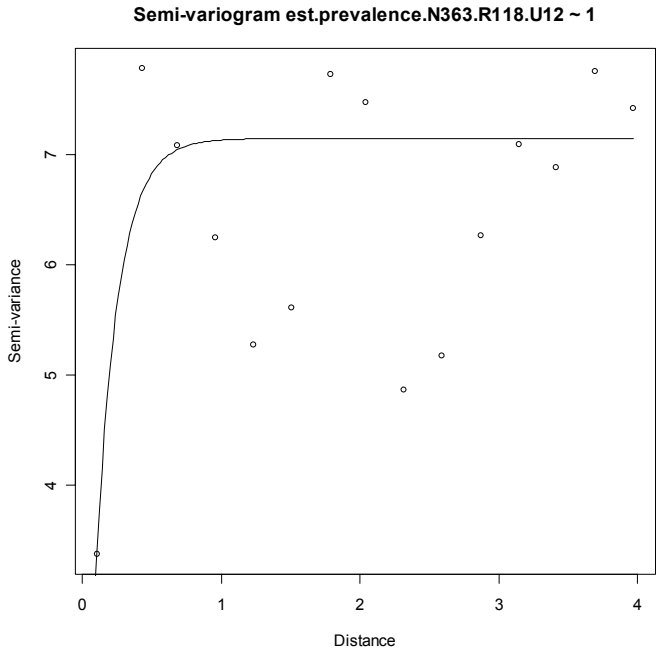
La nouvelle valeur est à saisir dans une fenêtre. Ici, nous avons choisi 0,025 degrés.

```
Une taille de cellule de 0.025 induit une grille de 308x458 cellules.
Cela vous convient-il ?

1: Oui
2: Non

Sélection : 1
```

Une fois la taille de la grille définie, la fonction calcule le semi-variogramme expérimental, ajuste un modèle de semi-variogramme aux données empiriques puis affiche les détails du modèle ajusté :

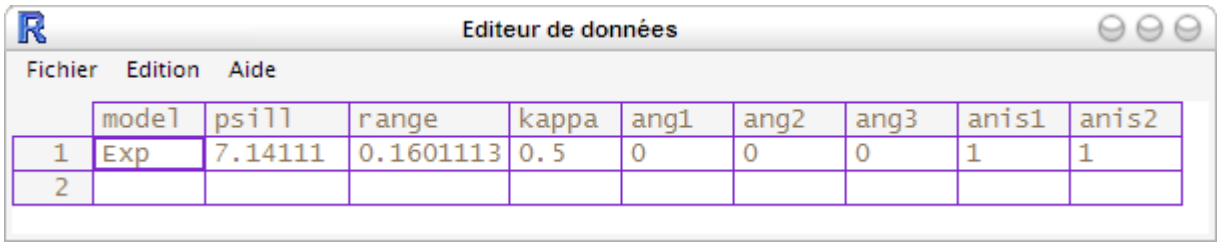
```
--- est.prevalence.N363.R118.U12 ~ 1 ---

  model   psill     range
1   Exp 7.14111 0.1601113
```

Le modèle en question est affiché dans une fenêtre graphique sous la forme d'une courbe. Les points représentent le semi-variogramme expérimental.



Le semi-variogramme expérimental étant plus ou moins régulier, l'ajustement n'a pas été optimal : le modèle ajusté croît trop rapidement vers une valeur plafond. Nous allons donc procéder à un ajustement manuel :

```
Ce modèle de variogramme convient-il ?

1: Oui
2: Non

Sélection : 2
Entrez un nouveau modèle.
```

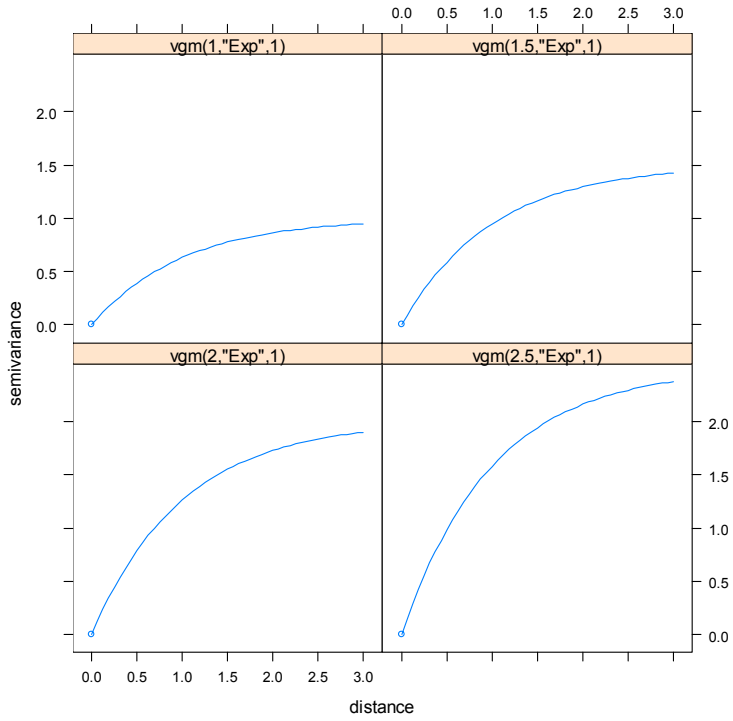


	model	psill	range	kappa	ang1	ang2	ang3	anis1	anis2
1	Exp	7.14111	0.1601113	0.5	0	0	0	1	1
2									

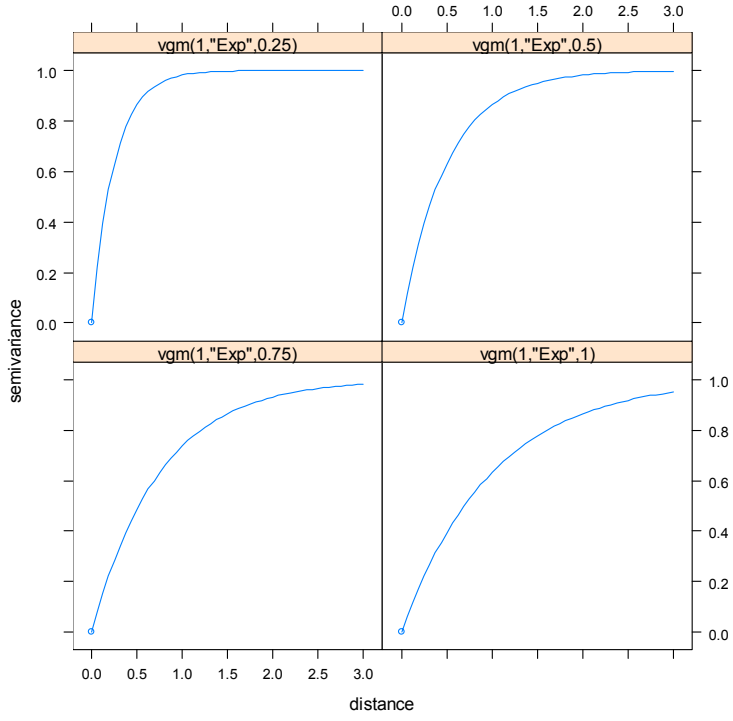
Une fenêtre s'ouvre permettant de modifier les paramètres du modèle. Pour plus de détails sur ces paramètres, consultez la documentation de la fonction vgm() du package gstat.

Pour un usage courant, seuls les paramètres *psill* et *range* seront à modifier. *psill* correspond à la valeur seuil vers laquelle le modèle tend et *range* à la rapidité avec laquelle le modèle s'approche de cette valeur seuil.

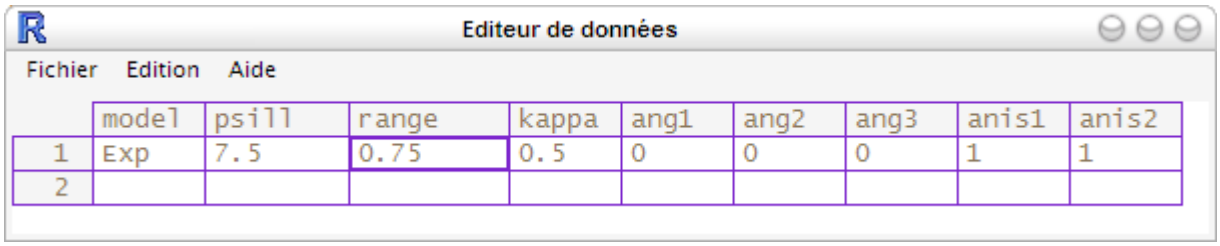
Ci-dessous, voici plusieurs modèles avec des valeurs de *psill* de 1, 1,5, 2 et 2,5. Plus la valeur de *psill* augmente, plus la courbe atteint une semi-variance élevée.



range influe sur la rapidité à laquelle le modèle s'approche de sa valeur seuil. Plus *range* est élevé, plus la croissance sera progressive. Ci-dessous, plusieurs modèles correspondant à des valeurs de *range* égales à 0,25, 0,5, 0,75 et 1.

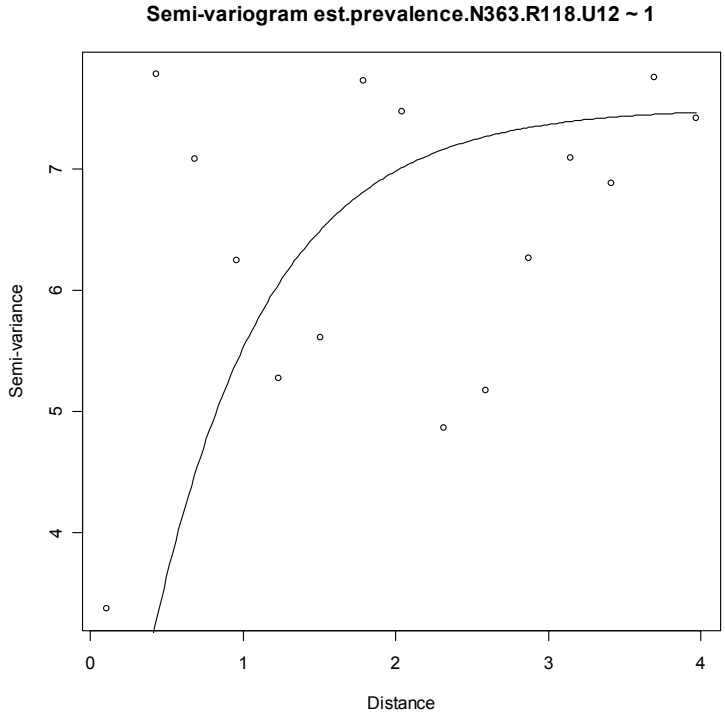


Dans le cas présent, nous allons modifier le modèle de manière à ce qu'il ait une croissance moins rapide (pour cela on augmente la valeur de *range*) et que sa valeur seuil soit un peu plus élevée (augmentation légère de *psill*). Nous allons donc entrer 7.5 pour *psill* et 0.75 pour *range*.



	model	psill	range	kappa	ang1	ang2	ang3	anis1	anis2
1	Exp	7.5	0.75	0.5	0	0	0	1	1
2									

Le graphique est alors modifié pour afficher la courbe du nouveau modèle.




```
--- est.prevalence.N363.R118.U12 ~ 1 ---

  model psill range
1   Exp   7.5  0.75

Ce modèle de variogramme convient-il ?

1: Oui
2: Non
```

Si l'on désire affiner encore l'ajustement, il est possible de modifier à nouveau les valeurs des paramètres du modèle. On peut alors procéder à plusieurs essais et avancer par tâtonnement. Le calcul de l'interpolation spatiale se lance une fois le modèle accepté.

```
Sélection : 1

Modèle de variogramme utilisé :
  model psill range
1   Exp   7.5  0.75
[using ordinary kriging]
```

Pour effectuer plusieurs interpolations en même temps, il suffit de spécifier une liste de formules au paramètre formula à l'aide de la fonction c(). L'ensemble des résultats sera alors regroupé en un seul fichier, permettant de réaliser aisément des cartes comparatives. D'autres indicateurs peuvent être interpolés, tels que l'indicateur de qualité ou le rayon des cercles de lissage. Dans l'exemple ci-dessous, la prévalence estimée est interpolée selon trois cas de figure (utilisation de la méthode des cercles avec seul le paramètre *N*, avec les paramètres *N* et *R* et avec les trois paramètres *N*, *R* et *U*) ainsi que l'indicateur de qualité et le rayon des cercles de lissage de l'estimation avec *N*, *R* et *U*.

```
> cm.krige <- krige.prev(
      cm.prev,
      formula = c(
            est.prevalence.N363.RInf.U0 ~1,
            est.prevalence.N363.R118.U0 ~ 1,
            est.prevalence.N363.R118.U12 ~ 1,
            quality.indicator.N363.R118.U12 ~1,
            circle.radius.N363.R118.U12 ~1
      ),
      boundary = cm.bounds,
)
```

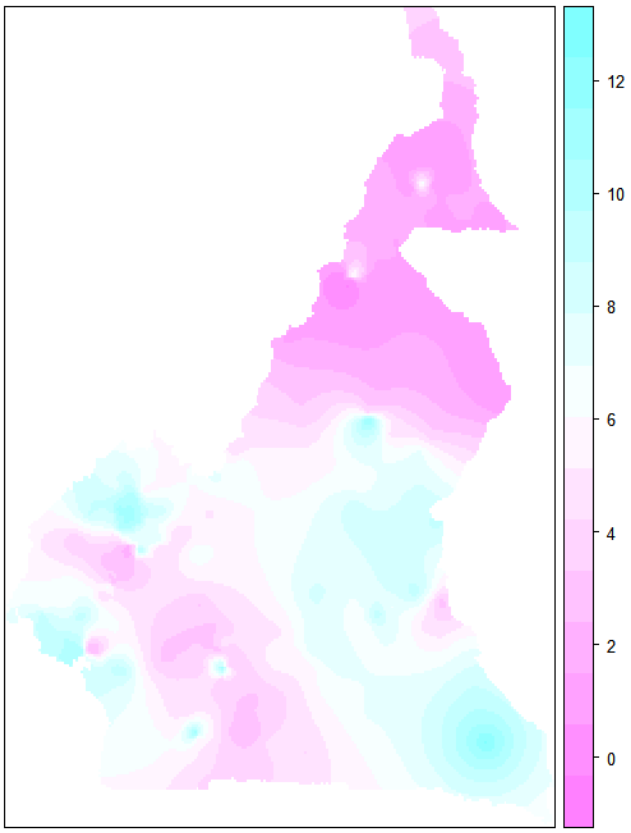
krige.prev() renvoie un objet de type *SpatialPixelsDataFrame*. Deux variables sont créées pour chacune des variables interpolées : l'une avec le suffixe *.pred* comportant pour chaque point de la grille la valeur de la prédiction ; l'autre avec le suffixe *.var* avec pour chaque point de la grille la variance de la prédiction (en cas de krigeage uniquement).

```
> str(cm.krige)
Formal class 'SpatialPixelsDataFrame' [package "sp"] with 7 slots
  ..@ data       :'data.frame': 71940 obs. of  12 variables:
  .. ..$ est.prevalence.N363.RInf.U0.pred    : num [1:71940] NA NA NA NA NA NA NA NA NA NA ...
  .. ..$ est.prevalence.N363.RInf.U0.var     : num [1:71940] NA NA NA NA NA NA NA NA NA NA ...
  .. ..$ est.prevalence.N363.R118.U0.pred    : num [1:71940] NA NA NA NA NA NA NA NA NA NA ...
  .. ..$ est.prevalence.N363.R118.U0.var     : num [1:71940] NA NA NA NA NA NA NA NA NA NA ...
  .. ..$ est.prevalence.N363.R118.U12.pred   : num [1:71940] NA NA NA NA NA NA NA NA NA NA ...
  .. ..$ est.prevalence.N363.R118.U12.var    : num [1:71940] NA NA NA NA NA NA NA NA NA NA ...
  .. ..$ quality.indicator.N363.R118.U12.pred: num [1:71940] NA NA NA NA NA NA NA NA NA NA ...
  .. ..$ quality.indicator.N363.R118.U12.var : num [1:71940] NA NA NA NA NA NA NA NA NA NA ...
  .. ..$ circle.radius.N363.R118.U12.pred    : num [1:71940] NA NA NA NA NA NA NA NA NA NA ...
  .. ..$ circle.radius.N363.R118.U12.var     : num [1:71940] NA NA NA NA NA NA NA NA NA NA ...
  ..@ coords.nrs : num(0)
  ..@ grid       :Formal class 'GridTopology' [package "sp"] with 3 slots
  .. .. ..@ cellcentre.offset: Named num [1:2] 8.49 1.65
  .. .. .. ..- attr(*, "names")= chr [1:2] "x" "y"
  .. .. ..@ cellsize         : Named num [1:2] 0.035 0.035
  .. .. .. ..- attr(*, "names")= chr [1:2] "x" "y"
  .. .. ..@ cells.dim        : Named int [1:2] 220 327
  .. .. .. ..- attr(*, "names")= chr [1:2] "x" "y"
  ..@ grid.index : int [1:71940] 1 2 3 4 5 6 7 8 9 10 ...
  ..@ coords     : num [1:71940, 1:2] 8.49 8.53 8.56 8.60 8.63 ...
  .. ..- attr(*, "dimnames")=List of 2
  .. .. ..$ : NULL
  .. .. ..$ : chr [1:2] "x" "y"
  ..@ bbox       : num [1:2, 1:2]  8.48  1.64 16.18 13.08
  .. ..- attr(*, "dimnames")=List of 2
  .. .. ..$ : chr [1:2] "x" "y"
  .. .. ..$ : chr [1:2] "min" "max"
  ..@ proj4string:Formal class 'CRS' [package "sp"] with 1 slots
  .. .. ..@ projargs: chr NA
```

9. Cartographier les résultats

Les objets du type *SpatialPixelsDataFrame* peuvent être facilement représentés à l'aide de la fonction spplot() du package sp. Le premier paramètre spécifie l'ensemble de données, le second la variable représenter (ou plusieurs variables s'il s'agit d'une liste). Si le second paramètre est omis, spplot() représentera avec la même échelle colorimétrique l'ensemble des variables contenues dans l'ensemble de données.

```
> spplot(cm.krige,'est.prevalence.N363.R118.U12.pred')
```

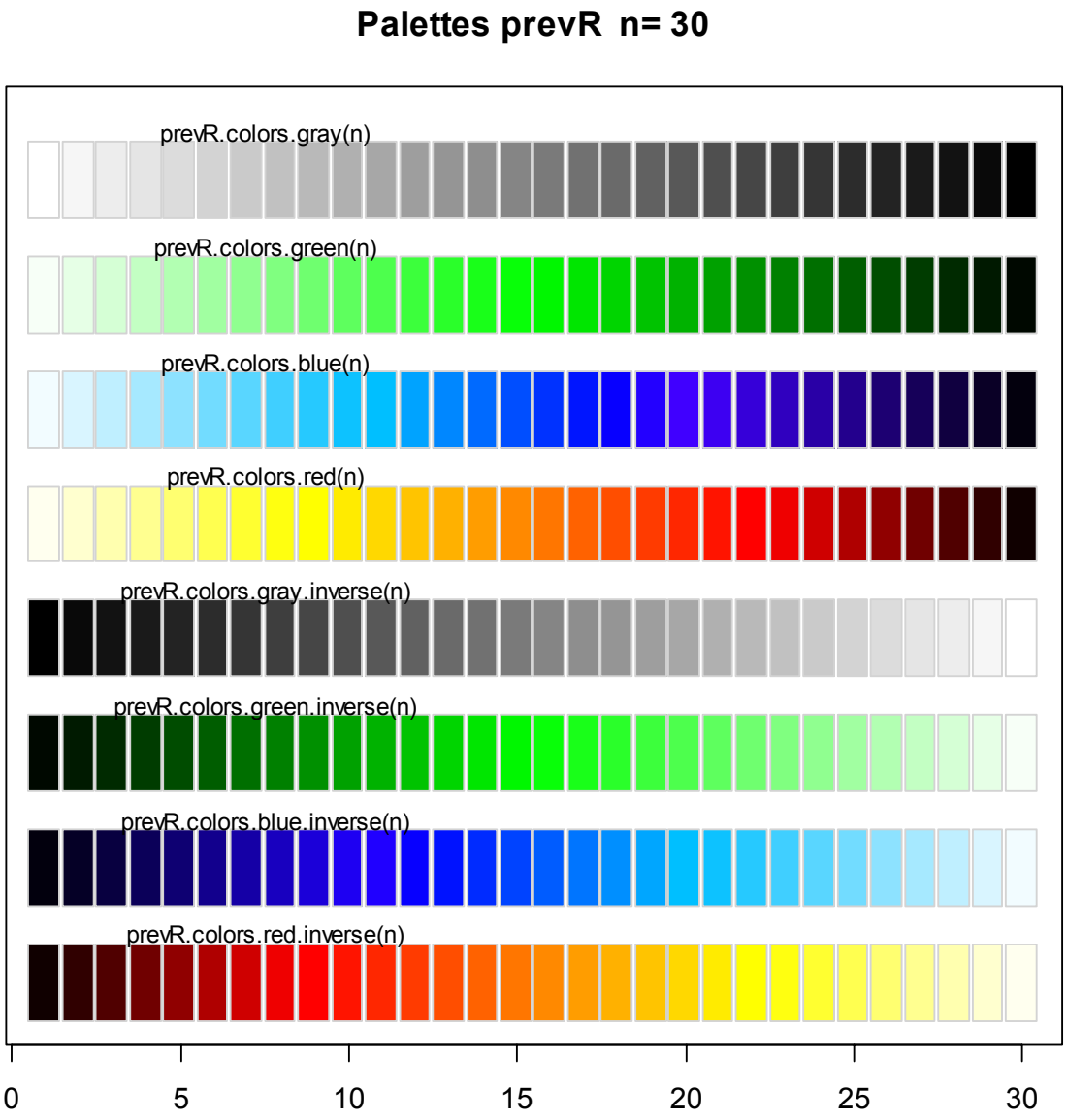


Il est possible de préciser un titre avec main et un sous-titre avec sub. Le paramètre cuts précise le nombre de plages de couleurs.

La fonction extract.col() peut être utile pour extraire certaines variables d'un ensemble de données. Voir la documentation de cette fonction.

La palette de couleurs à utiliser peut-être spécifiée avec col.regions. prevR fournit plusieurs palettes de couleurs, visible avec la fonction prevR.demo.pal().

```
> prevR.demo.pal(30)
```



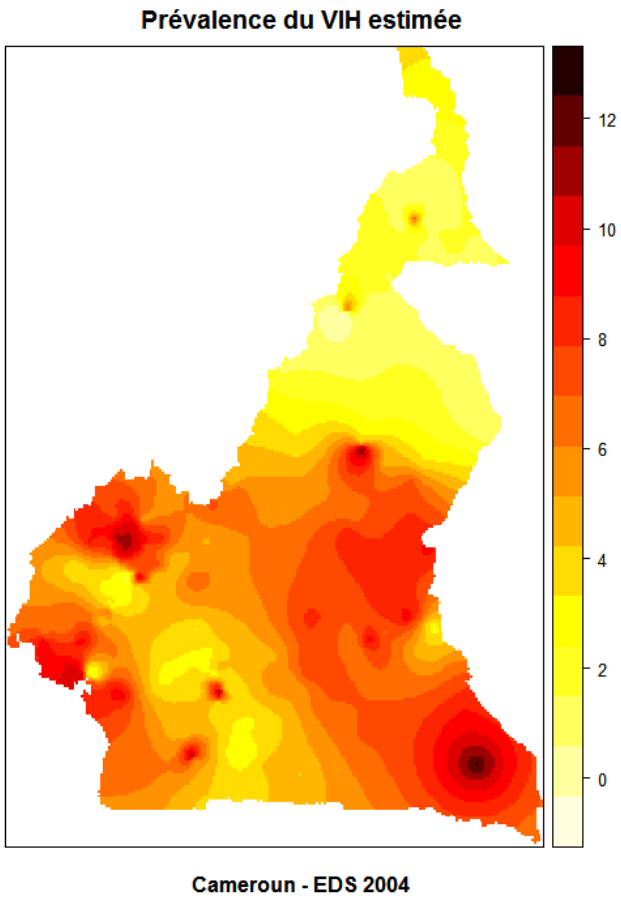
Ces palettes ont été conçues pour accentuer les contrastes en éclaircissant ou obscurcissant les valeurs extrêmes. D'autres palettes peuvent être utilisées. Consultez l'aide de la fonction rainbow() ou le package RColorBrewer. En appelant une fonction palette de couleurs avec spplot(), pensez à générer au moins une couleur de plus que de plages de niveaux.

Le nombre de plages de niveaux modifie le rendu global de la carte générée. Pour un rendu lisse, utilisez un nombre élevé de plages. Le nombre de plages est défini par le paramètre cuts.

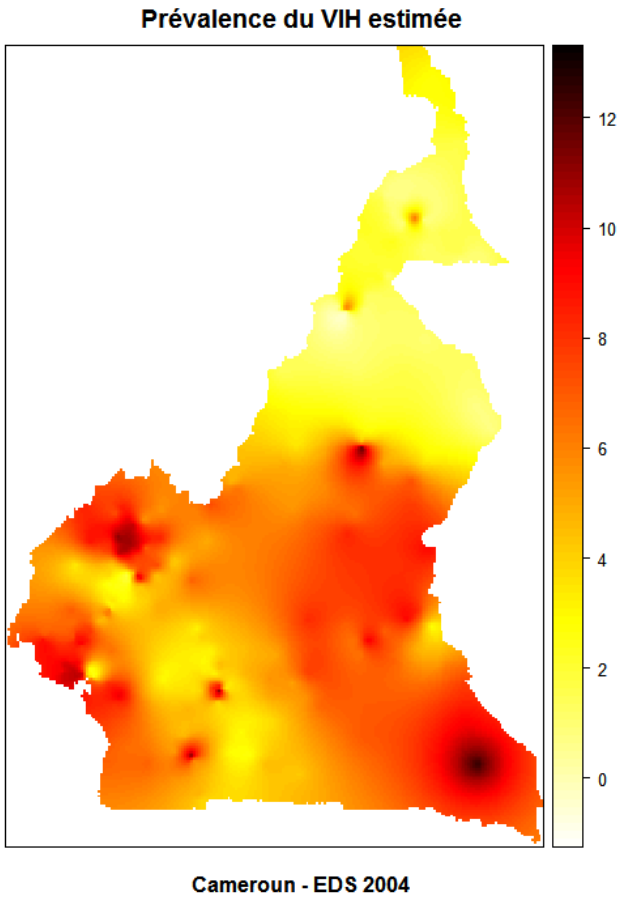
Il est possible de représenter plusieurs variables simultanément, avec la même échelle. Pour cela, il faut passer une liste de noms de variables à spplot() via le second paramètre appelé également zcol.

Inspirez-vous des exemples suivants :

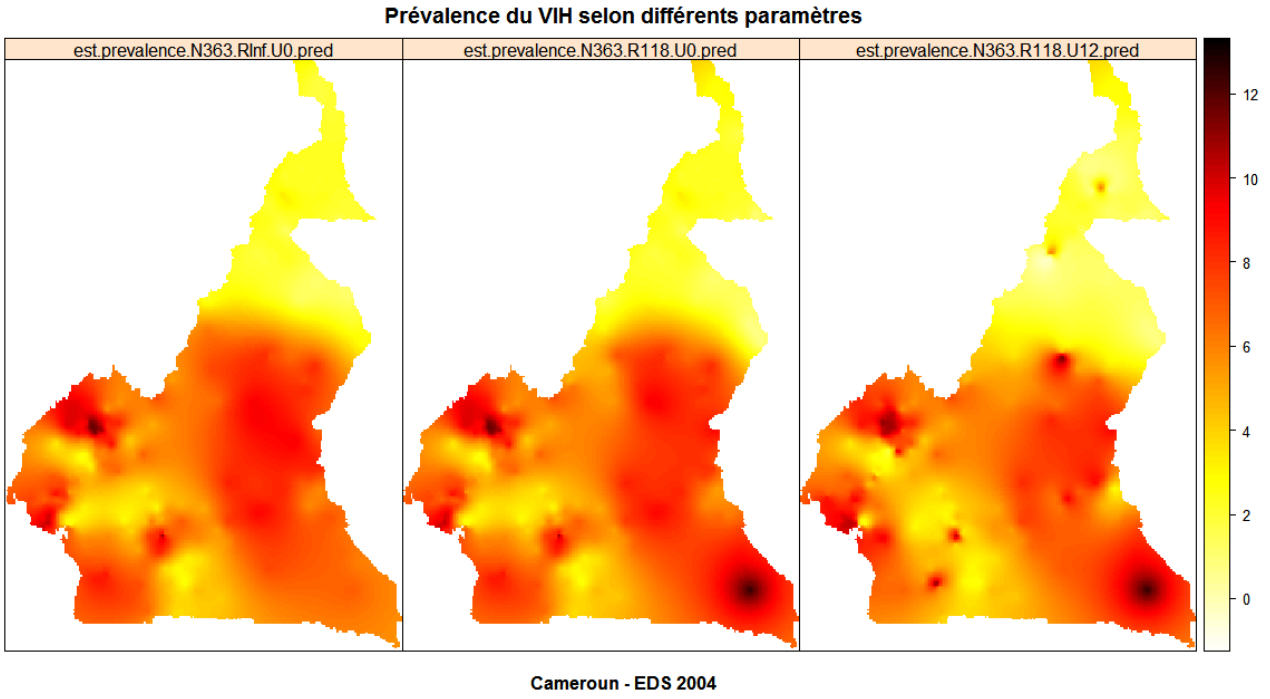
```
> spplot(
      cm.krige,
      'est.prevalence.N363.R118.U12.pred',
      cuts=15,
      col.regions=prevR.colors.red(16),
      main='Prévalence du VIH estimée',
      sub='Cameroun - EDS 2004'
)
```



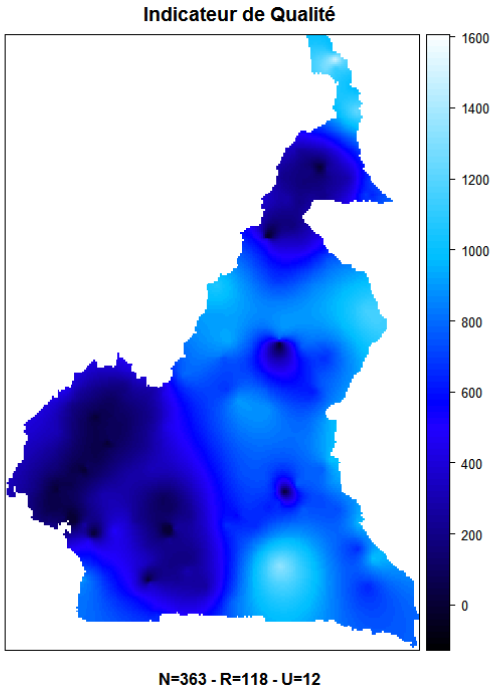

```
> spplot(
      cm.krige,
      'est.prevalence.N363.R118.U12.pred',
      cuts=100,
      col.regions=prevR.colors.red(101),
      main='Prévalence du VIH estimée',
      sub='Cameroun - EDS 2004'
)
```



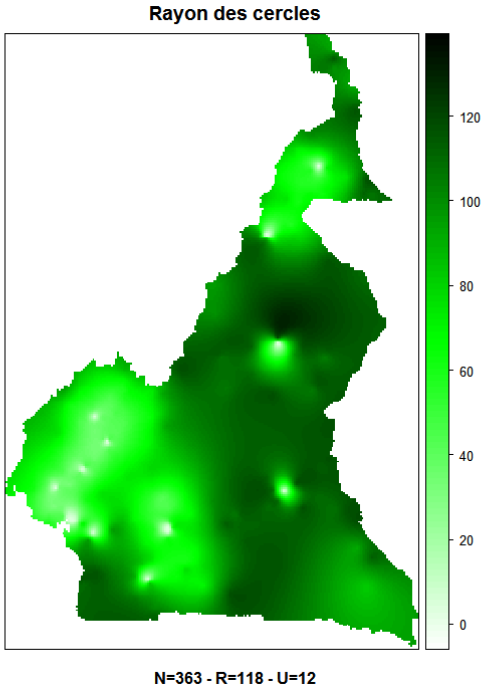
```
> spplot(
      cm.krige,
      c(
            'est.prevalence.N363.RInf.U0.pred',
            'est.prevalence.N363.R118.U0.pred',
            'est.prevalence.N363.R118.U12.pred'
      ),
      cuts=100,
      col.regions=prevR.colors.red(101),
      main='Prévalence du VIH selon différents paramètres',
      sub='Cameroun - EDS 2004'
)
```




```
> spplot(
      cm.krige,
      'quality.indicator.N363.R118.U12.pred',
      cuts=100,
      col.regions=prevR.colors.blue.inverse(101),
      main='Indicateur de Qualité',
      sub='N=363 - R=118 - U=12'
)
```



```
> spplot(
      cm.krige,
      'circle.radius.N363.R118.U12.pred',
      cuts=100,
      col.regions=prevR.colors.green(101),
      main='Rayon des cercles',
      sub='N=363 - R=118 - U=12'
)
```



La fonction spplot() crée par défaut une légende dont le minimum et le maximum diffèrent sensiblement du minimum et du maximum réel de la variable à représenter. Il est possible de connaître facilement le minimum et le maximum de chaque variable à l'aide la fonction summary().

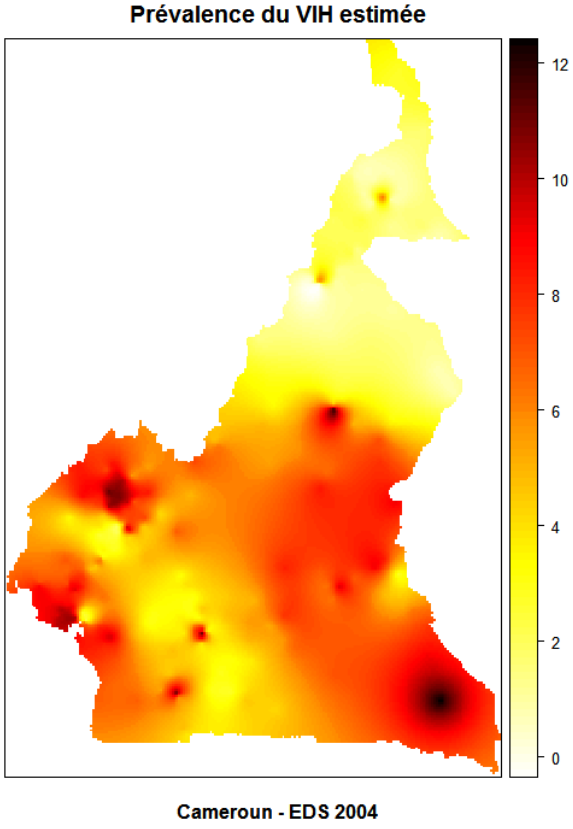
```
> summary(cm.krige)
Object of class SpatialPixelsDataFrame
Coordinates:
       min      max
x 8.477263 16.17726
y 1.635048 13.08005
Is projected: NA
proj4string : [NA]
Number of points: 71940

Data attributes:
 obs.prevalence.pred obs.prevalence.var  est.prevalence.N363.RInf.U0.pred
 Min.   :   -3.246   Min.   :3.958e-02   Min.   :    1.066
 1st Qu.:    1.617   1st Qu.:8.995e+00   1st Qu.:    4.237
 Median :    4.240   Median :1.491e+01   Median :    6.161
 Mean   :    4.897   Mean   :1.712e+01   Mean   :    5.753
 3rd Qu.:    7.152   3rd Qu.:2.361e+01   3rd Qu.:    7.301
 Max.   :   42.710   Max.   :5.892e+01   Max.   :   11.715
 NA's   :40816.000   NA's   :4.082e+04   NA's   :40816.000

 est.prevalence.N363.RInf.U0.var est.prevalence.N363.R118.U0.pred
 Min.   :4.288e-03               Min.   :7.145e-01
 1st Qu.:9.745e-01               1st Qu.:4.111e+00
 Median :1.615e+00               Median :6.069e+00
 Mean   :1.855e+00               Mean   :5.807e+00
 3rd Qu.:2.558e+00               3rd Qu.:7.666e+00
 Max.   :6.383e+00               Max.   :1.242e+01
 NA's   :4.082e+04               NA's   :4.082e+04

 est.prevalence.N363.R118.U0.var est.prevalence.N363.R118.U12.pred
 Min.   :5.033e-03               Min.   :   -0.3528
 1st Qu.:1.138e+00               1st Qu.:    3.4219
 Median :1.870e+00               Median :    5.6103
 Mean   :2.094e+00               Mean   :    5.2425
 3rd Qu.:2.907e+00               3rd Qu.:    7.0649
 Max.   :6.161e+00               Max.   :   12.4176
 NA's   :4.082e+04               NA's   :40816.0000
```

```
 est.prevalence.N363.R118.U12.var quality.indicator.N363.R118.U12.pred
 Min.   :5.888e-03                Min.   :  -20.76
 1st Qu.:1.331e+00                1st Qu.:  304.77
 Median :2.188e+00                Median :  666.37
 Mean   :2.449e+00                Mean   :  593.69
 3rd Qu.:3.400e+00                3rd Qu.:  818.85
 Max.   :7.207e+00                Max.   : 1498.96
 NA's   :4.082e+04                NA's   :40816.00

 quality.indicator.N363.R118.U12.var circle.radius.N363.R118.U12.pred
 Min.   :    46.17                   Min.   :    3.006
 1st Qu.: 10580.41                   1st Qu.:   75.258
 Median : 17796.29                   Median :  101.290
 Mean   : 21752.55                   Mean   :   92.755
 3rd Qu.: 29077.42                   3rd Qu.:  112.409
 Max.   :123239.31                   Max.   :  130.503
 NA's   : 40816.00                   NA's   :40816.000

 circle.radius.N363.R118.U12.var
 Min.   :7.916e-01
 1st Qu.:1.814e+02
 Median :3.051e+02
 Mean   :3.729e+02
 3rd Qu.:4.985e+02
 Max.   :2.113e+03
 NA's   :4.082e+04
```

Concernant la prévalence estimée pour *N=363, R=118* et *U=12*, le minimum et le maximum de l'interpolation spatiale correspondent respectivement à -0,3528 et 12,4176 %. La valeur négative du minimum peut surprendre. Elle correspond à des limites de prédiction en des points atypiques et ne doit donc pas être considérée comme une valeur précise localement, valeur qui serait aberrante. Cela doit s'interpréter comme un point de très faible prévalence proche de zéro[8].

On peut indiquer à spplot() les minimum et maximum à prendre en compte pour la représentation graphique à l'aide du paramètre at qui permet d'indiquer les limites des différents niveaux.

```
> spplot(
      cm.krige,
      'est.prevalence.N363.R118.U12.pred',
      cuts=100,
      col.regions=prevR.colors.red(101),
      main='Prévalence du VIH estimée',
      sub='Cameroun - EDS 2004',
      at = seq(-0.3529, 12.4177, length.out=100)
)
```



[8] Dans certains cas, on pourra considérer que les points présentant une prévalence estimée négative doivent être ramenés à la valeur 0. Pour cela, il suffit d'entrer la commande suivante :
`cm.krige$est.prevalence.N363.R118.U12.pred[cm.krige$est.prevalence.N363.R118.U12.pred<0] <- 0`

10. Exporter les résultats

10.1 Export vers un logiciel de statistiques

La majorité des logiciels de statistiques ainsi que les principaux tableurs tels que Excel ou OpenOffice Calc sont capables de lire des fichiers au format texte ou au format *dbf*.

Il est possible d'exporter facilement des tableaux de données (*data.frame*) ainsi que les résultats de krige.prev() à l'aide des fonctions write.table(), write.dbf() et write.txt().

- write.dbf() permet, comme son nom l'indique d'exporter au format *dbf*.
- write.table() est la fonction générique de R pour exporter au format texte, que soit tabulé ou de type *csv*. Elle dispose de nombreuses options.
- write.txt() permet d'appeler write.table() avec les options adéquates pour un export au format texte tabulé. Autrement dit, write.txt() génère un fichier texte dont les colonnes sont séparées par une tabulation, sans ajout des noms de ligne générés automatiquement par R, sans guillemets encadrant pour les valeurs textes. Seul le séparateur de décimales peut-être modifié. Par défaut, il s'agit du point. Pour certains logiciels, notamment Excel dans sa version française, il est nécessaire d'utiliser la virgule comme séparateur de décimal.

Quelques exemples :

```
> write.dbf(cm.prev,'cm_prev.dbf')
> write.dbf(cm.bounds, 'cm_bounds.dbf')

> write.txt(cm.prev, 'cm_prev.txt')
> write.txt(cm.prev, 'cm_prev_fr.txt', dec=',')
> write.csv(cm.prev, 'cm_prev.csv')
```

Les fichiers *dbf* ont une limitation. En effet, les noms de variables pour ce type de fichiers ne peuvent excéder dix caractères. Ainsi, au moment de l'export avec write.dbf(), les noms de colonnes trop longs du tableau de données seront tronqués. Or, si deux colonnes ont deux noms différents mais ont les mêmes dix premiers caractères, elles porteront le même nom dans le fichier exporté. Cela est fréquent dans un ensemble de données comme cm.prev où *est.prevalence.N338.R118.U12* et *est.prevalence.N338.RInf.Uo* seront tous deux renommés en *est_preval*. Afin de pouvoir différencier ces deux variables dans le fichier exporté, il est préférable de renommer manuellement les variables du tableau de données. Cela se fait nativement sous R à partir de la fonction names() (voir la documentation de cette fonction).

Pour rendre cette opération plus facile, il suffit d'utiliser la fonction check.names() fournie par prevR. Cette fonction vérifie la longueur des noms de colonne. Si un nom est trop long, une fenêtre s'ouvre permettant de modifier les noms de chaque colonne. La longueur des nouveaux noms est elle aussi vérifiée. Pendant cette opération, il est possible de supprimer certaines colonnes en les renommant *NULL*.

```
> str(cm.prev)
'data.frame':   466 obs. of  34 variables:
 $ cluster                          : int  1 10 100 101 102 103 104 105 106 107 ...
 $ x                                : num   9.72 13.57 11.23 14.71 11.55 ...
 $ y                                : num   4.04 10.25  4.74 10.43  3.88 ...
 $ residence                        : Factor w/ 2 levels "Rural","Urban": 2 1 2 1 2 2 2 2 1 2 ...
 $ region                           : num  3 5 2 5 12 7 8 12 7 3 ...
 $ region.name                      : Factor w/ 12 levels "ADAMAOUA","CENTRE",..: 3 5 2 5 12 7 8 ...
 $ longitude                        : num  NA NA NA NA NA NA NA NA NA NA ...
 $ latitude                         : num  NA NA NA NA NA NA NA NA NA NA ...
 $ n                                : num  9 22 18 8 17 36 57 16 16 10 ...
 $ nweight                          : num  10.8 34.8 28.7 12.7 22.1 ...
 $ obs.prevalence                   : num   0.00 13.63  0.00  0.00  5.81 ...
 $ dist.city                        : num    1.87  91.90 102.20  45.20   3.59 ...
 $ city.name                        : chr  "DOUALA" "MAROUA" "YAOUNDE" "MAROUA" ...
 $ urban.area                       : Factor w/ 2 levels "in urban area",..: 1 2 2 2 1 2 2 1 2 1 ...
 $ est.prevalence.N363.R118.U0      : num  5.46 1.94 3.16 2.49 7.61 ...
 $ circle.count.N363.R118.U0        : num  380 369 375 379 367 376 387 370 380 386 ...
 $ circle.radius.N363.R118.U0       : num   3.11 70.93 69.52 61.67  3.73 ...
 $ circle.nb.clusters.N363.R118.U0  : int  17 16 15 15 17 18 18 16 18 18 ...
 $ quality.indicator.N363.R118.U0   : num    0.498 261.887 249.583 195.371   0.727 ...
 $ est.prevalence.N363.RInf.U0      : num  5.46 1.94 3.16 2.49 7.61 ...
 $ circle.count.N363.RInf.U0        : num  380 369 375 379 367 376 387 370 380 386 ...
 $ circle.radius.N363.RInf.U0       : num   3.11 70.93 69.52 61.67  3.73 ...
 $ circle.nb.clusters.N363.RInf.U0  : int  17 16 15 15 17 18 18 16 18 18 ...
 $ quality.indicator.N363.RInf.U0   : num    0.498 261.887 249.583 195.371   0.727 ...
 $ est.prevalence.N363.R118.U12     : num  5.46 1.94 3.16 1.54 7.61 ...
 $ circle.count.N363.R118.U12       : num  380 369 375 384 367 356 387 370 371 386 ...
 $ circle.radius.N363.R118.U12      : num   3.11 70.93 69.52 72.50  3.73 ...
 $ circle.nb.clusters.N363.R118.U12: int  17 16 15 15 17 15 18 16 15 18 ...
 $ quality.indicator.N363.R118.U12 : num    0.498 261.887 249.583 268.231   0.727 ...
 $ est.prevalence.N363.RInf.U12     : num  5.46 1.94 3.16 1.54 7.61 ...
 $ circle.count.N363.RInf.U12       : num  380 369 375 384 367 393 387 370 371 386 ...
 $ circle.radius.N363.RInf.U12      : num   3.11 70.93 69.52 72.50  3.73 ...
 $ circle.nb.clusters.N363.RInf.U12: int  17 16 15 15 17 16 18 16 15 18 ...
 $ quality.indicator.N363.RInf.U12 : num    0.498 261.887 249.583 268.231   0.727 ...
```

```
> cm.prev.check <- check.names(cm.prev, lang='fr')
Certaines variables ont un nom dépassant 10 caractères. Veuillez entrer de nouveaux noms.
Pour supprimer une variable, entrez NULL.
```

avant saisie

Editeur de données

Fichier Edition Aide

	variable	size
1	cluster	7
2	x	1
3	y	1
4	residence	9
5	region	6
6	region.name	11
7	longitude	9
8	latitude	8
9	n	1
10	nweight	7
11	obs.prevalence	14
12	dist.city	9
13	city.name	9
14	urban.area	10
15	est.prevalence.N363.R118.U0	27
16	circle.count.N363.R118.U0	25
17	circle.radius.N363.R118.U0	26
18	circle.nb.clusters.N363.R118.U0	31
19	quality.indicator.N363.R118.U0	30
20	est.prevalence.N363.RInf.U0	27
21	circle.count.N363.RInf.U0	25
22	circle.radius.N363.RInf.U0	26
23	circle.nb.clusters.N363.RInf.U0	31
24	quality.indicator.N363.RInf.U0	30
25	est.prevalence.N363.R118.U12	28
26	circle.count.N363.R118.U12	26
27	circle.radius.N363.R118.U12	27
28	circle.nb.clusters.N363.R118.U12	32
29	quality.indicator.N363.R118.U12	31
30	est.prevalence.N363.RInf.U12	28
31	circle.count.N363.RInf.U12	26
32	circle.radius.N363.RInf.U12	27
33	circle.nb.clusters.N363.RInf.U12	32
34	quality.indicator.N363.RInf.U12	31

après saisie

Editeur de données

Fichier Edition Aide

	variable	size
1	cluster	7
2	x	1
3	y	1
4	residence	9
5	region	6
6	reg.name	11
7	NULL	9
8	NULL	8
9	n	1
10	nweight	7
11	obs.prev	14
12	dist.city	9
13	city.name	9
14	urban.area	10
15	eprev.NR	27
16	NULL	25
17	NULL	26
18	NULL	31
19	NULL	30
20	eprev.N	27
21	NULL	25
22	NULL	26
23	NULL	31
24	NULL	30
25	eprev.NRU	28
26	cc.NRU	26
27	cr.NRU	27
28	cnc.NRU	32
29	qual.NRU	31
30	NULL	28
31	NULL	26
32	NULL	27
33	NULL	32
34	NULL	31

```
> str(cm.prev.check)
'data.frame':   466 obs. of  19 variables:
 $ cluster   : int  1 10 100 101 102 103 104 105 106 107 ...
 $ x         : num   9.72 13.57 11.23 14.71 11.55 ...
 $ y         : num   4.04 10.25  4.74 10.43  3.88 ...
 $ residence : Factor w/ 2 levels "Rural","Urban": 2 1 2 1 2 2 2 2 1 2 ...
 $ region    : num  3 5 2 5 12 7 8 12 7 3 ...
 $ reg.name  : Factor w/ 12 levels "ADAMAOUA","CENTRE",..: 3 5 2 5 12 7 8 12 7 3 ...
 $ n         : num  9 22 18 8 17 36 57 16 16 10 ...
 $ nweight   : num  10.8 34.8 28.7 12.7 22.1 ...
 $ obs.prev  : num   0.00 13.63  0.00  0.00  5.81 ...
 $ dist.city : num    1.87  91.90 102.20  45.20   3.59 ...
 $ city.name : chr  "DOUALA" "MAROUA" "YAOUNDE" "MAROUA" ...
 $ urban.area: Factor w/ 2 levels "in urban area",..: 1 2 2 2 1 2 2 1 2 1 ...
 $ eprev.NR  : num  5.46 1.94 3.16 2.49 7.61 ...
 $ eprev.N   : num  5.46 1.94 3.16 2.49 7.61 ...
 $ eprev.NRU : num  5.46 1.94 3.16 1.54 7.61 ...
 $ cc.NRU    : num  380 369 375 384 367 356 387 370 371 386 ...
 $ cr.NRU    : num   3.11 70.93 69.52 72.50  3.73 ...
 $ cnc.NRU   : int  17 16 15 15 17 15 18 16 15 18 ...
 $ qual.NRU  : num    0.498 261.887 249.583 268.231   0.727 ...
```

```
> write.dbf(cm.prev.check, 'cm_prev.dbf')
```

10.2 Export vers un logiciel de cartographie (SIG)

La majorité des logiciels permettant de gérer des informations géolocalisées sont en capacité d'importer des fichiers au format *shapefile* et au format *asciigrid*. Ces deux formats ont été initialement développés par la société ESRI pour ses propres produits mais sont devenus des standards.

Extrait de l'encyclopédie Wikipedia (http://fr.wikipedia.org/wiki/Shapefile) :

> Le shapefile, ou "fichier de formes" est un format de fichier issu du monde des Systèmes d'Informations Géographiques (ou SIG). Initialement développé par ESRI pour ses logiciels commerciaux, ce format est désormais devenu un standard de facto, et largement utilisé par un grand nombre de logiciels libres (MapServer, Grass, Udig, MapGuide OpenSource ...) comme propriétaires.
> Il contient toute l'information liée à la géométrie des objets décrits, qui peuvent être :
> - des points
> - des lignes
> - des polygones
>
> Son extension est classiquement SHP, et il est toujours accompagné de deux autres fichiers de même nom, et d'extensions :
> - un fichier DBF, qui contient les données attributaires relatives aux objets contenus dans le shapefile
> - un fichier SHX, qui stocke l'index de la géométrie

Le format *shapefile* permet donc de décrire des points, des lignes ou des polygones. Le format *asciigrid* décrit, quant à lui, une grille de cellules et leurs valeurs.

Le package maptools fournit plusieurs fonctions pour lire et écrire dans ces formats. Le résultat de la fonction krige.prev() peut être directement exporté au format *asciigrid*.

```
> writeAsciiGrid(cm.krige, 'cm_krige_obs_prev.asc')
```

Le format *asciigrid* ne peut contenir qu'une seule grille, tandis que le tableau de données cm.krige en comporte plusieurs. Par défaut, writeAsciiGrid() exporte la première grille. Cependant, on peut lui indiquer, avec le paramètre attr, le nom ou le numéro de la grille à exporter.

L'extension usuelle des fichiers au format *asciigrid* est *.asc*.

```
> writeAsciiGrid(cm.krige, 'cm_krige_est_prev_N363.asc',
attr='est.prevalence.N363.RInf.U0.pred')
> writeAsciiGrid(cm.krige, 'cm_krige_quality_NRU.asc',
              attr=' quality.indicator.N363.R118.U12.pred')
```

Les exports au format *shapefile* sont un peu plus complexes à réaliser, les tableaux de données devant d'abord être convertis dans des formats géographiques.

Deux fonctions permettent d'automatiser ces opérations : write.boundary.shp() permet d'exporter les frontières d'un pays contenu dans un tableau de données à deux colonnes sous la forme d'un *shapefile* comportant un seul polygone. write.prev.shp() permet d'exporter au format *shapefile* un tableau de données où chaque ligne correspond à un point. Si les coordonnées des points ne sont pas contenues dans les colonnes x et y, il est possible de spécifier les colonnes adéquates avec le

paramètre coords (voir l'aide de cette fonction). Pour ces deux fonctions, il ne faut pas préciser l'extension des fichiers à créer dans leur nom, elle sera ajoutée automatiquement.

Par ailleurs, les données étant stockées dans des fichiers *dbf*, on retrouve la limitation évoquée plus haut spécifiant que la longueur des noms de colonnes ne doit pas dépasser dix caractères. Par défaut, la fonction check.names() est appliquée au tableau de données fourni à write.prev.shp(). Si le tableau de données comporte des colonnes dont le nom dépasse les dix caractères, alors l'utilisateur sera invité à renommer les noms des colonnes. Il est possible d'appeler write.prev.shp() sans que la longueur des noms de colonnes ne soit vérifiée en lui passant en paramètre check=FALSE.

```
> write.boundary.shp(cm.bounds, 'cm_bounds', 'Cameroun')
> write.prev.shp(cm.prev, 'cm_prev', lang='fr')
> write.prev.shp(cm.cities, 'cm_cities', lang='fr')
```

Les fichiers suivants sont alors créés dans le répertoire de travail :

- *cm_krige.asc,*
- *cm_bounds.shp, cm_bounds.bdf, cm_bounds.shx,*
- *cm_prev.shp, cm_prev.dbf, cm_prev.shx,*
- *cm_cities.shp, cm_cities.dbf* et *cm_cities.shx.*

10.3 Importer les résultats dans un SIG

La manière d'importer des fichiers *shapefile* et *asciigrid* diffère selon chaque logiciel. Nous vous renvoyons donc à la documentation spécifique de chacun d'eux.

Les principaux logiciels commerciaux sont en capacité d'importer des fichiers au format *shapefiles* ou *asciigrid*. Vous pouvez aussi avoir recours à des solutions SIG libres et/ou gratuites. Voici une liste des principaux SIG gratuits :

- SavGIS (http://www.savgis.org/). Développé depuis 1984 par l'IRD, SavGIS est distribué gratuitement en français, en anglais et en espagnol.
- GRASS (http://grass.itc.it/). Le plus connu des logiciels libres de cartographie, il est de plus en plus utilisé à travers le monde et permet de lire la majorité des formats de données existants. Par ailleurs, il existe des plugins permettant de faire interagir R avec GRASS.
- Quantum GIS (http://qgis.org/). Disponible en français, ce logiciel dispose d'une interface graphique relativement simple. Il peut lire les principaux formats de données et permet de réaliser facilement des cartes. Relativement simple, il permet de s'initier aux SIG.
- GMT (http://gmt.soest.hawaii.edu/). Il s'agit d'une bibliothèque logicielle permettant de réaliser des cartes vectorielles de haute qualité. Cependant, la prise en main peut être difficile dans la mesure où les cartes doivent être programmées en lignes de commandes.

Pour une présentation plus détaillée, nous vous recommandons d'aller visiter le site *framasoft* : http://www.framasoft.net/rubrique425.html.

Annexe 1 : exporter un graphique au format SVG

Pour cela, il vous faudra d'abord charger en mémoire le package *RSvgDevice*.

```
> library(RSvgDevice)
```

Ensuite, réaliser votre (vos) graphique(s) de manière habituelle. Si vous avez plusieurs fenêtres graphiques ouvertes, vous verrez que chacune possède un numéro. Dans l'en-tête de la fenêtre est indiqué si cette sortie graphique est actuellement active. Pour afficher la liste des sorties graphiques ouvertes, utilisez dev.list(). Pour savoir qu'elle est la sortie actuellement active, utilisez dev.cur().

```
> dev.list()
windows windows windows
      2       3       4
> dev.cur()
windows
      4
```

Supposons que nous souhaitons exporter en *svg* le graphique de la fenêtre 3. Nous devons tout d'abord rendre la fenêtre 3 active à l'aide de dev.set().

```
> dev.set(3)
windows
      3
> dev.cur()
windows
      3
```

Nous allons ensuite copier le contenu de la fenêtre courante (la 3 en l'occurrence) dans une sortie de type SVG, à l'aide de dev.copy(). Le paramètre devSVG permet de spécifier le type de sortie graphique désirée. Il nous faudra spécifier le nom du fichier *svg* qui doit être créé, *essai.svg* dans notre exemple. Ensuite, nous devrons fermer la sortie SVG créée à l'aide de dev.off(). Le fichier *svg* ne sera généré qu'à ce moment là. On écrira donc :

```
> dev.copy(devSVG, file='essai.svg')
devSVG
     5
> dev.off()
windows
      2
```

Le fichier *essai.svg* sera créé dans le répertoire de travail. Il pourra être lu notamment par Firefox et Inkscape.

Exemple : exporter la carte des clusters par milieu de résidence d'Alicante.

```
> data(alicante)
> map.clust(alicante.clust,alicante.bounds,lang='fr')
> library(RSvgDevice)
> dev.copy(devSVG,file='alicante-cluster.svg')
> dev.off()
```

Annexe 2 : appliquer une transparence avec Inkscape

Inkscape est un logiciel libre de dessin vectoriel, équivalent au logiciel commercial Adobe Illustrator. Il peut être téléchargé gratuitement depuis http://www.inkscape.org/. Il utilise de manière native le format *svg*.

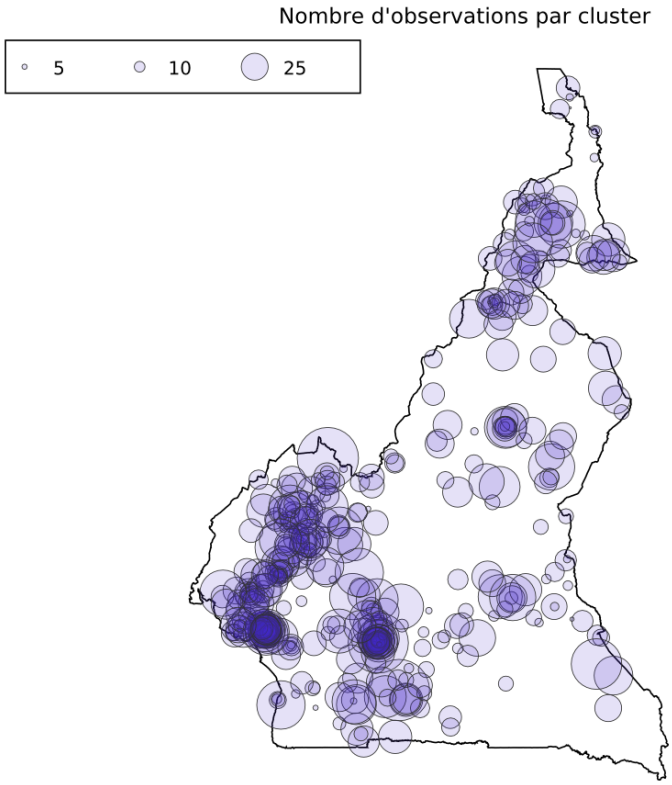
Nous reprendrons ici le graphique réalisé en 5.2, qui aura été préalablement exporté au format *svg* selon la méthode mentionnée dans l'annexe 1.

Ouvrez le fichier *svg* avec Inkscape.

Il nous faut tout d'abord sélectionner l'ensemble des cercles. Pour cela, choisissez *Édition* > *Rechercher*. Entrez *circle* dans le champ *ID*. Cela permettra de sélectionner tous les objets ayant le mot *circle* dans leur identifiant (ce qui est le cas par défaut pour un export depuis R). Cliquez sur *Recherchez* puis fermer la fenêtre de recherche.

Allez dans *Objet* > *Remplissage et contour*… Sous l'onglet *Remplissage*, modifiez le niveau de transparence (paramètre A). Vous pouvez également, si vous le souhaitez, modifier la couleur des cercles, ainsi que l'épaisseur et la couleur des contours des cercles. Dans le résultat présenté ci-dessous, nous avons opté pour une épaisseur des traits des cercles de 0,5 et une transparence de 35.

Le résultat peut ensuite être exporté dans différents formats pour intégration dans un document.



Annexe 7
Fonctions R utilisées pour la simulation d'EDS à partir d'Alicante et le calcul des N optimaux

La section 7.1 fournit le programme R utilisé pour le calcul des N optimaux dans le cadre d'une épidémie présentant une prévalence nationale de 30 %. Les programmes sont équivalents pour les autres niveaux épidémiques.

Les fonctions spécifiques utilisées pour ces simulations et donc, non inclues dans *prevR*, sont détaillées dans les sections suivantes.

Le calcul des N optimaux a commencé avant que le code de *prevR* soit entièrement réécris proprement afin de le rendre plus générique, bilingue et diffusable sous la forme d'un package. Ainsi, le code ci-après fait référence à la fonction *estimate.prev.rn.mult* qui n'est autre qu'une ancienne version de la fonction *estimate.prev* de *prevR*. Nous vous renvoyons donc à l'Annexe 4, section 4.3, pour plus de détails sur cette fonction.

7.1 Calcul des N optimaux pour une épidémie à 30 %

```
alicante.data.30.0 <- change.prev(alicante.data, 30.0)

clust <- c(300,360,432,518,622)
cible <- c(5000,6000,7200,8640,10368,12442,14930)

temp <- data.frame()
for (cl in 1:5) {
   for (ci in 1:7) {
       message("********** Epidémie 30% - Cluster ",clust[cl]," - Cible
",cible[ci]," **********")
       param.temp <- calcul.param (alicante.data.30.0, nb.tirages=100,
cluster=clust[cl], n=cible[ci],r.start=Inf,r.nb.step=1)
       minim.temp <- minim (param.temp, 30.0, cible[ci], clust[cl])
      temp <- rbind(temp,minim.temp)
      save.image("temp.Rdata")
   }
}
minim.30.0 <- temp
```

7.2 change.prev

```
change.prev <- function (data, prev) {
 new.data <- data
 prev.avant <- weighted.mean(data$prev,data$population)
 new.data$prev <- data$prev * prev / prev.avant
 message("La prévalence moyenne était de ",round(prev.avant,digits=2),"%. Elle
est maintenant de ",prev,"%.")
 return(new.data)
}
```

7.3 tirage

```
tirage <- function (data, cluster=400, n=8000, infos=TRUE, sur.ponderer=FALSE,
min.clust=15) {
 nb.cluster <- 0
 echantillon <- data.frame()
 pop <- sum(data$population)
 for (i in 1:length(levels(as.factor(data$strate)))) {
  strate <- data[data$strate==i,]
  row.names(strate) <- c(1:length(strate$x))
  pop.strate <- sum(strate$population)
  tx.strate <- pop.strate/pop
  nb.strate <- round(tx.strate*cluster, digits=0)
  if(nb.strate==0) nb.strate=1
```

```
  if (sur.ponderer&nb.strate<min.clust) {nb.strate<-min.clust}
  if (nb.strate>length(strate$x)) {nb.strate=length(strate$x)}
  if (infos) {
   message("-- Strate ",i," --")
   message("Cette strate représente ",round(100*tx.strate,digits=2),"% de la
population totale.")
   message(nb.strate," clusters ont été tirés dans cette strate.")
  }
  tir.strate <- sample(row.names(strate),nb.strate,prob=strate$population)
  echantillon.strate <- strate[tir.strate,]

  echantillon.strate$weight <- pop.strate / nb.strate

  if (i==1) { echantillon <- echantillon.strate }
  else { echantillon <- rbind(echantillon,echantillon.strate) }
  nb.cluster <- nb.cluster+nb.strate
 }

 echantillon$n <- rnorm(n=nb.cluster,mean=n/cluster, sd=n/(5*cluster))
 taille.echantillon <- sum(echantillon$n)
 echantillon$n <- round(echantillon$n*n/taille.echantillon, digits=0)

 if (length(echantillon[echantillon$n<2,]$n)>0) {echantillon[echantillon$n<2,]$n
<- 2}
 taille.echantillon <- sum(echantillon$n)
 total.weight <- sum(echantillon$weight)
 echantillon$nweight <- echantillon$weight*taille.echantillon/total.weight

 echantillon$cluster <- c(1:nb.cluster)
 for (i in (1:nb.cluster))
  { echantillon$positif[i] <-
rbinom(n=1,size=echantillon$n[i],prob=echantillon$prev[i]/100) }
 echantillon$prev_aleatoire <- echantillon$positif*100/echantillon$n
 result <-
data.frame(echantillon$cluster,echantillon$y,echantillon$x,echantillon$milieu,ech
antillon$n,echantillon$nweight,echantillon$prev_aleatoire,echantillon$prev,echa
ntillon$region)
 names(result) <-
c("CLUSTER","LATNUM","LONGNUM","RESIDENCE","N","NWEIGHT","PREV",
"PREV_REELLE","REGION")
 if (infos) {
  message("-- Tirage Final --")
  message(nb.cluster," ont été tirés au total.")
  taille.echantillon <- sum(result$N)
  message(taille.echantillon," individus ont été tirés au total.")
 }
 return(result)
}
```

7.4 ecarts.moyens

```
ecarts.moyens <- function (data, tir="NC") {
    result = data.frame()
    databis=data
    databis$ecarts<-abs(databis$PREV_EST-databis$PREV_REELLE)
    databis$N.MIN[is.na(databis$N.MIN)] = "NC"
    databis$R.MAX[is.na(databis$R.MAX)] = "NC"

    levels.n.min = levels(as.factor(data$N.MIN))
    nb.n.min = length(levels.n.min)
    if (nb.n.min == 0) {
        nb.n.min = 1
        levels.n.min = c("NC")
        }

    #Pour chaque n.min
    for (i in (1:nb.n.min)) {
        levels.r.max = levels(as.factor(data$R.MAX))
        nb.r.max = length(levels.r.max)
        if (nb.r.max == 0) {
            nb.r.max = 1
            levels.r.max = c("NC")
            }

        #Pour chaque r.max
        for (j in (1:nb.r.max)) {

            #On récupère la sous-base pour ce N.MIN et ce R.MAX
            temp <- databis[databis$N.MIN==levels.n.min[i],]
            temp <- temp[temp$R.MAX==levels.r.max[j],]

            #On calcul les indicateurs moyens
            N.MIN = temp$N.MIN[1]
            R.MAX = temp$R.MAX[1]
            ECART.MOYEN = mean(temp$ecarts)
            PREV.EST.MOY = weighted.mean(temp$PREV_EST,temp$NWEIGHT)
            TIRAGE = tir
            RADIUS.Q90 = quantile(temp$RADIUS,0.9,names=FALSE)
            RADIUS.MOYEN = mean(temp$RADIUS)

            one.result = data.frame(N.MIN, R.MAX, ECART.MOYEN,
PREV.EST.MOY, RADIUS.Q90, RADIUS.MOYEN, TIRAGE)
            result = rbind(result, one.result)
        }
    }

    result
}
```

7.5 calcul.param

```
calcul.param <- function (data, nb.tirages, n.start=50, n.by=50, n.nb.step=20,
r.start=50, r.by=25, r.nb.step=12,cluster, n) {
 result=data.frame()
 for (i in 1:nb.tirages) {
  message("TIRAGE N°",i," -------  Début --------")
  tir <- tirage (data=data, cluster=cluster, n=n, infos=FALSE,
sur.ponderer=FALSE)
  est <- estimate.prev.rn.mult(tir, n.start=n.start, n.by=n.by, n.nb.step=n.nb.step,
r.start=r.start, r.nb.step=r.nb.step,r.by=r.by)
  ecarts <- ecarts.moyens(est,tir=i)
  result = rbind(result, ecarts)
  message("TIRAGE N°",i," -------  Fin --------")
 }
 return(result)
}
```

7.6 minim

```
minim <- function (param, epidemie, n.cible, cluster) {
      EPIDEMIE <- epidemie
      N.CIBLE <- n.cible
      CLUSTER <- cluster
      result <- data.frame()

      for (i in 1:length(levels(as.factor(param$TIRAGE)))) {
         temp <- param[param$TIRAGE==i,]
         temp$N.MIN <- as.integer(as.character(temp$N.MIN))
         minim <- temp[temp$ECART.MOYEN==min(temp$ECART.MOYEN),]
         ECART.MOYEN <- minim$ECART.MOYEN
         N.MIN <- minim$N.MIN
         PREV.EST.MOY <- minim$PREV.EST.MOY
         RADIUS.Q90 <- minim$RADIUS.Q90
         RADIUS.MOYEN <- minim$RADIUS.MOYEN
         TIRAGE <- i
        one.result <- data.frame(TIRAGE, ECART.MOYEN, N.MIN,
PREV.EST.MOY, RADIUS.Q90, RADIUS.MOYEN, EPIDEMIE, N.CIBLE,
CLUSTER)
         result <- rbind (result, one.result)
      }
      return(result)
}
```

Références bibliographiques des annexes

Remarques préliminaires :

Lorsque le nombre d'auteurs est supérieur à six, seuls les trois premiers auteurs sont listés, suivi de l'abréviation *et al.* Si nous avions connaissance de l'existence d'une version électronique accessible gratuitement sur le web, l'adresse internet a été mentionnée entre parenthèses à la suite du document concerné. Les noms de villes sont suivis d'un code à deux lettres spécifiant le pays concerné. Il s'agit du code international *ISO 3166-1-alpha-2*. Un tableau de correspondance avec les noms de pays peut être obtenu sur *http://www.iso.org/iso/fr/prods-services/iso3166ma/02iso-3166-code-lists/list-fr1.html* ou bien encore *http://fr.wikipedia.org/wiki/ISO_3166-1*.

Dans le corps du texte, les références sont mentionnées, entre parenthèses, par le nom du premier auteur et l'année de publication. Lorsque plusieurs documents ont le même premier auteur et la même année de publication, une lettre minuscule a été ajoutée à l'année de publication afin de les distinguer (par exemple : 1983a, 1983b). Lorsque la référence suit une citation d'un passage précis d'un ouvrage, la ou les pages dont est extraite la citation sont mentionnées sous la forme *p. 210* pour la page 210 ou *p. 185-190* pour les pages 185 à 190.

Dans les notes de bas de page, la première occurrence d'une référence bibliographique est détaillée mais a été allégée de certaines informations (telles que les particularités de l'édition ou le nombre de pages) afin d'éviter d'allonger inutilement les notes de bas de page dans la mesure où l'ensemble de ces informations sont disponibles ci-après. À partir de la seconde occurrence d'une même référence, seuls sont mentionnés l'auteur et le titre court de l'ouvrage.

ALVAREZ M., OYONARTE S., RODRIGUEZ P. M. et HERNANDEZ J. M., « Estimated risk of transfusion-transmitted viral infections in Spain », *Transfusion*, n°42(8), 2002, pages 994-998.

BLYTH C. R. et STILL H. A., « Binomial confidence intervals », *Journal of the American Statistical Association*, n°78, 1983, pages 108-116.

BOISSON E., NICOLL A., ZABA B. et RODRIGUES L. C., « Interpreting HIV seroprevalence data from pregnant women », *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, n°13(5), 1996, pages 434-439.

BORGDORFF M., BARONGO L., VAN JAARSVELD E. *et al.*, « Sentinel surveillance for HIV-1 infection: how representative are blood donors, outpatients with fever, anaemia, or sexually transmitted diseases, and antenatal clinic attenders in Mwanza Region, Tanzania? », *AIDS*, n°7(4), 1993, pages 567-572.

BOUZITAT C., BOUZITAT P. et PAGÈS G., *Statistique, Probabilités, Estimation ponctuelle : cours et exercice d'application*, Cujas, Paris (FR), 1990, 224 pages.

BRENNER D. J. et QUAN H., « Exact Confidence Limits for Binomial Proportions-Pearson and Hartley Revisited », *The Statistician*, n°39(4), 1990, pages 391-397.

BROCKLEHURST P. et FRENCH R., « The association between maternal HIV infection and perinatal outcome: a systematic review of the literature and meta-analysis », *British Journal of Obstetrics and Gynaecology*, n°105(8), 1998, pages 836-848.

BUSCH M. P., LEE L. L., SATTEN G. A. *et al.*, « Time course of detection of viral and serologic markers preceding human immunodeficiency virus type 1 seroconversion: implications for screening of blood and tissue donors », *Transfusion*, n°35(2), 1995, pages 91-97.

BUSCH M. P., KLEINMAN S. H., JACKSON B. *et al.*, « Committee report. Nucleic acid amplification testing of blood donors for transfusion-transmitted infectious diseases: Report of the Interorganizational Task Force on Nucleic Acid Amplification Testing of Blood Donors », *Transfusion*, n°40(2), 2000, pages 143-159.

BUSCH M. P., GLYNN S. A., STRAMER S. L. *et al.*, « A new strategy for estimating risks of transfusion-transmitted viral infections based on rates of detection of recently infected donors », *Transfusion*, n°45(2), 2005, pages 254-264.

BUVE A., LAGARDE E., CARAEL M. *et al.*, « Interpreting sexual behaviour data: validity issues in the multicentre study on factors determining the differential spread of HIV in four African cities », *AIDS*, n°15 Suppl 4, 2001, pages S117-126.

CARPENTER L. M., NAKIYINGI J. S., RUBERANTWARI A. *et al.*, « Estimates of the impact of HIV-1 infection on fertility in a rural Ugandan population cohort », *Health Transition Review*, n°7(Supplement 2), 1997, pages 113-126.

CARPENTER L. M., KAMALI A., PAYNE M. *et al.*, « Independent effects of reported sexually transmitted infections and sexual behavior on HIV-1 prevalence among adult women, men, and teenagers in rural Uganda », *Journal of Acquired Immune Deficiency Syndromes*, n°29(2), 2002, pages 174-180.

CASSIGNOL C., « Note sur la construction d'intervalles de confiance pour la proportion de défectueux d'un lot à partir d'échantillons d'effectifs peu élevés », *Revue de statistique appliquée*, n°2(3), 1954, pages 43-55. (http://www.numdam.org/item?id=RSA_1954__2_3_43_0)

CHANGALUCHA J., GROSSKURTH H., MWITA W. *et al.*, « Comparison of HIV prevalences in community-based and antenatal clinic surveys in rural Mwanza, Tanzania », *AIDS*, n°16(4), 2002, pages 661-665.

CLOPPER C. J. et PEARSON E. S., « The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial », *Biometrika*, n°26(4), 1934, pages 404-413.

COPAS J. B., « Exact Confidence Limits for Binomial Proportions-Brenner & Quan Revisited », *The Statistician*, n°41(5), 1992, pages 569-572.

CRAMPIN A. C., GLYNN J. R., NGWIRA B. M. *et al.*, « Trends and measurement of HIV prevalence in northern Malawi », *AIDS*, n°17(12), 2003, pages 1817-1825.

DUMAS M., « Choix et détermination pratique d'intervalles de confiance », *Revue de statistique appliquée*, n°3(3), 1955, pages 85-101. (http://www.numdam.org/item?id=RSA_1955__3_3_85_0)

FABIANI M., ACCORSI S., LUKWIYA M. *et al.*, « Trend in HIV-1 prevalence in an antenatal clinic in North Uganda and adjusted rates for the general female population », *AIDS*, n°15(1), 2001, pages 97-103.

FABIANI M., FYLKESNES K., NATTABI B., AYELLA E. O. et DECLICH S., « Evaluating two adjustment methods to extrapolate HIV prevalence from pregnant women to the general female population in sub-Saharan Africa », *AIDS*, n°17(3), 2003, pages 399-405.

FANG C. T., FIELD S. P., BUSCH M. P. et HEYNS ADU P., « Human immunodeficiency virus-1 and hepatitis C virus RNA among South African blood donors: estimation of residual transfusion risk and yield of nucleic acid testing », *Vox Sanguinis*, n°85(1), 2003, pages 9-19.

FYLKESNES K., NDHLOVU Z., KASUMBA K., MUBANGA MUSONDA R. et SICHONE M., « Studying dynamics of the HIV epidemic: population-based data compared with sentinel surveillance in Zambia », *AIDS*, n°12(10), 1998, pages 1227-1234.

FYLKESNES K., MUSONDA R. M., SICHONE M. *et al.*, « Declining HIV prevalence and risk behaviours in Zambia: evidence from surveillance and population-based surveys », *AIDS*, n°15(7), 2001, pages 907-916.

GAGNON P., *Intervalles de confiance pour une différence de deux proportions*, mémoire pour l'obtention du grade de Maître ès Sciences (M. Sc.), sous la direction de BÉLISLE C., Université de Laval, Faculté des Sciences et de Génie, Québec (CA), 2006, 94 pages. (http://www.theses.ulaval.ca/2006/24060/24060.pdf)

GARCIA-CALLEJA J. M., ZANIEWSKI E., GHYS P. D., STANECKI K. et WALKER N., « A global analysis of trends in the quality of HIV sero-surveillance », *Sexually Transmitted Infections*, n°80 Suppl 1, 2004, pages I25-I30.

GLYNN J. R., BUVE A., CARAEL M. *et al.*, « Factors influencing the difference in HIV prevalence between antenatal clinic and general population in sub-Saharan Africa », *AIDS*, n°15(13), 2001, pages 1717-1725.

GLYNN S. A., KLEINMAN S. H., WRIGHT D. J. et BUSCH M. P., « International application of the incidence rate/window period model », *Transfusion*, n°42(8), 2002, pages 966-972.

GOSH B. K., « A comparison of some approximate confidence intervals for the binomial parameter », *Journal of the American Statistical Association*, n°74, 1979, pages 894-900.

GRAY R. H., WAWER M. J., SERWADDA D. *et al.*, « Population-based study of fertility in women with HIV-1 infection in Uganda », *Lancet*, n°351(9096), 1998, pages 98-103.

GREGSON S., ZHUWAU T., ANDERSON R. M., CHIMBADZWA T. et CHIWANDIWA S. K., « Age and religion selection biases in HIV-1 prevalence data from antenatal clinics in Manicaland, Zimbabwe », *Central African Journal of Medicine*, n°41(11), 1995, pages 339-346.

GREGSON S., TERCEIRA N., KAKOWA M. *et al.*, « Study of bias in antenatal clinic HIV-1 surveillance data in a high contraceptive prevalence population in sub-Saharan Africa », *AIDS*, n°16(4), 2002, pages 643-652.

HUNTER S. C., ISINGO R., BOERMA J. T. *et al.*, « The association between HIV and fertility in a cohort study in rural Tanzania », *Journal of Biosocial Science*, n°35(2), 2003, pages 189-199.

JACKSON D. J., NGUGI E. N., PLUMMER F. A. *et al.*, « Stable antenatal HIV-1 seroprevalence with high population mobility and marked seroprevalence variation among sentinel sites within Nairobi, Kenya », *AIDS*, n°13(5), 1999, pages 583-589.

JOLION J.-M., *Probabilités et Statistique - Cours de troisième année*, Lyon (FR), INSA, Département génie Industriel, 2006, 120 pages. (http://rfv.insa-lyon.fr/~jolion/PS/poly_stat.pdf)

KRIEGER J. N., COOMBS R. W., COLLIER A. C. *et al.*, « Fertility parameters in men infected with human immunodeficiency virus », *Journal of Infectious Diseases*, n°164(3), 1991, pages 464-469.

KWESIGABO G., KILLEWO J. Z., URASSA W. *et al.*, « Monitoring of HIV-1 infection prevalence and trends in the general population using pregnant women as a sentinel population: 9 years experience from the Kagera region of Tanzania », *Journal of Acquired Immune Deficiency Syndromes*, n°23(5), 2000, pages 410-417.

LEWIS J. J., RONSMANS C., EZEH A. et GREGSON S., « The population impact of HIV on fertility in sub-Saharan Africa », *AIDS*, n°18(Supplement 2), 2004, pages S35-43.

LOUA A., SOW E. M., MAGASSOUBA F. B., CAMARA M. et BALDE M. A., « Evaluation du risque infectieux residuel chez les donneurs de sang au Centre national de transfusion sanguine de Conakry », *Transfusion Clinique et Biologique*, n°11(2), 2004, pages 98-100.

MARTIN P. M., GRESENGUET G., HERVE V. M. *et al.*, « Decreased number of spermatozoa in HIV-1-infected individuals », *AIDS*, n°6(1), 1992, pages 130.

NDINYA-ACHOLA J. O., WAMOLA I. A., NAGELKERKE N. et AL. E., « Impact of post-partum counselling of HIV infected women of their subsequent reproduction behaviour », *Kenya AIDS Technical Bulletin*, n°1, 1990.

NEWCOMBE R. G., « Two-sided confidence intervals for the single proportion: comparison of seven methods », *Statistics in Medicine*, n°17(8), 1998, pages 857-872.

NTOZI J. P., « Widowhood, remarriage and migration during the HIV/AIDS epidemic in Uganda », *Health Transition Review*, n°7(Supplement), 1997, pages 125-144.

OUATTARA H., SIRANSY-BOGUI L., FRETZ C. *et al.*, « Residual risk of HIV, HVB and HCV transmission by blood transfusion between 2002 and 2004 at the Abidjan National Blood Transfusion Center », *Transfusion Clinique et Biologique*, n°13(4), 2006, pages 242-245.

PILLONEL J., LAPERCHE S., SAURA C., DESENCLOS J. C. et COUROUCE A. M., « Trends in residual risk of transfusion-transmitted viral infections in France between 1992 and 2000 », *Transfusion*, n°42(8), 2002, pages 980-988.

ROSS A., MORGAN D., LUBEGA R. *et al.*, « Reduced fertility associated with HIV: the contribution of pre-existing subfertility », *AIDS*, n°13(15), 1999, pages 2133-2141.

RYDER R. W., BATTER V. L., NSUAMI M. *et al.*, « Fertility rates in 238 HIV-1-seropositive women in Zaire followed for 3 years post-partum », *AIDS*, n°5(12), 1991, pages 1521-1527.

SCHREIBER G. B., BUSCH M. P., KLEINMAN S. H. et KORELITZ J. J., « The risk of transfusion-transmitted viral infections. The Retrovirus Epidemiology Donor Study », *New England Journal of Medicine*, n°334(26), 1996, pages 1685-1690.

SCHWARTLANDER B., STANECKI K. A., BROWN T. *et al.*, « Country-specific estimates and models of HIV and AIDS: methods and limitations », *AIDS*, n°13(17), 1999, pages 2445-2458.

SHANG G., SEED C. R., WANG F., NIE D. et FARRUGIA A., « Residual risk of transfusion-transmitted viral infections in Shenzhen, China, 2001 through 2004 », *Transfusion*, n°47(3), 2007, pages 529-539.

THE UNAIDS REFERENCE GROUP ON ESTIMATES MODELLING AND PROJECTIONS, « Improved methods and assumptions for estimation of the HIV/AIDS epidemic and its impact: Recommendations of the UNAIDS Reference Group on Estimates, Modelling and Projections », *AIDS*, n°16(9), 2002, pages W1-14.

THE UNAIDS REFERENCE GROUP ON ESTIMATES MODELLING AND PROJECTIONS, *Improving parameter estimation, projection, methods, uncertainty estimation and epidemic classification*, Report of a meeting, Prague (CZ), 29 novembre - 1er décembre, UNAIDS, 2006, 25 pages. (http://www.epidem.org/Publications/Prague2006report.pdf)

TOBI H., VAN DEN BERG P. B. et DE JONG-VAN DEN BERG L. T., « Small proportions: what to report for confidence intervals? », *Pharmacoepidemiology and drug safety*, n°14(4), 2005, pages 239-247.

TSAI W. Y., CHI Y. et CHEN C. M., « Interval estimation of binomial proportion in clinical trials with a two-stage design », *Statistics in Medicine*, n° en cours d'impression, DOI: 10.1002/sim.2930, 2007.

UNAIDS/WHO, *Reconciling antenatal clinic-based surveillance and population-based survey estimates of HIV prevalence in sub-Saharan Africa*, UNAIDS/WHO, Genève (CH), 2003, 30 pages.

VELATI C., ROMANO L., BARUFFI L. *et al.*, « Residual risk of transfusion-transmitted HCV and HIV infections by antibody-screened blood in Italy », *Transfusion*, n°42(8), 2002, pages 989-993.

VESSEREAU A., « Sur l'intervalle de confiance d'une proportion logique «classique» et logique «bayesienne» », *Revue de statistique appliquée*, n°26(2), 1978, pages 5-31. (http://www.numdam.org/item?id=RSA_1978__26_2_5_0)

VOLLSET S. E., « Confidence intervals for a binomial proportion », *Statistics in Medicine*, n°12(9), 1993, pages 809-824.

WALKER N., GARCIA-CALLEJA J. M., HEATON L. *et al.*, « Epidemiological analysis of the quality of HIV sero-surveillance in the world: how well do we track the epidemic? », *AIDS*, n°15(12), 2001, pages 1545-1554.

WIDY-WIRSKI R., BERKLEY S., DOWNING R. *et al.*, « Evaluation of the WHO clinical case definition for AIDS in Uganda », *JAMA*, n°260(22), 1988, pages 3286-3289.

WILSON E. B., « Probable inference, the law of succession, and statiscal inference », *Journal of the American Statistical Association*, n°22, 1927, pages 209-212.

WONNACOTT T. H. et WONNACOTT R. J., *Statistique : économie - gestion - sciences - médecine*, Economica, réimpression de 1995 de la traduction française de 1991 à partir de la quatrième édition américaine de 1990, 1972 pour la première édition originale, Paris (FR), 1990, 920 pages.

YATES F., « Contingency table involving small numbers and the Khi² test », *Supplement to the Journal of the Royal Statistical Society*, n°1(2), 1934, pages 217-235.

ZABA B. W. et GREGSON S., « Measuring the impact of HIV on fertility in Africa », *AIDS*, n°12 Suppl 1, 1998, pages S41-S50.

ZABA B. W., CARPENTER L. M., BOERMA J. T. *et al.*, « Adjusting ante-natal clinic data for improved estimates of HIV prevalence among women in sub-Saharan Africa », *AIDS*, n°14(17), 2000, pages 2741-2750.

Table des matières des annexes

Liste des tableaux annexes

Liste des figures annexes

Liste des équations annexes

Prévalences du VIH en Afrique : validité d'une mesure

Les prévalences nationales du VIH sont estimées, en Afrique subsaharienne, à partir de deux sources : la surveillance sentinelle des femmes enceintes et les enquêtes nationales en population générale (EDS). En plusieurs endroits, les résultats divergent, questionnant la validité de chaque approche. Quelles portée, limites et signification objective peuvent être accordées aux diverses observations, chacune appréhendant le réel sous un angle différent ?
Les EDS constituent un bon indicateur du niveau des épidémies à l'échelle national et régional, voir infrarégional grâce au recours à des techniques d'analyse spatiale en composantes d'échelles. Mais leur fréquence est inadaptée pour mesurer les évolutions à court terme. La surveillance sentinelle, estimateur local des ordres de grandeur, peut être un indicateur de tendances sous certaines conditions. Cependant, si la situation actuelle commence à être mieux connue, une mesure réelle des dynamiques est encore hors de notre portée.

Mots-Clés

VIH/SIDA, Afrique subsaharienne, prévalence, validité épistémologique, EDS (Enquêtes Démographiques et de Santé), surveillance sentinelle, interpolation spatiale, épidémiologie.

HIV prevalence in Africa: validity of a measurement

HIV national prevalence is estimated, in sub-Saharan Africa, from two main data sources: sentinel surveillance of pregnant women and national population-based surveys (DHS). In several countries, results differ, questioning each of those approaches' efficiency. What range, limitations and objective significance can be granted to the diverse observations, each apprehending the real from a different point of view?
DHS constitute a good indicator of prevalence levels at a national or a regional scale. Intraregional variations can be reproduced by using spatial analysis techniques. But DHS frequencies are inaccurate to measure short-term evolution. Sentinel surveillance, local estimator of magnitude of epidemics, can be a tendency indicator under certain conditions. However, if the present situation is only just starting to be better-known, a real measurement of HIV dynamics is still out of our reach.

Key Words

HIV/AIDS, sub-Saharan Africa, prevalence, epistemological validity, DHS (Demographic and Health Surveys, sentinel surveillance, spatial interpolation, epidemiology.

Résumé

Key Words

Mots Clés

Abstract