Contribution des modèles à classes latentes à l'étude de la répartition spatio-temporelle des vecteurs de Paludisme et à l'étude temporelle de l'observance aux antirétroviraux chez les patients VIH

Olayidé Boussari

▶ **To cite this version:**

Olayidé Boussari. Contribution des modèles à classes latentes à l'étude de la répartition spatio-temporelle des vecteurs de Paludisme et à l'étude temporelle de l'observance aux antirétroviraux chez les patients VIH. Statistiques [math.ST]. Université Claude Bernard - Lyon I; Université d'Abomey-Calavi (Bénin), 2014. Français. NNT : 2014LYO10095 . tel-01617239

HAL Id: tel-01617239
https://theses.hal.science/tel-01617239

Submitted on 16 Oct 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

N° d'ordre 95 Année 2014

THESE DE L'UNIVERSITE DE LYON

délivrée par

L'UNIVERSITE CLAUDE BERNARD LYON 1

et préparée en cotutelle avec

L'UNIVERSITE D'ABOMEY-CALAVI (BENIN)

École doctorale Évolution Écosystèmes Microbiologie Modélisation (ED 341)

DIPLOME DE DOCTORAT

(arrêté du 7 août 2006)

soutenue publiquement le 16 Juin 2014

par

Olayidé BOUSSARI

Contribution des modèles à classes latentes à l'étude de la répartition spatio-temporelle des vecteurs de Paludisme et à l'étude temporelle de l'observance aux antirétroviraux chez les patients VIH.

Directeur de thèse: **René ECOCHARD**

Co-directeur de thèse: **Noël FONTON**

JURY : M. André GARCIA (Directeur de recherche)
M. Jean-François ETARD (Directeur de recherche)
M. Aliou DIOP (Professeur)
M. Norbert HOUNKONNOU (Professeur)

UNIVERSITE CLAUDE BERNARD - LYON 1

Président de l'Université	**M. François-Noël GILLY**
Vice-président du Conseil d'Administration	M. le Professeur Hamda BEN HADID
Vice-président du Conseil des Etudes et de la Vie Universitaire	M. le Professeur Philippe LALLE
Vice-président du Conseil Scientifique	M. le Professeur Germain GILLET
Directeur Général des Services	M. Alain HELLEU

COMPOSANTES SANTE

Faculté de Médecine Lyon Est – Claude Bernard	Directeur : M. le Professeur J. ETIENNE
Faculté de Médecine et de Maïeutique Lyon Sud – Charles Mérieux	Directeur : Mme la Professeure C. BURILLON
Faculté d'Odontologie	Directeur : M. le Professeur D. BOURGEOIS
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : Mme la Professeure C. VINCIGUERRA
Institut des Sciences et Techniques de la Réadaptation	Directeur : M. le Professeur Y. MATILLON
Département de formation et Centre de Recherche en Biologie Humaine	Directeur : M. le Professeur P. FARGE

COMPOSANTES ET DEPARTEMENTS DE SCIENCES ET TECHNOLOGIE

Faculté des Sciences et Technologies	Directeur : M. le Professeur F. De MARCHI
Département Biologie	Directeur : M. le Professeur F. FLEURY
Département Chimie Biochimie	Directeur : Mme le Professeur H. PARROT
Département GEP	Directeur : M. N. SIAUVE
Département Informatique	Directeur : M. le Professeur S. AKKOUCHE
Département Mathématiques	Directeur : M. le Professeur A. GOLDMAN
Département Mécanique	Directeur : M. le Professeur H. BEN HADID
Département Physique	Directeur : Mme S. FLECK
Département Sciences de la Terre	Directeur : Mme la Professeure I. DANIEL
UFR Sciences et Techniques des Activités Physiques et Sportives	Directeur : M. C. COLLIGNON
Observatoire des Sciences de l'Univers de Lyon	Directeur : M. B. GUIDERDONI
Polytech Lyon	Directeur : M. P. FOURNIER
Ecole Supérieure de Chimie Physique Electronique	Directeur : M. G. PIGNAULT
Institut Universitaire de Technologie de Lyon 1	Directeur : M. C. VITON
Institut Universitaire de Formation des Maîtres	Directeur : M. A. MOUGNIOTTE
Institut de Science Financière et d'Assurances	Administrateur provisoire : M. N. LEBOISNE

A Olabayo

Remerciements

Plusieurs personnes, de près ou de loin et à des degrés divers ont contribué à l'aboutissement du présent travail. Je tiens à adresser mes sincères remerciements :

Au Professeur René Ecochard pour avoir accepté de diriger mon travail de thèse. Tu m'as initié à la recherche et j'ai tellement appris à tes côtés. Aussi ton soutien, semblable à celui d'un père, ne m'a jamais fait défaut.

Au Professeur Noël Fonton qui a codirigé cette thèse et dont les conseils à tous les niveaux m'ont assez aidé.

Au Professeur Norbert Hounkonnou pour sa disponibilité et ses conseils si constructifs.

Au Professeur Elisabeth Gassiat pour m'avoir proposé et convaincu d'aller en thèse et aussi pour avoir accepté de juger le fruit de ces quelques années de travail.

Au Professeur Aliou Diop pour avoir accepté de rapporter ce travail. Je reste marqué par votre sens d'humilité et le cadre convivial que vous avez su donner à nos quelques échanges.

Au Docteur André Garcia pour l'intérêt que vous avez toujours porté à ce travail et pour avoir accepté de faire partie du jury.

Au Docteur Vincent Corbel pour m'avoir accueilli dans l'équipe de l'UMR224/IRD à Cotonou (Bénin) ; vous avez été pour beaucoup dans le déroulement et l'aboutissement de ce travail. C'est pour moi l'occasion de dire aussi un sincère merci à vos successeurs Franck Remoué et Cédric Pennetier à la tête de cette équipe.

Au Professeur Martin Akogbéto, directeur du CREC, pour m'avoir accepté dans cette structure durant mes travaux de thèse.

Aux autorités et au personnel de la CIPMA (Chaire UNESCO Cotonou) pour leur appui tout au long de cette thèse. Je pense en particulier au Docteur Ezinvi Baloïtcha dont je peux, sans aucune exagération, qualifier la disponibilité de légendaire.

A tous les enseignants du master de Statistique Appliquée au Vivant de Cotonou, en particulier le Professeur Simplice Dossou-Gbété avec qui j'ai véritablement pris goût aux statistiques.

Au Docteur Jean-Christophe Ernould qui promptement a permis d'établir mes premiers contacts avec le Professeur René Ecochard et qui m'a permis de faire mes premiers pas dans l'enseignement des Biostatistiques à travers le Réseau Epidémiologie et Développement (RED/IRD).

A tous les membres de l'équipe UMR224/IRD au Bénin ; particulièrement aux futurs chercheurs de l'équipe devenus chercheurs entre temps (ils se reconnaitront certainement) avec qui j'ai partagé et continue de partager sur le plan scientifique mais aussi et surtout sur le plan humain.

A tous les membres du Service de Biostatistique des Hospices Civils de Lyon qui m'ont cordialement accueilli dans leur dynamique et solidaire équipe durant mes séjours à Lyon. Je pense particulièrement à Stéphanie, à Michèle et à Mariethé qui ont tant fait pour moi... Une mention spéciale à Jean avec qui j'ai partagé d'agréables et inoubliables moments d'échange et de travail.

Aux autorités du SCAC de l'ambassade de France à Cotonou et à celles des Hospices Civils de Lyon, de même qu'aux responsables du projet Edulink et ceux de l'UMR224/IRD au Bénin, pour leurs soutiens financiers.

A Eric pour ton précieux soutien pendant les derniers moments de mise en forme de ce document.

A Sylviane pour ton soutien et surtout pour avoir supporté mes longues journées de travail et mes longues périodes d'absence pendant ces trois dernières années.

A mes parents, à mes frères et à ma sœur pour leurs constants soutien et encouragements durant toutes ces années de dur labeur.

Enfin à Dieu tout Puissant, je veux témoigner ma plus grande reconnaissance. C'est vrai, Père, tu es le commencement et la fin.

"...all models are wrong, but some are useful."

George E.P. Box (1919-2013)

Résumé

Ce travail est construit autour de deux problématiques de santé relatives aux deux plus grandes pandémies qui sévissent en Afrique sub-saharienne : i) l'hétérogénéité rencontrée dans la répartition spatiale et temporelle des vecteurs de paludisme ; ii) la variabilité dans l'observance au traitement antirétroviral par des personnes vivant avec le virus de l'immunodéficience humaine. Sur le plan méthodologique, ces deux problèmes se rapportent à la prise en compte de l'hétérogénéité dans la modélisation de données issues de mesures répétées ; ils nécessitent en outre le développement d'outils statistiques permettant de distinguer à partir des données, des sous groupes (de localités, d'individus...) homogènes indispensables pour rendre plus efficientes les mesures de santé souvent déployer par les praticiens dans le cadre de la lutte contre le paludisme ou le VIH/SIDA.

Les modèles de mélanges finis, grâce à leur flexibilité, sont des outils capables de fournir non seulement de bonnes estimations en présence d'une grande hétérogénéité dans les observations mais aussi une bonne partition des unités statistiques. Nous les distinguons, parmi d'autres méthodes, comme étant adaptés aux problématiques du présent travail.

Deux applications de ces modèles aux données issues de capture de moustiques ont permis de modéliser la répartition spatiale et temporelle de vecteurs de paludisme et de dégager une méthode simple d'évaluation d'impact de mesures de lutte antivectorielle. Nous introduisons la notion de «trajectoires de variances» dans une troisième application portant sur des données d'observance aux traitements antirétroviraux par des personnes vivant avec le virus de l'immunodéficience humaine.

Mots clés : hétérogénéité ; mesures répétées ; modèles de mélange ; classification non supervisée ; vecteurs de paludisme ; antirétroviraux ; VIH/SIDA.

Abstract

This work focuses on two health issues relating to two major pandemics in sub-Saharan Africa : i) the heterogeneity encountered in the spatial and temporal distribution of malaria vectors ; ii) the variability in adherence to antiretroviral treatment by people living with the human immunodeficiency virus. Methodologically, these two problems are related to the consideration of the heterogeneity in the modeling of data from repeated measurements. They also require the development of statistical tools to distinguish from the data, homogeneous clusters of localities, individuals... that are needed to make more efficient health measures often deployed by practitioners in the fight against malaria and HIV/AIDS.

The finite mixture models, due to their flexibility, are statistical tools that not only provide good estimates in the presence of heterogeneity in the observations but also a good classification of statistical units. We show that they are able to deal with the problematics of our study.

The spatial and temporal distributions of malaria vectors are modeled through two different applications of finite mixture models and a simple tool to evaluate the impact of vector control methods is generated. We introduce a "variance trajectories" method in a third application of finite mixture models to data on adherence to antiretroviral therapy by people living with human immunodeficiency virus.

Keywords : heterogeneity ; repeated measurements ; mixture models ; unsupervised classification ; malaria vectors ; antiretroviral ; HIV/AIDS.

Table des matières

Introduction

Le souci d'étudier l'évolution d'un phénomène sur un ensemble d'unités (sujets, machines, villes...) amène souvent, dans de nombreuses études, à effectuer des mesures répétées du phénomène sur cet ensemble. La mesure du niveau d'anxiété chez des travailleurs de différentes entreprises, la mesure de la pression artérielle systolique sur des patients et sous différentes conditions (assis, couché, après un effort,...) ou encore l'évolution sur 50 années du nombre annuel de jours de pluies dans différentes localités sont quelques exemples de mesures répétées. Elles conduisent à des données dites multiniveaux ou hiérarchisées ou encore des données en clusters. Les données longitudinales qui sont des mesures répétées dans le temps sur un ensemble d'unités constituent le cas le plus répandu des mesures répétées.

Différents outils statistiques ont été développés pour prendre en compte les structures de dépendance des observations ou de leurs appartenances à des groupes donnés dans le cas des données répétées. Les modèles à effets aléatoires sont les plus répandus de ces outils. L'idée générale dans ces modèles est que les dépendances observées dans les données proviennent de variables qui ne peuvent être directement observées, donc des variables latentes. De ce fait, les modèles à effets aléatoires peuvent être regardés comme des modèles à variables latentes. Le modèle à effets aléatoires le plus simple et le plus ancien est celui de l'analyse de la variance qui remonte vraisemblablement à Airy (1861) puis Fisher (1925). Il est défini par la formule :

$$y_{ij} = \beta + b_i + e_{ij} \tag{1}$$

où :

l'indice i, $1 \leqslant i \leqslant n$, désigne une unité statistique particulière et j, $1 \leqslant j \leqslant n_i$, une observation sur l'unité i. y_{ij} désigne les variables dépendantes ou les réponses

observées, b_i une variable latente continue (l'effet aléatoire) spécifique à l'unité i et e_{ij} les résidus. b_i et e_{ij} sont supposés gaussiens. β désigne la moyenne des réponses et constitue la partie fixe du modèle.

En introduisant des covariables x_{ij} et z_i respectivement dans la partie fixe et la partie aléatoire du modèle (1), on obtient la forme générale des modèles linéaires mixtes (Laird and Ware, 1982) dont une écriture est la suivante :

$$y_{ij} = x'_{ij}\beta + z_i b_i + e_{ij} \tag{2}$$

Dans les cas où l'on ne peut pas faire directement l'hypothèse de normalité pour les lois des erreurs sur les variables dépendantes (réponses binaires ou résultant de comptages par exemple), on applique le principe du modèle linéaire généralisé (McCullagh and Nelder, 1989) pour obtenir un modèle linéaire généralisé mixte. Une forme d'un tel modèle est la suivante :

$$g\left[\mathbb{E}(y_{ij}|b_i)\right] = x'_{ij}\beta + z_i b_i \tag{3}$$

avec g la fonction de lien et $\mathbb{E}$ désignant l'espérance mathématique.

Toutefois, il existe des cas de données répétées dont la distribution ne peut être approchée par des modèles à effets aléatoires faisant l'hypothèse forte d'une distribution multinormale des effets aléatoires. Il devient alors nécessaire d'aller au-delà des modèles linéaires généralisés mixtes et cela requiert l'utilisation par exemple d'une ou de plusieurs variables latentes discrètes. Les modèles ainsi obtenus sont appelés modèles à classes latentes ou modèles de mélange (Droesbeke et al., 2013) dans lesquels les estimations se font généralement par la méthode du maximum de vraisemblance (MV) non paramétrique (voir par exemple (Aitkin, 1996) et (Aitkin, 1999)), ce qui leur confère une très grande flexibilité. Les modèles de mélanges offrent entre autres avantages, la possibilité de classer les unités statistiques en utilisant les différences entre les coefficients de régression et de proposer une interprétation

significative à cette partition. Dans le cas d'un mélange simple utilisant une seule variable latente discrète à K catégories, le modèle s'écrit comme suit :

$$g\left[\mathbb{E}(y_{ij}|i \in k)\right] = x'_{ij}\beta + b_k \tag{4}$$

avec les $\mathbf{P}(i \in k) = \pi_k$ correspondant aux proportions des différentes catégories du mélange. Elles sont dites probabilités *a priori* et vérifient $\sum_{k=1}^{K} \pi_k = 1$.

Nous proposons dans ce mémoire de thèse un cadre d'application des modèles de mélange à travers deux types de données longitudinales relatives aux deux plus grandes pandémies en Afrique subsaharienne et montrons ainsi l'intérêt de ces types de modèles en santé publique.

La première partie de ce travail donne quelques rappels généraux sur le paludisme puis le VIH/SIDA. Dans chaque cas, le problème de modélisation d'un jeu de données relatif à chacune de ces maladies est posé ; l'objectif étant de répondre à certains questionnements à l'origine des collectes de ces données. Il s'agit d'une part de données issues de captures répétées de moustiques dans le cadre de la lutte contre le paludisme et d'autre part de données provenant de mesures répétées de l'observance chez des personnes vivant avec le virus de l'immunodéficience humaine (VIH) et mis sous traitement antirétroviral (ARV). Cette partie montre ensuite comment ces types de données peuvent être convenablement traités par des modèles à classes latentes et se termine par un développement sur les fondements théoriques des modèles à classes latentes et des méthodes d'estimation permettant de faire de l'inférence à partir de ces modèles.

La seconde partie expose successivement les traitements concrets, à l'aide de modèles à classes latentes, des différents jeux de données présentés dans la première partie. Il consiste donc en la présentation de trois articles de recherche répondant aux problématiques posées dans la première partie.

Nous concluons notre travail en pointant l'intérêt de nos résultats dans l'amélioration

des méthodes actuelles de lutte antivectorielle et dans l'amélioration de la prise en charge des personnes vivant avec le VIH/SIDA. Nous discutons aussi des limites des méthodes de traitement statistiques que nous avons adoptées et ouvrons sur quelques perspectives.

Première partie

Contextes, Problématiques, Modèles à classes latentes

Chapitre 1

Paludisme et problématique de la répartition spatiale et temporelle des vecteurs

1.1 Généralités sur le Paludisme

Le paludisme est une maladie infectieuse à transmission vectorielle faisant intervenir trois acteurs : un homme (jouant le rôle d'hôte) est infecté par un protozoaire parasite du genre *Plasmodium* qui lui a été transmis par la piqûre d'un moustique vecteur du genre *Anopheles*.

La découverte de l'agent causal du paludisme par Laveran (1880) puis celle du rôle vecteur de l'anophèle dans sa transmission par Ross (1897) et Grassi (1899) remontent à bien plus d'un siècle. Le paludisme se caractérise par des épisodes fébriles aigus et peut être mortel. Les symptômes (fièvre, maux de tête, frissons et vomissements) apparaissent au bout de sept jours ou plus (généralement 10 à 15 jours) après la piqûre de moustique infectante.

L'intensité de la transmission du paludisme dépend de facteurs liés au parasite, au vecteur, à l'hôte humain et à l'environnement. La transmission est plus intense aux endroits où les espèces de vecteurs ont une durée de vie relativement longue et piquent plutôt les êtres humains que les animaux. Par exemple, la longue durée de vie et la forte préférence pour l'homme des espèces africaines de vecteurs expliquent que plus de 90% des décès par paludisme enregistrés dans le monde surviennent en

Afrique subsaharienne (WHO, 2014). La transmission dépend aussi des conditions climatiques qui peuvent influer sur l'abondance et la survie des moustiques, telles que le régime des précipitations, la température et l'humidité. L'immunité humaine est un autre facteur important, en particulier chez les adultes dans les zones de transmission modérée à intense. L'immunité se développe après des années d'exposition et réduit le risque que l'infection palustre cause des troubles sévères. C'est la raison pour laquelle la plupart des décès par paludisme en Afrique surviennent chez de jeunes enfants, tandis que, dans les zones de faible transmission et où la population est peu immunisée, tous les groupes d'âge sont exposés.

Depuis des décennies, d'importants efforts ont été déployés pour venir à bout du paludisme mais ils n'ont malheureusement pas permis d'éviter que la maladie demeure de nos jours une des premières causes de mortalité dans le monde et en particulier en Afrique sub-saharienne. En effet, l'Organisation Mondiale de la Santé (OMS) rapporte que 3,3 millions de décès imputables au paludisme ont été évités entre 2001 et 2012, et que 90% des décès évités concernaient des enfants de moins de cinq ans en Afrique subsaharienne (WHO, 2013). Cependant, le même rapport estime qu'en 2012, environ 207 millions de cas et 627000 décès étaient imputables au paludisme.

Le diagnostic et le traitement précoces du paludisme permettent de réduire l'intensité et la transmission de la maladie et aussi d'éviter qu'elle ne devienne mortelle. La lutte antivectorielle reste le principal moyen de prévention qui permet de réduire de façon significative la transmission du paludisme au niveau communautaire. La maladie peut également être prévenue au moyen d'antipaludiques. Toutefois La résistance du parasite aux antipaludiques et celle du vecteur aux méthodes de lutte antivectorielle (voir par exemple (Dondorp et al., 2009) et (Ranson et al., 2011)) demeurent des problèmes récurrents dans la lutte contre le paludisme et expliquent en majorité les difficultés d'éradication de la maladie.

1.2 Vecteurs

Le paludisme est transmis par des moustiques du genre *Anopheles*. La reconnaissance des espèces vectrices est capitale pour mesurer le rôle joué par chacune d'elles dans la transmission, pour identifier et donc « cibler » les vecteurs dans un programme de lutte (Carnevale et al., 2009). Compte tenu des critères de classification taxinomique des anophèles, les espèces appartenant à un même groupe sont très proches morphologiquement et présentent des différences à au moins un stade de leur développement tandis que les espèces appartenant à un même complexe sont identiques sur le plan morphologique à tous les stades de leur développement.

Hambach (2004) a rapporté, que le genre *Anopheles* comprend environ 484 espèces sur la planète dont une soixantaine est vectrice de *Plasmodium*. L'Afrique sub-saharienne regorge d'une trentaine d'espèces vectrice de *Plasmodium* sur environ les 150 espèces d'anophèles dénombrées sur le continent. La transmission du paludisme en Afrique sub-saharienne est majoritairement assurée (95%) par 5 espèces : *An. gambiae s.s.* (*An. coluzzii* et *An. gambiae* (Coetzee et al., 2013)), *An. arabiensis, An. funestus s.s, An. moucheti* et *An. nili* (Mouchet et al., 2004). Ce sont les femelles d'anophèle qui s'alimentent sur des humains (Coluzzi et al., 1979). La longévité moyenne est de 3 à 4 semaines pour les principaux vecteurs de paludisme en Afrique sub-saharienne (Gillies, 1961; Gillies and Wilkes, 1965).

Les larves d'*An. gambiae s.s.* et *An. arabiensis* se développent de préférence dans les eaux douces, peu chargées en matière organique, peu profondes, calmes, ensoleillées et sans végétation (Pages et al., 2007). Les femelles d'*An. gambiae s.s* et *An. arabiensis* piquent généralement à l'intérieur des maisons (endophagie). *An. funestus* qui a besoin d'ombre pour se développer, abonde au niveau des marais à végétation dressée et au niveau des rizières et assure en fin de saison des pluies, le relais d'*An. gambiae* dans la transmission (Djènontin et al., 2010) ; ceci explique qu'*An. funestus* soit un vecteur principal du paludisme en Afrique.

1.3 Agent pathogène et son cycle de vie

Lors de la prise de son repas sanguin, l'anophèle libère dans le sang humain les parasites responsables du paludisme chez l'homme. Ces parasites sont des protozoaires appartenant à 5 espèces du genre Plasmodium :

- *P. falciparum* est celui qui est responsable des formes cliniques potentiellement mortelles, qui développe des résistances aux antipaludiques et qui est le plus largement répandu à travers le monde. Dans les régions équatoriales, il est transmis toute l'année avec cependant des recrudescences saisonnières. Dans les régions sub-tropicales, il ne survient qu'en période chaude et humide. Presque trois-quarts des cas d'infection due au *P. falciparum* dans le monde sont enregistrés en Afrique (Snow et al., 2005).

- *P. vivax* est très largement répandu en Amérique du Sud et en Asie mais beaucoup plus rarement observé en Afrique à cause d'une incompatibilité des antigènes du groupe sanguin. L'affection par *P. vivax* est classiquement considérée comme bénigne (fièvre tierce bénigne), toutefois elle peut être une source d'anémie chez l'enfant.Une caractéristique particulière de l'infection à *P. vivax* est que ce parasite peut rester dans le foie sous forme d'hypnozoites pendant de longues périodes et être relégué progressivement par un phénomène de relapses (Imwong et al., 2012).

- *P. ovale* sévit en Afrique intertropicale (Centre et Ouest) et dans certaines régions du Pacifique et provoque une fièvre tierce bénigne.

- *P. malariae* sévit principalement en Afrique subsaharienne et de manière beaucoup plus sporadique. Ce parasite est également présent en Asie du Sud Est. L'affection par *P. malariae* est caractérisée par une fièvre quarte bénigne mais peut parfois entraîner des complications rénales.

- *P. knowlesi* était seulement rencontre chez les singes en Asie du Sud-Est mais il a été rapporté recemment chez l'homme dans la même région (Cox-Singh and Singh, 2008). Au microscope *P. knowlesi* ressemble à *P. malariae* mais contrairement à ce dernier, il peut être létal pour l'homme. Toutefois, il est à ce jour, sensible à la

simple chloroquine.

Le cycle biologique des *Plasmodium* se déroule alternativement entre deux hôtes : l'homme qui joue le rôle d'hôte intermédiaire où se déroule la partie asexuée du cycle et le moustique anophèle qui représente l'hôte définitif où a lieu la reproduction sexuée (Figure 1.1). Les *Plasmodium* pénètrent dans l'organisme humain sous forme de sporozoïtes à la faveur d'une piqûre d'un moustique infectant. Les sporozoïtes se multiplient ensuite dans les cellules du foie après y avoir été transportés par la circulation sanguine. Libérés dans le sang sous forme de mérozoïtes, ils envahissent les globules rouges et deviennent des schizontes dont la multiplication entraîne l'éclatement des globules rouges. Il en résulte des accès de fièvre paludéenne. Les schizontes peuvent alors infecter d'autres globules rouges ou se transformer en gamétocytes mâles et femelles.

Par la suite, un moustique se contamine par piqûre, en absorbant du sang contenant des gamétocytes qui passeront à l'étape de gamètes mâles et femelles dans le tube digestif de l'insecte. Ces gamètes fusionnent en un œuf libre, mobile appelé ookinète qui quittera la lumière du tube digestif pour se fixer ensuite à la paroi externe de l'estomac et se transforme en oocyste. Les cellules parasitaires se multiplient à l'intérieur de cet oocyste, produisant des centaines de sporozoïtes qui migrent ensuite dans les glandes salivaires du moustique, d'où ils pourront contaminer un nouvel individu lors d'un repas sanguin.

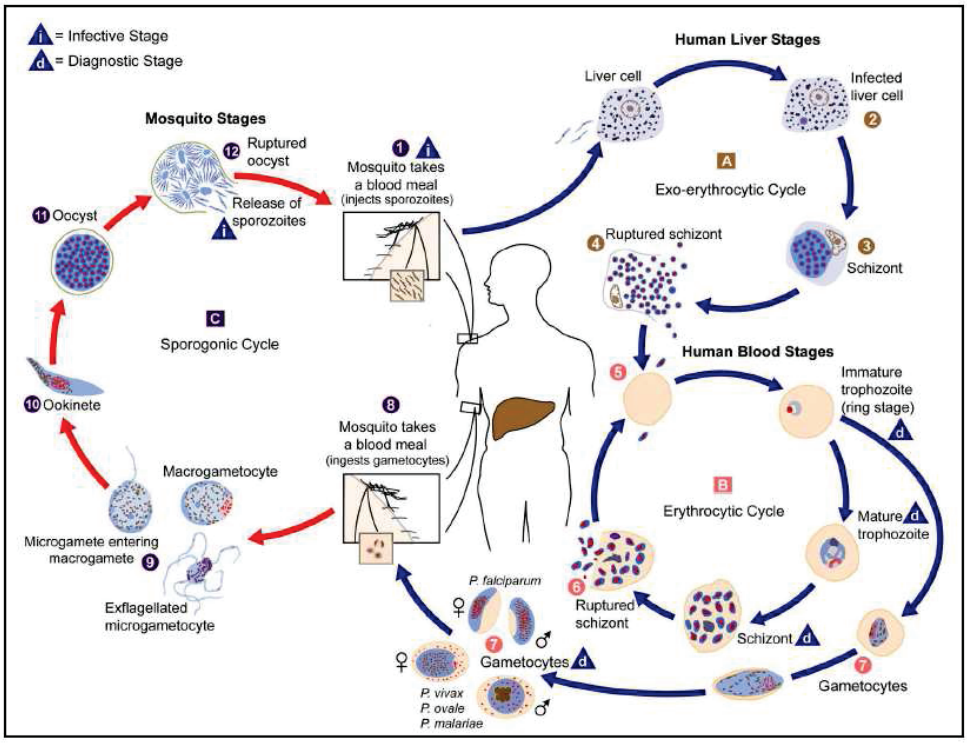


FIGURE 1.1 – Cycle de développement et de reproduction des *Plasmodium spp.* (CDC 2002)

1.4 Lutte antivectorielle

L'éradication du paludisme n'est envisageable qu'avec une lutte antivectorielle (LAV) réussie. Le premier plan de lutte intégrée contre le paludisme remonte au début du vingtième siècle et fut très largement concentré sur la lutte contre les vecteurs (Gorgas, 1910). L'histoire de la LAV est faite d'une succession de grands espoirs de réussite et de désillusions liées essentiellement à la capacité des vecteurs à développer des mécanismes de résistance aux méthodes de LAV (voir par exemple (Corbel and N'Guessan, 2013)). Nous pouvons donner ici l'exemple du DDT (Dichloro-Diphenyl- Trichlororethane) qui avait aidé à éradiquer le paludisme dans plusieurs pays européens et américains au début des années 50 ; puis l'apparition dès 1951, de résistances au DDT chez des anophèles vecteurs (Livadas, 1951; Livadas and Georgopoulos, 1953). Les ambitions de la LAV dans les temps actuels se déclinent en six points (Najera and Zaim, 2005) : la prévention et la lutte contre

les épidémies, l'élimination des nouveaux foyers dans les zones indemnes de paludisme, la prévention des pics saisonniers de transmission, le contrôle de la transmission dans les zones à risques, la réduction de la transmission dans les zones où la chimiorésistance des plasmodies est élevée, la lutte contre le paludisme endémique.

Selon l'élément de la biologie du vecteur qu'on cherche à réduire, une méthode de LAV peut être plus adaptée qu'une autre. Le tableau 1.1 donne une classification des méthodes de LAV les plus couramment utilisées en fonction de l'élément de la biologie du vecteur qui est visé.

TABLE 1.1 – Méthodes de lutte anti-vectorielle et paramètres de la biologie du vecteur visé

	Élément de la biologie du vecteur ciblé			
Méthode de lutte	Densité	Contact hôte-vecteur	Longévité	Compétence vectorielle
Lutte anti-larvaire	X			
Moustiquaires imprégnées	X	X	X	
Aspersions intra-domiciliaires	X		X	
Bâches murales imprégnées	X		X	
Grillage de fenêtres		X		
Répulsifs		X		
Pulvérisations spatiales	X		X	
Lutte génétique	X			X

Source : Thèse Moiroux (2012)

De façon générale, le choix d'une mesure d'intervention dans une région devra tenir compte :

- des caractéristiques entomologiques, épidémiologiques et écologiques (niveaux de sensibilité aux insecticides des populations d'anophèles par exemple) qui déterminent l'efficacité des mesures envisageables,
- de la situation économique et surtout environnementale qui influe sur la durabilité et l'efficacité de l'effort de lutte. S'il est indispensable d'avoir une bonne connaissance de la situation qui prévaut dans une région avant la mise en place des interventions, le choix des indicateurs d'impacts est aussi très important pour le suivi et l'évaluation des opérations de LAV.

1.5 Problématique de santé

L'objectif de la LAV en particulier la lutte contre les vecteurs adultes revient à minimiser, voire éliminer, les contacts hôtes/vecteurs pour réduire, voire stopper la transmission (Pages et al., 2007). Les mesures de contact hôte-vecteur sont donc un précieux indicateur dans l'évaluation des méthodes de LAV dans une région donnée. Plusieurs méthodes permettent d'évaluer le contact hôte-vecteur (Coffinet et al., 2009; Silver, 2008), la plus utilisée est la capture sur appât humain. Les résultats de la capture sur appât humain comme de bien d'autres méthodes peuvent révéler des situations très variables à l'intérieur d'une région aussi modeste soit-elle en superficie. Il est donc indispensable de distinguer ces spécificités à l'intérieur d'une même région afin de rendre les mesures de LAV plus ciblées et permettre aussi une meilleure évaluation de ces mesures. Un outil indispensable pour la LAV aujourd'hui serait donc de distinguer des zones aussi fines que possible en termes i) d'intensité du contact hôte-vecteur, ii) de profil (annuel par exemple) du contact hôte-vecteur.

1.6 Données de capture d'anophèles issues du projet REFS

Les données que nous exploitons dans le cadre du présent travail proviennent d'un projet dénommé REFS «Recherche en Entomologie, Formation et Stratégie». L'un des objectifs du projet était l'évaluation à l'échelle communautaire de quatre stratégies (bras de traitement) de LAV combinant deux traitements insecticides résiduels à l'intérieur des habitations : i) bras «PNLP» = distribution de Moustiquaires Imprégnées d'insecticide à Longue Durée d'action (MILD) de type PermaNet2® à tous les couchages où dorment des enfants âgés de 0 à 5 ans ; ii) bras «MILD» = distribution de MILD à tous les couchages ; iii) bras «PID» = traitement avec du bendiocarb à la dose de 400 mg/m^2 de la face intérieure des murs, des portes et des fenêtres des cases habitées et distribution des MILD de type

PermaNet2® à tous les couchages où dorment des enfants âgés de 0 à 5 ans ; iv) bras «MILD+BI» = distribution des MILD à tous les couchages et mise en place à l'intérieur de toutes les cases habitées (à environ 1/3 supérieur des murs) des bâches plastiques murales imprégnées de bendiocarb (BI) à la dose de 200 mg/m^2.

Dans le but de cette évaluation, des collectes de moustiques ont été effectuées dans un ensemble de trois communes du sud Bénin dans le département de l'atlantique, à l'ouest de Cotonou. Cette zone est connue pour être une zone de faible transmission et de mésoendemie palustre (Djènontin et al., 2010; Damien et al., 2010). Au total 28 villages, à raison de 7 villages par bras de traitement (voir Thèse Djènontin (2011) pour plus de détails), ont été retenus pour la collecte suivant les critères ci-après : population humaine supérieure à 300, nombre d'enfants de moins de 5 ans supérieur à 60, distance minimale de 2 km entre villages choisis, regroupement des habitations dans le village, absence de centre de santé et accessibilité du village. Les captures de moustiques ont été effectuées d'Octobre 2007 à Décembre 2009.

Au total, 17 missions de capture de moustique ont été réalisées dans chacun des 28 villages. Dans chaque village, 4 maisons sont choisies au hasard et dans chaque maison il y a un captureur posté à l'intérieur et un autre à l'extérieur. Ainsi dans chaque village et par mission, il y avait 8 postes de capture. La durée qui sépare 2 missions successives est d'environ 6 semaines. Chaque mission est composée de 2 nuits successives de capture. Les moustiques collectés sont regroupés par heure de capture, chaque nuit de capture allant de 10 heures du soir à 06 heures du matin. A côté des collectes principales que constituent les captures de moustique, plusieurs autres données ont été enregistrées. Il s'agit essentiellement de données environnementales : les différents gîtes naturelles ou artificiels et temporaires ou définitifs, la distance du village à un cours d'eau douce ou stagnante, les coordonnées géographiques des points de capture, la pluviométrie sous différentes formes, la population du village et sa superficie, le nombre et le type d'habitations, la nature de sol, le taux de couverture et d'utilisation des moustiquaires imprégnées d'insecticides à longue durée

d'action,... Un travail d'identification des moustiques au laboratoire a succédé aux collectes et à permis de dégager les vecteurs de paludisme (Djènontin et al., 2010).

1.7 Problématiques méthodologiques

Comme évoqué dans la section 1.5 , une politique de LAV, pour avoir des chances d'être efficace doit tenir compte au mieux des spécificités de chaque zone, de chaque sous-zone. Il en est de même pour tout système d'évaluation de méthodes de LAV. Dans le cas particulier de la lutte contre les vecteurs adultes qui induit une évaluation du niveau de contact homme-vecteur, il est indispensable de se munir d'outils à même de prendre en compte l'importante hétérogénéité à la fois spatiale et temporelle qui caractérise les populations de vecteur. L'apport de l'outil statistique dans l'atteinte de cet objectif devra donc consister à développer des modèles capables à la fois de i) coller au mieux à la vraie distribution des vecteurs ou du contact homme-vecteur au niveau de chaque unité statistique ii) regrouper les unités statistiques en sous-ensembles assez homogènes en termes de niveau moyen (aspect spatial) et/ou de profil (aspect temporel) de la densité de vecteur ou du contact homme/ vecteur. Les modèles de mélange incorporent en leur sein un double objectif qui semble tout à fait approprié au problème qui vient d'être posé : d'une part ils sont un bon outil de modélisation en présence d'hétérogénéité grâce à leur flexibilité et d'autre part ils sont capables de fournir une partition assez satisfaisante des données.

Chapitre 2

VIH/SIDA et problématique de la variabilité de l'observance aux antirétroviraux

2.1 Historique de l'épidémie

Les premiers cas d'infection par le virus de l'immunodéficience humaine (VIH) ont été diagnostiqués au début des années 1980 (CDC, 1981; Korber et al., 2000) bien que certaines études semblent soutenir que les premiers cas d'infections remontent à des décennies plus tôt. L'impact de la maladie, supposée restreinte à des groupes à risque, a été initialement sous-estimé et ce n'est qu'à partir de 1985 que l'humanité prendra conscience de l'évidence de la pandémie du VIH. La décennie qui suit est alors marquée par d'importantes mobilisations des pouvoirs publics pour lutter contre la pandémie avec des mesures comme le dépistage de l'infection, la prévention de la contamination, la mise sur le marché des premiers ARV, la création de structures spécifiques pour faire face au «désastre» (Caraël, 2006; Fee and Parry, 2008; Merson et al., 2008; Plummer et al., 1991). Mais l'absence de traitements réellement efficaces n'a pas permis de freiner de façon remarquable l'impact de la pandémie. A partir de 1996, les traitements antirétroviraux hautement actifs (HAART pour Highly Active Antiretroviral Therapy) ont été disponibles en Europe et en Amérique du Nord modifiant considérablement le pronostic de l'infection chez les patients dans ces régions (Mocroft et al., 2003; Palella Jr et al., 1998). Contrairement aux pays à fort revenu, l'accès au HAART a longtemps été problématique et d'un niveau faible dans les

pays de l'Afrique subsaharienne pour diverses raisons (Kebba, 2003; Yazdanpanah, 2004).

A ce jour, l'organisation Mondiale de la Santé (OMS) estime à plus de 36 millions le nombre de décès liés à l'infection par le VIH jusqu'à ce jour. En 2012, il y avait environ 35,3 [32,2–38,8] millions de personnes vivant avec le VIH dans le monde et 69% de ces personnes vivaient en Afrique subsaharienne (ONUSIDA, 2013). Si l'épidémie semble être maîtrisée dans les pays à fort revenu, son impact en terme humain, social et économique sur le continent africain demeure majeur.

2.2 Cycle du VIH et mode d'infection

Le SIDA ou syndrome d'immunodéficience acquise, est la maladie qu'entraine l'infection par le VIH. Le virus peut se transmettre par le contact étroit et non protégé avec les liquides organiques d'un sujet infecté : sang, lait maternel, sperme et sécrétions vaginales. La gravité du SIDA est liée au fait que le VIH infecte spécifiquement des cellules du système immunitaire garant de l'intégrité biologique. Le VIH est un virus d'une centaine de nanomètres de diamètre, formé de 2 capsides et limité par une enveloppe hérissée de protéines de surface (Figure 2.1). Son programme génétique est constitué d'ARN et d'une enzyme (la transcriptase inverse) qui permet la synthèse d'ADN à partir de l'ARN viral dans la cellule infectée. Il s'en suit alors la production de nombreuses particules virales à l'intérieur de la cellule infectée ; la mort de cette cellule provoquera leur dissémination dans l'organisme.

Les différentes phases de l'infection par le VIH traduisent différents aspects de la réponse immunitaire.

La première phase ou primo-infection fait suite à une contamination par le virus. Les cellules infectées migrent dans certains organes (en particulier les ganglions lymphatiques) qui constituent de véritables réservoirs du virus. Les symptômes sont alors ceux d'une maladie virale bénigne.

La deuxième phase qualifiée de phase asymptotique, se traduit par la mise en place des réponses immunitaires. Des anticorps anti-VIH sont détectés dans le sang du sujet deux semaines à quelques mois après la contamination ; leur présence définit le caractère séropositif du sujet pour le VIH. Des lymphocytes T cytotoxiques (globules blancs particuliers) apparaissent dans le sang du sujet contaminé pour lutter contre les cellules infectées par le VIH. La phase asymptotique oppose une apparente stabilité, à l'échelle de l'organisme, à d'importantes modifications à l'échelle cellulaire (le virus continue à se multiplier, la population de lymphocytes T diminue progressivement).

La troisième phase ou phase symptomatique, est qualifiée de SIDA déclaré. Elle est caractérisée par un déficit immunitaire très fort ou immunodépression. Ceci conduit au développement de maladies opportunistes (pneumonies, diarrhées, dégénérescence du système nerveux, cancer). L'association de ces maladies opportunistes correspond au SIDA.

2.3 Antirétroviraux et problématique de santé

L'apparition de la trithérapie dès la fin des années 90 (premiers essais en Europe et en Amérique du Nord) ont fait naître l'espoir de traitements qui pouvaient changer radicalement le pronostic des patients au stade SIDA (Hammer et al., 1997; Palella Jr et al., 1998; Mocroft et al., 2003). Mais, le coût très élevé en rendait l'accès difficile dans les pays à ressources limitées. Plusieurs actions ont été menées notamment par les associations des personnes vivant avec le VIH (PVVIH) et les organisations non gouvernementales, pour l'accès aux ARV dans les pays à ressources limitées. La revendication pour l'accès aux ARV a pris ainsi à partir de 1999 un caractère très médiatisé qui conduit à l'internationalisation du débat avec à la clé la reconnaissance en 2001 et 2002 par la commission des droits de l'homme des Nations Unies de l'accès aux traitements contre le SIDA comme une composante fondamentale des droits de toute personne de jouir du meilleur état de santé physique et

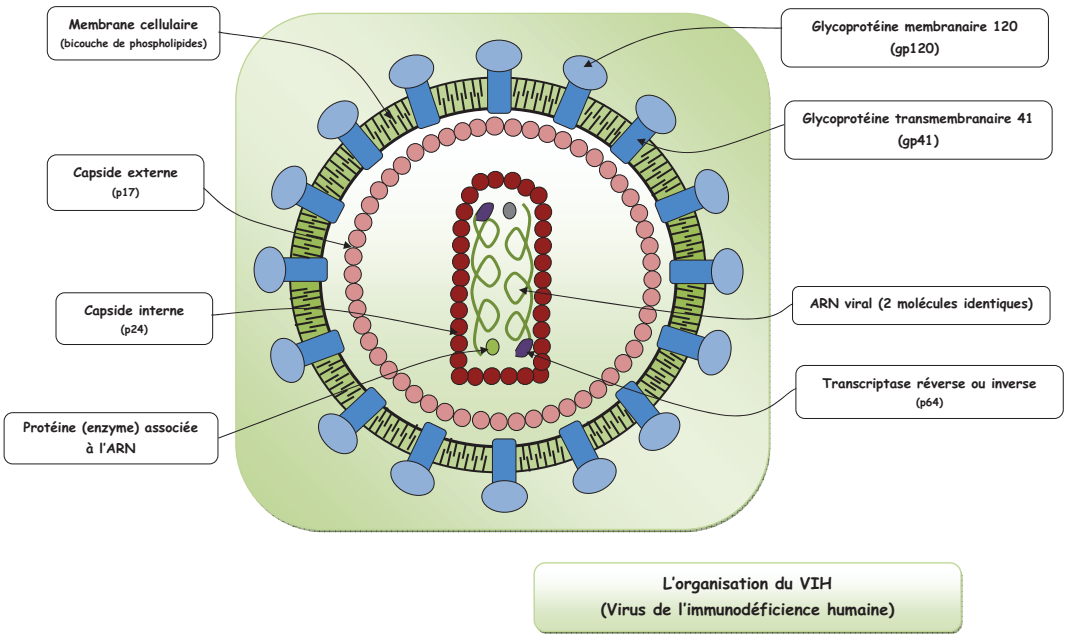


FIGURE 2.1 – Schéma de la structure du VIH (Fabrice Morales ; Banque de schémas – SVT, académie de Dijon)

mental (United Nations 2002). La création de l'ONUSIDA quelques années plus tôt marqua le début d'une période de mise en commun des ressources pour la lutte contre l'épidémie de VIH-1, en particulier dans les pays à ressources limitées qui va se traduire par une importante réduction des coûts des ARV, voire leur gratuité dans ces pays. Si la question de l'accès au traitement ARV semble se poser avec de moins en moins d'acuité, la problématique de l'observance au traitement, elle, reste d'actualité. L'observance dans les traitements de l'infection par le VIH, doit être supérieure à 95%, sous peine d'échappement virologique (Bangsberg et al., 2001; Mills et al., 2006; WHO, 2003; Castro, 2005). C'est une exigence assez sévère qui implique que tout oubli ou refus occasionnels de prise du traitement peut avoir des conséquences sur le contrôle à long terme de la maladie. Or la complexité du traitement, les effets indésirables, les composantes socio-culturelles... sont autant de raisons qui induisent différents profils d'observance chez les patients.

Si plusieurs travaux ont montré le lien entre le niveau d'observance aux ARV et

la suppression de la charge virale dans le plasma, la reconstitution des CD4, la progression de la maladie chez les patients ou encore la mortalité (Bangsberg et al., 2001; Haubrich et al., 1999; Palella Jr et al., 1998; Mannheimer et al., 2002; Nachega et al., 2006; Abaasa et al., 2008), à l'inverse, l'impact de la variabilité de l'observance aux ARV (fluctuations du niveau d'observance au cours du temps) sur la réponse immuno-virologique ne semble pas encore avoir fait l'objet d'étude : une partie du présent travail de thèse tente de combler ce vide.

2.4 Données de travail : Cohorte ANRS 1215

A la faveur des multiples combats pour l'accès aux ARV dans les pays à ressources limitées, la fin des années 90 a vu naître en Afrique subsaharienne quelques projets allant dans ce sens. L'initiative sénégalaise d'accès aux ARV (ISAARV) mise en place dès 1998 par le gouvernement sénégalais est l'un des tous premiers programmes de prise en charge des PVVIH sur le continent. Une des composantes de ce programme était une cohorte financée par l'ANRS (cohorte ANRS 1215) dont l'objectif général était l'évaluation à long terme de l'impact des ARV délivrés au sein de l'ISAARV (Taverne et al., 2012; Desclaux et al., 2003). Les objectifs spécifiques étaient d'évaluer l'efficacité clinique et biologique des traitements, la tolérance clinique et biologique des traitements, l'observance et l'émergence de résistances virales. La cohorte ANRS 1215 a été mise en place entre 1998 et 2002 à partir des 404 premiers patients inclus dans l'ISAARV. Parmi les patients inclus, 80 provenaient de deux essais cliniques (ANRS 1204/IMEA 011 et ANRS 1206/IMEA 012) (Landman et al., 2003). Le tableau 2.1 résume les critères d'inclusion des patients.

Globalement, les patients de la cohorte ANRS 1215 étaient à un stade avancé de la maladie (plus de 50% étaient au stade CDC C), leur âge médian était de 37 ans et les femmes constituaient presque 55% de la cohorte. Chaque patient est revu, dans le même centre de santé (hôpital Fann - Dakar) par intervalle de un à deux mois maximum après l'initiation aux ARV. En Novembre 1999, la mesure de

TABLE 2.1 – Critères d'inclusion des patients

	Critère
Patients hors essai	Stade CDC* C. Stade CDC B et CD4 < 350 cellules/mm^3. Stade CDC A, charge virale $> 10^5$ copies/mL et CD4 < 350 cellules/mm^3.
Essai ANRS 1204/IMEA 011	CD4 entre 50 et 500 cellules/mm^3 et charge virale $> 3 \times 10^4$ copies/mL.
Essai ANRS 1206/IMEA 012	CD4 < 350 cellules/mm^3 et charge virale $> 3 \times 10^4$ copies/mL.

*Centers for Diseases Control

l'observance au traitement a démarré pour les 180 premiers patients de la cohorte puis en Mai 2004 pour les 224 autres. A chaque visite du patient, le pharmacien du centre procède à l'estimation de l'observance en comptant les comprimés ramenés par le patient et recueille par interview à l'aide d'un questionnaire, les éventuelles raisons de non observance. L'observance est obtenue par le rapport du nombre de comprimés pris par le nombre de comprimés prescrits. Par ailleurs, tous les six mois, chaque patient subit des examens biologiques incluant entre autres la mesure du taux de CD4 et celle de la charge virale.

2.5 Problématique méthodologiques

Comme évoqué dans la section 2.3, en fonction des contraintes socio-économiques et en particulier dans le contexte africain, l'observance aux ARV peut être très variable d'un patient à l'autre ou chez le même patient au cours du temps. Cette situation peut résulter aussi des effets indésirables liés aux différentes lignes thérapeutiques étant donné que le niveau de tolérance est variable d'un patient à l'autre. Différents profils d'observance peuvent donc apparaitre au sein d'une population de patients VIH mis sous ARV. Ces profils peuvent être de deux ordres : le niveau moyen d'ob-

servance et la variabilité de l'observance qui peuvent avoir des conséquences sur la réponse immunovirologique chez les patients. Ici aussi, comme pour le contact homme/vecteur dans le cas du paludisme, se pose la nécessité de dégager différents profils d'observance. Sur le plan méthodologique, le problème posé revient donc à proposer un modèle qui dégagerait une classification des patients en termes de niveau moyen et de variabilité de l'observance. Le même modèle doit nécessairement avoir une seconde qualité : celui de coller au mieux dans chaque classe de patients aux observations relatives aux deux caractéristiques de l'observance ici évoquées. Les modèles à classes latentes semblent un outil approprié pour répondre à cet objectif.

Chapitre 3

Modèles à classes latentes

3.1 Une méthode flexible pour la modélisation

Dans le cas des données longitudinales, de données de survie multidimensionnelles ou encore de données provenant de répétitions de mesures expérimentales, on utilise des modèles à effets aléatoires pour prendre en compte la dépendance des observations ou leur appartenance à des groupes donnés. Les dépendances observées et prises en compte dans ces modèles proviennent de variables qui ne peuvent être directement observées, ces variables sont dites «latentes». Dans nombre d'analyses de régression comme les modèles linéaires mixtes et les modèles linéaires généralisés mixtes, on modélise les variations des paramètres d'une unité (statistique) à l'autre en introduisant une ou des variable(s) latente(s) continue(s), supposant alors une distribution multinormale des effets aléatoires. Mais il devient de plus en plus fréquent d'utiliser dans ces genres d'analyses de régression, une ou plusieurs variables latentes discrètes et on obtient une configuration connue sous le nom de modèles à classes latentes ou modèles de mélange (Droesbeke et al. (2013) ; Skrondal and Rabe-Hesketh (2004) ; McLachlan and Peel (2000)). Une telle approche se justifie par trois arguments :

i) L'hypothèse forte d'une distribution multinormale des effets aléatoires n'est pas possible dans certains modèles à effets aléatoires. Les estimations se font alors par la méthode du maximum de vraisemblance non paramétrique dans laquelle on utilise une approximation discrète de cette distribution.

ii) C'est un moyen pratique d'éviter des intégrations par approximation numérique dans le cas où le nombre de dimensions peut être important (situation fréquente avec

les modèles linéaires généralisés mixtes).

iii) C'est un outil permettant le classement des unités (statistiques) en utilisant les différences entre les coefficients. Par exemple, il est possible de classer des villages suivant le niveau moyen du contact homme-vecteur dans le cadre du paludisme ou encore de classer des patients VIH suivant leurs profils d'observance aux ARV.
En concentrant les trois avantages ci-dessus énumérés, les modèles de mélange offrent une remarquable flexibilité et leur utilité dans la modélisation de phénomènes présentant une grande hétérogénéité est indéniable (voir McLachlan and Peel (2000) et ses références). Les modèles de mélange sont devenus aujourd'hui un outil populaire et utilisé avec succès dans un nombre croissant de disciplines incluant l'astronomie, la biologie, la génétique, la médecine, l'économie, la reconnaissance d'images, le marketing...
Nous présentons dans les parties qui suivent quelques fondements théoriques sur les modèles de mélange et nous limitons strictement au cas de mélanges finis. Les meilleures références auxquelles on peut se rapporter sur le sujet, à notre avis, semblent être Skrondal and Rabe-Hesketh (2004) et McLachlan and Peel (2000).

3.2 Présentation et interprétation générative

Notons π_k des quantités telles que $\sum_{k=1}^{K} \pi_k = 1$, avec $0 < \pi_k \leqslant 1$ et f_k, $k = 1, \ldots, K$ des densités de probabilités définies sur un espace χ ; alors :

$$g = \sum_{k=1}^{K} \pi_k f_k \tag{3.1}$$

définit sur χ une densité de mélange fini à K composantes, les f_k. Le poids de la k^{ieme} composante vaut alors π_k. Cette définition générale laisse la possibilité aux f_k d'être de natures différentes.

Mais très souvent, on suppose que les f_k proviennent d'une même famille paramétrique et ne diffèrent que par la valeur α_k d'un paramètre. La densité de mélange fini (3.1)

s'écrit alors :

$$g(.,\theta) = \sum_{k=1}^{K} \pi_k f(.,\alpha_k) \tag{3.2}$$

où les appartiennent à un même espace de paramètres et $\theta = (\pi_1, \ldots, \pi_k, \alpha_1, \ldots, \alpha_k)$ désigne le paramètre du mélange.

L'idée sous-jacente à la construction d'un modèle de mélange revient à considérer une population qui se subdivise en K sous-populations (les composantes) dont les proportions respectives (généralement inconnues) sont les poids π_k. Un échantillon de cette population serait constitué des couples (X_i, Z_i) pour $i = 1, \ldots, n$, où $X_i = x_i$ correspond à la mesure faite sur l'individu i et $Z_i = k$ le numéro de la sous population de laquelle provient i. En supposant qu'on échantillonne uniquement dans la sous-population k et que X est discret, on obtiendrait : $P(X = x|Z = k) = f(x, \alpha_k)$. Le paramètre α_k est généralement inconnu et propre à la sous-population k. Aussi, comme l'échantillonnage est aléatoire, on a $P(Z = k) = \pi_k$ et on déduit que la loi jointe du couple aléatoire (X, Z) est donnée par $P(X = x, Z = k) = f(x, \alpha_k)\pi_k$. En réalité les Z_i ne sont pas observés (d'où la notion de variable latente) ; seul est observé l'échantillon $X_1, \ldots, X_n$ dont la loi est obtenue comme loi marginale de (X, Z) par : $g(x, \theta) = P[X = x] = \sum_{k=1}^{K} \pi_k f(x, \alpha_k)$.

Ainsi, générer une donnée $x \in \chi$ selon l'équation (3.2) revient à choisir l'une des composantes du mélange (par exemple la k^{ieme} avec la probabilité π_k) puis à générer x selon $f(., \alpha_k)$.

De façon inverse, à toute donnée $x \in \chi$ générée selon l'équation (3.2), correspond une réalisation z de la variable Z (variable latente) indiquant la composante dont x est issue, dans le cas où cette composante est connue.

Remarque :

Lorsque le but de la modélisation est de montrer l'évolution au cours du temps d'un phénomène, on fait dépendre les $f(., \alpha_k)$ du temps. Dans ce cas on utilise le vocable de modèle à classes latentes de trajectoires ou simplement de modèle à trajectoires latentes.

3.3 Modèles de mélange et classification

La classification non supervisée, connue aussi sous le nom de classification automatique ou clustering, a pour principal objectif de partitionner un ensemble de n objets $(x_1, \ldots, x_n)$ d'un espace en K classes « homogènes ». D'autres structures peuvent aussi être recherchées par la classification, comme par exemple les hiérarchies, qui sont des emboîtements de partitions.

La partition recherchée peut être représentée sous la forme d'un tableau disjonctif complet $n \times K$ défini par : $z_{ik} = 1$ si l'individu x_i est dans la classe k et $z_{ik} = 0$ sinon.

L'objectif de la classification n'est pas seulement formel. L'idée derrière cette démarche est souvent d'aider le praticien à analyser des données. Le regroupement en classes est pour lui une façon de synthétiser pour isoler l'information pertinente qu'il est difficile d'appréhender directement en présence de données parfois nombreuses, éventuellement décrites par de multiples dimensions dans des espaces eux-mêmes un peu complexes. Mieux, la classification peut fournir au praticien une aide supplémentaire permettant de faciliter l'interprétation des partitions elles-mêmes. En effet, le calcul de statistiques comme par exemple une moyenne par classe sur la partition obtenue permettra de dégager la classe « des grands », la classe « des modérés », la classe « des petits »...

3.3.1 Approches probabilistes de la classification

La recherche de partition en classification conduit souvent à des algorithmes conçus d'un point de vue heuristique et utilisant des critères métriques. En effet, les deux algorithmes les plus utilisés en classification sont basés sur l'inertie inter-classe d'une partition, c'est-à-dire la somme des inerties de chaque classe : il s'agit de l'algorithme des k-means (centres-mobiles) pour la recherche de partitions et l'algorithme de classification ascendante hiérarchique de Ward pour la recherche de hiérarchies. Cette approche peine souvent à se justifier en ce qui concerne le choix de la métrique et le critère utilisés : la métrique doit mesurer la dissimilarité entre les objets de l'ensemble à classifier tandis que le critère défini à partir de cette métrique doit mesurer le degré de cohésion et de séparation des classes. Pour contourner cette difficulté, une approche plus statistique, qui utilise des modèles probabilistes de classification pour formaliser l'idée intuitive de la notion de classe naturelle, a fait son chemin depuis quelques années. Elle offre un cadre d'interprétation statistique à certains critères métriques dont les différentes variantes n'étaient pas toujours bien claires.

Très souvent, la classification est faite sur un échantillon et les conclusions obtenues sont étendues à la population dont est issu l'échantillon. Dans ce cas, la classification n'a pas de sens sans un recours à un modèle probabiliste permettant de justifier cette inférence. Toute approche probabiliste de la classification non supervisée fait l'hypothèse que les données constituent un échantillon aléatoire d'une population et s'appuie sur l'analyse de la distribution de probabilités de cette population pour définir une classification. Nous présentons très brièvement ici, deux approches probabilistes de la classification.

Approches non paramétriques

Elles ne supposent aucune hypothèse sur la distribution de probabilités et s'appuient toutes sur la forme de cette distribution. Dans le cas de données continues par

exemple, cette distribution est caractérisée par sa fonction de densité qui est alors utilisée pour définir la notion de classes. Hartigan (1975) distingue ainsi une classe de forte densité comme un sous ensemble connexe de points de densité supérieure à un certains seuil et obtient un arbre hiérarchique de classes en faisant varier ce seuil. La présence de données hétérogènes et donc de classes peut s'illustrer par la présence de plusieurs maxima de la densité. Des classes modales peuvent ainsi être définies par la recherche de ces maxima et l'affectation des points de l'espace de référence à chacun d'entre eux. L'estimation de la distribution inconnue à partir des données est nécessaire à l'application de ces méthodes dont les plus courantes s'appuient sur une estimation non paramétrique de la densité (méthode des plus proches voisins, méthode des noyaux, utilisation de l'histogramme). Cette démarche a donné lieu à de nombreux algorithmes parmi lesquels on pourrait ranger les algorithmes de classification hiérarchique de Lerman (1981) définis à partir de la notion de vraisemblance du lien.

Approches paramétriques

Elles font des hypothèses sur la distribution de probabilité induisant alors une classification et formalisant ainsi la notion de classes (naturelles). Le modèle de mélange se range dans cette catégorie et est sans nul doute le modèle paramétrique le plus utilisé en classification. D'autres approches paramétriques de la classification automatique existent, comme par exemple les modèles fonctionnels à effet fixe et les processus ponctuels.

Les modèles fonctionnels à effet fixe se caractérisent par l'équation :

$$\text{Données} = \text{Structure} + \text{Erreur},$$

où la structure est inconnue mais fixe et l'erreur est aléatoire. Ces modèles peuvent être appliqués aisément à la classification en choisissant une structure adéquate.

L'exemple le plus simple de ces modèles consiste à supposer que les données sont des vecteurs $x_1, \ldots, x_n$ de $\mathbb{R}^d$ et de considérer le modèle $x_i = y_i + \varepsilon_i$, dans lequel on impose aux y_i d'appartenir à un ensemble de K centres $m_1, \ldots, m_K$ et aux erreurs ε_i de suivre une loi centrée de même variance. (Bollen, 1989) semble être une bonne référence traitant des modèles fonctionnels.

Les processus ponctuels sont des modèles utilisés en statistique spatiale pour des données qui peuvent être, par exemple, la répartition des arbres dans une forêt ou des anophèles dans un village. Certains de ces processus correspondent à une organisation en agrégats et peuvent être considérés comme des modèles probabilistes associés à une classification. Le plus utilisé est le processus de Neyman-Scott (Neyman and Scott, 1958) qui peut être interprété comme une génération des données en trois étapes :
Etape 1- K points $m_1, \ldots, m_K$ sont tirés au hasard suivant une distribution uniforme sur une région convexe ;
Etape 2- les tailles $n_1, \ldots, n_K$ des classes sont tirées au hasard, par exemple à l'aide d'une distribution de Poisson ;
Etape 3- Pour chaque classe k, n_k points sont tirés au hasard en utilisant une distribution sphérique centrée en m_k, par exemple une distribution gaussienne de moyenne m_k.

3.3.2 Modèles de mélange pour la classification

L'utilisation des modèles de mélange à des fins de classification automatique se rencontre dans des domaines très variés car cette approche correspond souvent à l'idée intuitive que l'on peut se faire d'une population composée de plusieurs classes. En outre, elle possède d'importantes similitudes avec des méthodes de référence comme l'algorithme des k-means et grâce à sa flexibilité, cette approche permet d'intégrer assez naturellement nombre de situations particulières. Pour obtenir une

partition des données initiales à partir d'un modèle de mélange dont la densité est spécifiée par la formule (3.2), on peut utiliser l'approche ML (Maximum Likelihood) ou l'approche CML (Classification Maximum Likelihood). Il s'agira donc dans les deux approches d'estimer le paramètre $\theta = (\pi_1, \ldots, \pi_k, \alpha_1, \ldots, \alpha_k)$.

Approche ML

Le principe dans cette approche de la classification automatique à l'aide du modèle de mélange consiste à d'abord estimer le paramètre du modèle par la méthode du maximum de vraisemblance puis à déterminer une partition en rangeant chaque individu dans la classe la plus probable conditionnellement à cette estimation. En considérant un échantillon $x = (x_1, \ldots, x_n)$, la vraisemblance du mélange s'écrit

$$L(x, \theta) = \prod_{i=1}^{n} g(x_i, \theta) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f(x_i, \alpha_k) \tag{3.3}$$

et on déduit $l(\theta)$ la log-vraisemblance associée :

$$l(\theta) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k f(x_i, \alpha_k) \right) \tag{3.4}$$

La maximisation de cette log-vraisemblance peut alors être effectuée via l'algorithme EM qui s'appuie sur la notion de données complètes. Pour le modèle de mélange, ces données complètes sont obtenues en complétant l'échantillon $x = (x_1, \ldots, x_n)$ par les composants d'origine $z = (z_1, \ldots, z_n)$ de chaque individu de l'échantillon avec $z_i \in 1, \ldots, K$ représentant le numéro du composant de l'individu i. Toutefois, la partition définie par z sera représentée par la matrice de classification $(z_{ik})_{i=1:n,k=1:K}$ avec z_{ik} égal à 1 si l'individu i appartient à la classe k et 0 sinon. La vraisemblance des données complètes est alors obtenue comme suit :

$$L(x, z, \theta) = \prod_{i=1}^{n} \prod_{k=1}^{K} [\pi_k f(x_i, \alpha_k)]^{z_{ik}} \tag{3.5}$$

D'où l'écriture suivante de la log-vraisemblance complète :

$$l_c(z,\theta) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log(\pi_k f(x_i,\alpha_k)) \tag{3.6}$$

L'algorithme EM consiste alors à maximiser par itération la fonction :

$$\begin{aligned} Q(\theta,\theta') &= \mathbb{E}(l_c(\theta,z)|x,\theta') \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{E}(z_{ik}|x,\theta')\log(\pi_k f(x_i,\alpha_k)) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} t_{ik}\log(\pi'_k f(x_i,\alpha'_k)) \end{aligned} \tag{3.7}$$

qui correspond à la log-vraisemblance attendue de θ'. Dans cette écriture, t_{ik} désigne la probabilité *a posteriori* d'appartenance de l'individu i à la classe k lorsque la valeur du paramètre est θ' :

$$\begin{aligned} t_{ik} = \mathbb{E}(z_{ik}|x,\theta') &= P(z_{ik}=1|x,\theta') \\ &= \frac{\pi'_k f(x_i,\alpha'_k)}{\sum_{q=1}^{K}\pi'_q f(x_i,\alpha'_q)} \end{aligned} \tag{3.8}$$

Les étapes de l'algorithme EM pour l'estimation de θ sont alors les suivantes :

- initialisation : choix arbitraire d'une solution initiale $\theta^{(0)}$;
- étape E (espérance) : calcul des probabilités d'appartenance des x_i aux classes conditionnellement au paramètre courant :

$$t_{ik}^{(c)} = \frac{\pi_k^{(c)} f(x_i,\alpha_k^{(c)})}{\sum_{q=1}^{K}\pi_q^{(c)} f(x_i,\alpha_q^{(c)})} \tag{3.9}$$

- étape M (Maximisation) : la vraisemblance est maximisée conditionnellement aux t_{ik} ; les proportions des différentes classes sont alors obtenues par la relation :

$$\pi_k^{(c+1)} = \frac{1}{n}\sum_{i=1}^{n} t_{ik}^{(c)} \tag{3.10}$$

tandis que les paramètres $\alpha_k^{(c+1)}$ sont obtenus par résolution des équations de vraisemblance qui dépendent du modèle de mélange retenu. Les étapes E et M sont répétées jusqu'à la convergence de l'algorithme.
A la convergence de l'algorithme EM, une partition est alors déduite par la méthode du maximum a posteriori (ou méthode du MAP) : selon cette méthode, chaque individu est rangé dans la classe maximisant la probabilité *a posteriori* t_{ik} d'appartenance de l'individu i à la classe k calculée à partir des paramètres estimés.

Approche CML

La partition obtenue dans l'approche ML apparaît comme un résultat déduit de l'estimation du paramètre de la loi de mélange. L'approche CML (Symons, 1981) estime simultanément la partition z et le paramètre θ. Dans cette approche, l'idée revient à maximiser en θ et z la log-vraisemblance complète, habituellement appelée vraisemblance classifiante et qui s'écrit comme suit :

$$l_C(z,\theta) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log(\pi_k f(x_i,\alpha_k)) \tag{3.11}$$

On recherche alors une partition de l'échantillon de telle sorte que chaque classe k soit identifiable à un sous-échantillon issu de la loi $f(x_i,\alpha_k)$.
Plusieurs propositions ont été faites pour l'introduction de la partition z dans le critère de vraisemblance. Ainsi, Scott and Symons (1971) définissent un critère dans lequel les proportions π_k n'apparaissent pas, une sorte de vraisemblance classifiante restreinte dont l'écriture est la suivante :

$$l_{CR}(z,\theta) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log f(x_i,\alpha_k) \tag{3.12}$$

Mais ce critère a tendance à donner des classes de mêmes proportions. Symons (1981) le modifiera pour proposer finalement une log-vraisemblance des données

complétées qui intègre un terme de pénalité dans le critère précédent comme suit :

$$l_C(z,\theta) = l_{CR}(z,\theta) + \sum_{k=1}^{K} n_k \log \pi_k \tag{3.13}$$

avec n_k le cardinal de la classe k et $\sum_{k=1}^{K} n_k \log \pi_k$ le terme de pénalité qui disparaît si on impose aux proportions d'être toutes identiques.

L'algorithme CEM (Classification EM) introduit par Celeux and Govaert (1992) permet de maximiser la vraisemblance classifiante. C'est une version classifiante de l'algorithme EM obtenue en incorporant entre les étapes E et M, une étape C de classification qui permet de transformer les probabilités conditionnelles t_{ik} en une classification discrète. Ainsi l'algorithme CEM est défini de la manière suivante :

- initialisation : choix arbitraire d'une solution initiale $\theta^{(0)}$;
- étape E : calcul des probabilités conditionnelles $t_{ik}^{(c)}$ comme dans l'algorithme EM ;
- étape C : la partition $z^{(c+1)}$ est obtenue en appliquant la méthode du MAP qui consiste à ranger chaque x_i dans la classe maximisant $t_{ik}^{(c)}$, c'est-à-dire que :

$$z_{ik}^{(c+1)} = \begin{cases} 1 & \text{si } k = \arg\max_{q=1:K} t_{iq}^{(c)} \\ 0 & \text{sinon} \end{cases} \tag{3.14}$$

- étape M : maximisation de la vraisemblance conditionnellement aux $z_{ik}^{(c+1)}$: on obtient les estimations du maximum de vraisemblance des π_k et des α_k en utilisant les classes de la partition $z^{(c+1)}$ comme sous-échantillons. Les proportions sont alors obtenues par la formule $\pi_k^{(c+1)} = \frac{n_k^{(c+1)}}{n}$ avec $n_k^{(c+1)}$ désignant le cardinal de la k^{eme} classe de $z^{(c+1)}$; le calcul des $\alpha_k^{(c+1)}$ est fonction du modèle de mélange retenu.

Remarque

Généralement, il est préférable d'utiliser l'approche mélange (ML) que l'approche classification (CML). En effet, en déterminant à chaque itération les paramètres à partir d'échantillons tronqués du modèle de mélange, l'approche classification fournit une estimation biaisée et non convergente en raison du paramètre à estimer qui croît rapidement avec la taille de l'échantillon. Toutefois l'approche classifiante peut fournir de meilleurs résultats dans les cas où les classes sont bien séparées avec des effectifs faibles.

3.3.3 Formule d'Hathaway

Dans les méthodes dites de classification floue (on parle aussi de classification probabiliste), l'appartenance, vraie ou fausse, d'un sujet à une classe est remplacée par un degré d'appartenance. Sur le plan formel, une classification floue sera donc caractérisée par une matrice $p = (p_{ik})_{i=1:n,k=1:K}$ telle que :

$$\text{a.) } p_{ik} \in [0,1] \qquad \text{b.) } \sum_{k=1}^{K} p_{ik} = 1 \qquad \text{c.) } \sum_{i=1}^{n} p_{ik} \neq 0 \tag{3.15}$$

Une mesure de la séparation des classes probabilistes déterminées par la classification floue est son entropie définie par :

$$\text{Ent}(p) = -\sum_{i=1}^{n}\sum_{k=1}^{K} p_{ik} \log p_{ik} \tag{3.16}$$

On a $0 \leqslant \text{Ent}(p) \leqslant n \log K$ et on montre aisément que lorsque $\text{Ent}(p)$ est proche de 0, les classes probabilistes déterminées par p sont séparées tandis qu'elles sont mélangées lorsque $\text{Ent}(p)$ est proche de $n \log K$.

Hathaway (1986) donne, dans le cas particulier des mélanges, une interprétation de l'algorithme EM qui consiste à considérer EM comme un algorithme optimisant un critère qui porte à la fois sur le paramètre du mélange et une classification floue des données.

Dans un contexte où l'échantillon $x = (x_1, \ldots, x_n)$ est supposé issu d'un mélange paramétrique comme spécifié à l'équation (3.2), la méthode du MAP permet de faire correspondre à toute valeur θ du paramètre, une partition floue (voir 3.8) qui est un objet de même nature que p.

Hathaway propose dans son interprétation de l'algorithme EM, de dissocier la partition floue p de la donnée et la partition t relative au paramètre θ. Cette démarche repose sur deux quantités :

la log-vraisemblance d'une valeur θ du paramètre, complétée de la partition p de la donnée,

$$l_C(p, \theta) = \sum_{i=1}^{n} \sum_{k=1}^{K} p_{ik} \log(\pi_k f(x_i, \alpha_k)) \tag{3.17}$$

et le critère

$$C(p, \theta) = l_C(p, \theta) + \mathrm{Ent}(p) \tag{3.18}$$

La quantité $l_C(p, \theta)$ coïncide avec la log-vraisemblance complétée de θ (voir 3.6) si p est la vraie partition $(z_{ik})_{i=1:n,k=1:K}$ de la donnée et la quantité $l_C(p, \theta)$ correspond à la log-vraisemblance attendue de θ' (voir (3.7)) si p est la partition floue $(t_{ik})_{i=1:n,k=1:K}$ induite par la valeur θ du paramètre.

On peut aisément montrer que la quantité $C(p, \theta)$ peut être réécrite sous la forme

$$C(p, \theta) = l(\theta) - KL(p, t) \tag{3.19}$$

où $l(\theta)$ est la log-vraisemblance du paramètre θ comme définie dans (3.4) et

$$KL(p, t) = \sum_{i=1}^{n} \sum_{k=1}^{K} p_{ik} \log\left(\frac{p_{ik}}{t_{ik}}\right) \tag{3.20}$$

désigne l'écart dit de Kullback entre les deux partitions probabilistes p et t. Ainsi, l'étape M d'EM reviendrait à optimiser $C(t, \theta')$ par rapport à θ' (t étant la partition floue induite par la valeur courante θ du paramètre). L'étape E d'EM détermine

suivant (3.19), la partition p la plus proche au sens de (3.20), de la partition t induite par la valeur courante θ du paramètre.
La formule d'Hathaway montre donc que l'algorithme EM consiste de façon alternative
a) à déterminer la partition floue p la plus proche au sens de (3.20), de celle induite par la valeur θ du paramètre et
b) à augmenter la vraisemblance du paramètre, complétée de cette partition floue.

3.4 Choix du modèle et du nombre de classes

Modéliser des données hétérogènes par un mélange paramétrique doit pouvoir atteindre à la fois deux objectifs : l'adéquation du modèle retenu aux données et l'obtention d'une partition acceptable des données c'est-à-dire qui distingue au mieux les différents sous-groupes en présence au sein de l'échantillon. Dans la pratique, les poursuites des deux objectifs sont assez liées. Le problème posé revient à sélectionner une solution optimale parmi une famille $\mathcal{M}$ de modèles sélectionnés et pour différentes valeurs du nombre de classes K. Il s'agit donc de trouver des critères qui permettent de dégager le meilleur couple (M, K) pour $M \in \mathcal{M}$ et $1 \leqslant K \leqslant K_{max}$, avec $K_{max} = n^{0.3}$ (voir Bozdogan (1993), n étant la taille de l'échantillon.

3.4.1 Critère AIC

Akaïke (1974) propose le critère AIC (Akaike Information Criterion), qui consiste à minimiser l'écart de Kullback moyen entre le vrai modèle et le modèle postulé. Il s'agit plus précisément de trouver le modèle M, K minimisant l'espérance suivante :

$$\mathbb{E}_{x,x'}\left[\log g\left(x'|\hat{\theta}\right)\right] \tag{3.21}$$

où $x = (x_1, \ldots, x_n)$ est un échantillon iid issu de la vraie distribution g inconnue, x' est un vecteur aléatoire issu de g et indépendant des x_i précédents et $\hat{\theta}$ est l'estima-

teur du maximum de vraisemblance de θ obtenu avec l'échantillon x.
En notant $l_{M,K} = \log g(x|M, K, \hat{\theta})$ la log-vraisemblance maximum pour M et K et $\nu_{M,K}$ le nombre de paramètres libres du modèle M, K ; on obtient comme approximation de l'espérance précédente, le critère AIC d'Akaike comme suit :

$$AIC(M) = l_{M,K} - \nu_{M,K} \tag{3.22}$$

Ce critère n'est valide que si on suppose que les données son issues d'une distribution appartenant à l'ensemble des modèles en compétition. Il est constitué d'un premier terme de vraisemblance qui est favorable au choix de modèles complexes et d'un second terme égal au nombre de paramètres du modèle qui pénalise les modèles trop complexes : ce critère peut donc être vu comme un critère de vraisemblance pénalisé.

3.4.2 Critère BIC

Une approche classique pour choisir un modèle dans le cadre bayésien est de sélectionner celui qui maximise la vraisemblance intégrée

$$(\hat{M}, \hat{K}) = \arg\max_{M,K} g(x|M, K) \tag{3.23}$$

En notant $\theta_{M,K}$ l'espace du paramètre du modèle M, K et $\pi(\theta|M, K)$ une loi *a priori* non informative sur le paramètre θ de ce modèle, la vraisemblance intégrée s'écrit :

$$g(x|M, K) = \prod_{i=1}^{n} g(x_i|M, K, \theta) \tag{3.24}$$

Sur la base d'une approximation asymptotique de cette vraisemblance intégrée (valide sous certaines conditions de régularité), Schwarz (1978) propose le critère

BIC (Bayesian Information Criterion) :

$$BIC(M,K) = -l_{M,K} + \frac{\nu_{M,K}}{2}\log n \tag{3.25}$$

où $l_{M,K}$ et $\nu_{M,K}$ sont définies comme dans (3.22) et n est la taille de l'échantillon. Le Critère BIC apparaît ainsi comme un critère de vraisemblance pénalisée. En termes de critère de choix de modèle, BIC a un bon comportement dans le cas où les modèles de mélange sont vus comme un outil semi-paramétrique d'estimation de densité comme dans le cas où ces modèles sont considérés comme un outil dédié à la classification non supervisée. Pour le choix du nombre de composants, dans le contexte d'estimation de densité, BIC a également un bon comportement. En outre, Il a été prouvé par Keribin (2000) que le critère BIC est consistant pour les modèles de mélange.

Cependant, ce critère ne prend pas en compte directement l'aspect classification et peut souvent avoir tendance à surestimer K sans tenir compte de la séparation des classes. Le critère ICL (prochain paragraphe) vient à bout de cette limite de BIC en offrant l'avantage de choisir le nombre de classes K de façon à obtenir une partition des données avec des classes suffisamment séparées.

3.4.3 Critère ICL

Le critère ICL (Integrated Completed Likelihood) proposé par Biernacki et al. (2000) intègre un objectif de classification (ce que ne fait pas BIC) sans perdre de vue la modélisation des données. Il repose sur une approximation non plus de la vraisemblance intégrée du paramètre mais de sa vraisemblance complétée et intégrée comme suit :

$$g(x,z|M,K) = \int_{\theta_{M,K}} g(x,z|M,K,\theta)\pi(\theta|M,K)d\theta \tag{3.26}$$

où $z = (z_{ik})_{i=1:n,k=1:K}$ désigne une partition déterministe de x.

Par analogie à l'élaboration de BIC, la log-vraisemblance intégrée complétée peut être approchée par l'opposé de la quantité suivante Biernacki et al. (2000) :

$$ICL(M,K) = -\log g(x|M,K,\hat{\theta}) + \frac{\nu_{M,K}}{2}\log n - \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{z_{ik}} \log t_{ik} \tag{3.27}$$

où $\hat{\theta}$ désigne l'estimateur du maximum de vraisemblance de θ obtenu avec l'échantillon x , $\nu_{M,K}$ le nombre de paramètres libres du modèle M, K et $\hat{z}_{ik}$ l'estimation de z_{ik}obtenue par MAP. Le modèle qui minimise (3.20) est alors considéré comme le meilleur des modèles. On peut remarquer aisément d'après (3.27) que le critère ICL apparaît sous la forme d'un critère pénalisé par un terme d'entropie :

$$ICL(M,K) = BIC(M,K) + \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{z}_{ik} \log t_{ik} \tag{3.28}$$

Une version du critère ICL qui a pour objectif d'augmenter la pénalisation imposée aux composantes qui se recouvrent consiste à remplacer dans (3.28), $\hat{z}_{ik}$ par t_{ik}.

Deuxième partie

Traitement des données

Chapitre 4

Répartition spatiale du contact homme-vecteur dans le cadre du paludisme

4.1 Hétérogénéité spatiale du contact homme-vecteur

Dans ce chapitre, nous nous intéressons à la modélisation de la distribution spatiale de vecteurs de paludisme dans un contexte de présence de méthodes de LAV. Cette modélisation porte sur des comptes d'anophèles issus des données présentées dans la section 6 du chapitre 1 et a fait l'objet d'une publication.

De façon concrète dans ce travail, nous nous intéressons à l'hétérogénéité spatiale du contact homme-vecteur dans une zone constituée de 28 villages choisis selon des critères présentés à la section 1.6. Le contact homme-vecteur est mesuré à travers des comptes de vecteurs réalisés par la méthode de capture sur appât humain. Les vecteurs considérés sont *An. funestus*, *An. gambiae* et *An. coluzzii*, connus pour être les vecteurs qui assurent la quasi-totalité de la transmission du paludisme dans la zone (Djènontin et al., 2010). La période de recueil considérée dans ce travail correspond à une période d'après mise en place de 4 méthodes de luttes anti-vectorielles et s'étale sur une année calendaire (de Janvier à Décembre 2009).

L'étude de la relation moyenne – variance a révélé une nette surdispersion (variance > moyenne) des comptes d'anophèles excluant la possibilité d'approcher la dis-

tribution des anophèles par une distribution poissonienne qui elle, postule l'égalité des deux statistiques (moyenne et variance).

Il s'est révélé que l'une des causes de cette surdispersion est l' «excès» de zéro dans les données, comparativement à la proportion de zéro attendue d'un point de vue poissonien. Le modèle dit «Zero Inflated Poisson» (ZIP) (Johnson and Kotz, 1969; Lambert, 1992) est l'une des méthodes développées pour gérer la surdispersion due à une pléthore de zéro. Le ZIP n'est en fait qu'un cas particulier des modèles de mélange. C'est une combinaison d'une distribution de Poisson et d'une masse de Dirac en zéro comme on peut le constater aisément dans sa fonction densité donnée par :

$$f(y, \lambda, \pi) = \pi \mathbf{I}_{\{0\}}(y) + (1 - \pi) f_p(y, \lambda) \tag{4.1}$$

où, $\mathbf{I}_{\{0\}}$ est la masse de Dirac en zéro, $f_p(y, \lambda) = \frac{\lambda^y}{y!} \exp(-\lambda)$ est la fonction densité d'une distribution de Poisson de moyenne λ et π et $(1 - \pi)$ sont les proportions du mélange.

Bien que le ZIP a permis de bien prendre en compte la surabondance de zéro dans les comptes d'anophèles, il ne parvient pas a gérer complètement la surdispersion donc l'hétérogénéité présente au niveau de ces données, d'où la nécessité d'aller au-delà d'un mélange à deux distributions.

Une autre façon de prendre en compte cette hétérogénéité est de considérer la distribution des comptes d'anophèles comme une distribution à deux niveaux : le paramètre de la distribution des observations (premier niveau) est supposé avoir lui-même une certaine distribution (deuxième niveau). C'est le cas de la distribution binomiale négative (NB) qui suppose que les observations sont issues d'un mélange continu de Poisson dont les moyennes sont des réalisations d'une loi Gamma. Cette approche s'est révélée suffisamment efficace pour prendre en compte l'hétérogénéité dans les comptes d'anophèles et produit des prédictions qui collent de façon assez satisfaisante aux distributions des observations.

Seulement l'hypothèse d'un mélange continu de lois de Poisson faite dans la NB semble assez forte et pas appropriée au contexte de notre étude. En effet, il est question de définir des sous-ensembles de villages assez homogènes du point de vue du niveau moyen de contact homme-vecteur. Ceci devra permettre à terme de comparer les efficacités des 4 méthodes de lutte implémentées dans l'ensemble des 28 villages.

Nous considérons alors des mélanges de plus de deux distributions de Poisson tout en ne faisant pas d'hypothèse particulière sur la distribution des paramètres des différentes distributions de Poisson participant au mélange. Ces paramètres sont alors estimés par une approche non paramétrique du maximum de vraisemblance (Aitkin, 1996) à travers l'algorithme EM. Nous retrouvons ainsi les modèles de mélange non paramétriques ou modèles à classes latentes dont le cadre théorique a fait l'objet du chapitre 3.

Le critère ICL a révélé qu'un mélange à quatre composants semble meilleur en termes du compromis d'une bonne adéquation du modèle aux données et d'un partitionnement satisfaisant des villages. Ce modèle a donc tenu compte d'un paramètre aléatoire spécifique de chaque village. Des covariables environnementales et descriptives des villages sont aussi prises en compte dans le modèle et leurs effets sur le niveau moyen du contact homme-vecteur ont été établis.

La méthode du MAP a ensuite permis de regrouper les villages en quatre classes qui peuvent être ordonnées en fonction des coefficients aléatoires fournis par le modèle. Ces coefficients peuvent être vus comme un «proxi» de la partie non expliquée de la distribution du contact homme-vecteur, par les covariables prises en compte dans le modèle. Les méthodes de LAV en présence dans les villages n'ont pas été intégrées comme covariables dans le modèle et nous avons fait l'hypothèse qu'elles pourraient être à la base de l'hétérogénéité traduites par les coefficients aléatoires que nous pouvons identifier aux différentes classes. Nous avons donc testé le lien entre le regroupement des villages par méthodes de LAV et la classification des villages obtenue du modèle à quatre classes latentes. Ce lien s'est révélé non

significatif conduisant à l'un des plus importants résultats de notre travail à savoir que parmi les quatre méthodes de LAV en présence dans la zone, il n'y a pas de méthode de lutte significativement plus efficace que les autres en termes d'impact sur l'intensité moyenne du contact homme-vecteur.

Les détails de cette première partie de notre travail sont présentés sous forme d'article dans la section qui suit.

4.2 Article1 : Use of a mixture statistical model in studying malaria vectors density

OPEN ACCESS Freely available online

PLOS ONE

Use of a Mixture Statistical Model in Studying Malaria Vectors Density

Olayidé Boussari[1,2,3,4,5]*, Nicolas Moiroux[6,7,8], Jean Iwaz[2,3,4,5], Armel Djènontin[9,10], Sahabi Bio-Bangana[10], Vincent Corbel[7,8,9], Noël Fonton[1], René Ecochard[2,3,4,5]

1 International Chair in Mathematical Physics and Applications, Université d'Abomey-Calavi, Abomey-Calavi, Bénin, **2** Service de Biostatistique, Hospices Civils de Lyon, Lyon, France, **3** Université de Lyon, Lyon, France, **4** Université Lyon 1, Villeurbanne, France, **5** Laboratoire de Biométrie et Biologie Evolutive, Centre National de la Recherche Scientifique – Unité Mixte de Recherche 5558, Villeurbanne, France, **6** Maladies Infectieuses et Vecteurs Écologie, Génétique, Évolution et Contrôle, Institut de Recherche pour le Développement, Montpellier, France, **7** Université Montpellier 1, Montpellier, France, **8** Université Montpellier 2, Montpellier, France, **9** Maladies Infectieuses et Vecteurs Écologie, Génétique, Évolution et Contrôle, Institut de Recherche pour le Développement, Cotonou, Bénin, **10** Centre de Recherche en Entomologie de Cotonou, Ministère de la Santé, Cotonou, Bénin

Abstract

Vector control is a major step in the process of malaria control and elimination. This requires vector counts and appropriate statistical analyses of these counts. However, vector counts are often overdispersed. A non-parametric mixture of Poisson model (NPMP) is proposed to allow for overdispersion and better describe vector distribution. Mosquito collections using the Human Landing Catches as well as collection of environmental and climatic data were carried out from January to December 2009 in 28 villages in Southern Benin. A NPMP regression model with "village" as random effect is used to test statistical correlations between malaria vectors density and environmental and climatic factors. Furthermore, the villages were ranked using the latent classes derived from the NPMP model. Based on this classification of the villages, the impacts of four vector control strategies implemented in the villages were compared. Vector counts were highly variable and overdispersed with important proportion of zeros (75%). The NPMP model had a good aptitude to predict the observed values and showed that: i) proximity to freshwater body, market gardening, and high levels of rain were associated with high vector density; ii) water conveyance, cattle breeding, vegetation index were associated with low vector density. The 28 villages could then be ranked according to the mean vector number as estimated by the random part of the model after adjustment on all covariates. The NPMP model made it possible to describe the distribution of the vector across the study area. The villages were ranked according to the mean vector density after taking into account the most important covariates. This study demonstrates the necessity and possibility of adapting methods of vector counting and sampling to each setting.

Citation: Boussari O, Moiroux N, Iwaz J, Djènontin A, Bio-Bangana S, et al. (2012) Use of a Mixture Statistical Model in Studying Malaria Vectors Density. PLoS ONE 7(11): e50452. doi:10.1371/journal.pone.0050452

Editor: Luciano A. Moreira, Centro de Pesquisas René Rachou, Brazil

Received August 7, 2012; **Accepted** October 22, 2012; **Published** November 21, 2012

Copyright: © 2012 Boussari et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The Ministère Français des Affaires Etrangères (MAEE) supported the project FSP/REFS N°2006-22 that generated the data analyzed in this paper. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: olayide.boussari@chu-lyon.fr

Introduction

Malaria is still a major public health issue in Sub-Saharan Africa. In 2010, this region bore 91% of the global disease death burden estimated to 655,000 deaths [1]. Studying the risk of vector transmission is at the basis of every survey about the importance of malaria in a given zone. The overarching goal of vector control is to decrease the transmission of the malaria parasite *Plasmodium spp* to humans by mosquito vectors of the genus *Anopheles*. Among the recommendation of the World Health Organization (WHO) to fight malaria, the deployment of long-lasting insecticidal mosquito nets (LLIN) and indoor residual spraying (IRS) at national scale has shown important reductions of malaria burden although evidences of malaria resurgence have been recorded in several African countries [1,2].

The most common indicator to evaluate vector control interventions such as LLIN and IRS relies on malaria transmission through estimation of the Entomological Inoculation Rate (EIR). EIR is the product of the Human Biting Rate (HBR; number of bites of malaria vectors per human per unit time) and the prevalence of *Plasmodium* infection in mosquitoes. HBR is usually measured using the Human Landing Catches (HLC) counting technique that is the method of reference to quantify the human-vector contact [3].

In 28 villages of Southern Benin, a recent cluster randomized controlled trial (RCT) aiming at comparing the efficacy of combined LLIN and carbamate IRS or carbamate-treated plastic sheeting (CTPS) with a background of LLIN coverage did not show benefits of the combination for reducing HBR and EIR [4]. In the study area, high variations in the density of malaria vectors were observed in time and space [5] and there were many localities with zero mosquitoes collected during several nights.

The most ancient and popular statistical distribution used to describe count data is the Poisson distribution that assumes equidispersion of the counts. However, in real datasets, these counts are often overdispersed [6–8] and there are various means to demonstrate it [9–11]. Among the causes of overdispersion is

the excess of zeros. Within the context of malaria vectors counts, the excess of zeros may result from the absence of mosquitoes at some locations (houses, village…) or during some period of time (dry season, cold temperatures…).

To deal with such overdispersed data with excess zeros, Johnson and Kotz [12] introduced the zero-inflated Poisson model (ZIP); i.e., a Poisson mixture model that combines a point mass at zero with a Poisson count distribution. Later, Lambert [13] extended this model to allow for covariates. Another way to deal with count overdispersion is the use of the negative binomial (NB) model or, better, the zero-inflated negative binomial (ZINB) model constructed on the same principle as that of the ZIP.

Besides these well-known models, other finite mixture distribution models have been proposed (e.g., McLachlan and Peel [14]) and have been the object of numerous applications. In fact, these models extend the previous ones; instead of considering a mixture of two distributions as with the ZIP or the ZINB, they consider a mixture of three or more Poisson or NB distributions. In addition, a non-parametric approach of the maximum likelihood introduced by Aitkin [15] has shown to be an excellent tool to allow for overdispersion. An extension of this approach by the same author [16] allowed its application to repeated measurements. Thus, a non-parametric mixture of Poisson model (NPMP) seems adapted to take into account the frequent changes in vector counts in various sites of a study zone.

In the present work, we assessed the ability of Poisson, NB, ZIP, ZINB and NPMP to fit the distribution of counts of malaria vectors measured in 28 villages in southern Benin where a clinical trial was implemented to evaluate the efficacy of vector control interventions for malaria prevention [4]. Using a multivariate NPMP, we introduced a classification of the villages based on the mean vector density after adjustment for a set of environmental and climatic covariates. Then, we assessed the relationship between this classification and the vector interventions implemented in the villages. The results of this work will help design site-specific malaria vectors sampling.

Methods

Mosquito collection

The data analyzed in the present study stem from mosquito collections carried out every 6 weeks between January and December 2009 (i.e. 8 surveys) in 28 villages of the sanitary region of Ouidah-Kpomassè-Tori (OKT) in South Benin. Of the 58 villages screened at the baseline, 28 were enrolled. The other villages were excluded because they did not fulfill inclusion criteria i.e. distance between two villages >2 km, population size between 250 and 500 inhabitants with non-isolated habitations and absence of any local health care centre.

Entomological surveys were performed using the HLC technique, on two successive nights (22:00 to 06:00) at four sites (both indoor and outdoor) per village. Collectors were hourly rotated along collection sites and/or position (indoor/outdoor). Malaria vectors collected on humans were identified using morphological keys [17]. Only *Anopheles gambiae* and *Anopheles funestus* mosquito counts were considered in the present work because these are the main malaria vectors in West Africa [18–20] and practically the only present in the study area [5].

These villages were divided into four groups (seven villages per group) where four different vector control measures were implemented (see Corbel et al. [4] for details): i) targeted-coverage LLIN (TLLINs) destined to protect pregnant women and children <6 years old (the reference group); ii) universal-coverage LLIN destined to protect sleeping units (ULLINs), iii) TLLINs plus full IRS of carbamate every eight months (TLLIN+IRS), and ULLIN plus full CTPS taped to the upper part of the walls (ULLIN+CTPS).

Ethics statement

The IRD (Institut de Recherche pour le Développement) Ethics Committee and the National Research Ethics Committee of Benin approved the study (CNPERS, reference number IRB00006860). The study was also registered with Current Controlled Trials, number ISRCTN07404145. All necessary permits were obtained for the described field studies. No mosquito collection was done without the approval of the village chief, the owner and occupants of the collection house. Mosquito collectors gave their written informed consent and were treated free of charge for malaria presumed illness throughout the study.

Demographic, geographic and environmental data

The following data were collected: the average distance (in km) from each village to the nearest freshwater body (Toho lake), the presence of market gardening 2 km around each village, the presence of cattle farms inside the village, the presence of water conveyance in the village, and the population density. The layout (or structure) of each village was described by the distribution of its clusters of houses, these clusters being separated by vegetated strips. Two modalities were then considered: single-cluster vs. multi-cluster villages. Daily rainfall data from 8 weather stations were spatially interpolated to compute the cumulated rainfall (in mm) and the number of rainy days in each village during the 15 days preceding each survey. The Normalized Difference Vegetation Index (NDVI) was derived from a "Satellite pour l'Observation de la Terre (SPOT-5)" satellite image acquired on 12/28/2003. The mean NDVI was computed in a buffer area of 50 m diameter around each mosquito collection site (house).

Checking overdispersion and excess of zero in the data

The mean-variance relationship regarding the number of collected malaria vectors was analyzed graphically to explore data dispersion. A linear relationship of slope 1 (variances equal to means) indicated a Poisson distribution without overdispersion whereas a linear relationship with slope >1 or a quadratic relationship indicated overdispersion. We also assess the "excess of zero" through a graphical representation of the distribution of vector counts.

Approximation of the distribution of the data

The approximations of the distribution of the number of collected malaria vectors by the Poisson, ZIP, NB, ZINB and NPMP distributions were compared using the maximum likelihood (ML) estimation. Poisson, ZIP, NB and ZINB models were fitted using the function ***nlm*** [21,22] in the 'R' software version 2.14.0. Parameters of the NPMP model were estimated with the function ***alldist*** [15,23] which used an EM algorithm [24]. In this approach, the dispersion of the data is described by a probability law that does not take into account the hierarchical structure of the data. This is thus a "marginal" model. A graphical representation of comparison results is used to show the counts as well as the predictions given by each of the above-cited distributions.

Multivariate Analysis

Given the hierarchical structure of the data collection system, another NPMP model was considered to allow for various components of the variance of the counts. In this model, the counts of malaria vectors were assessed according to environmen-

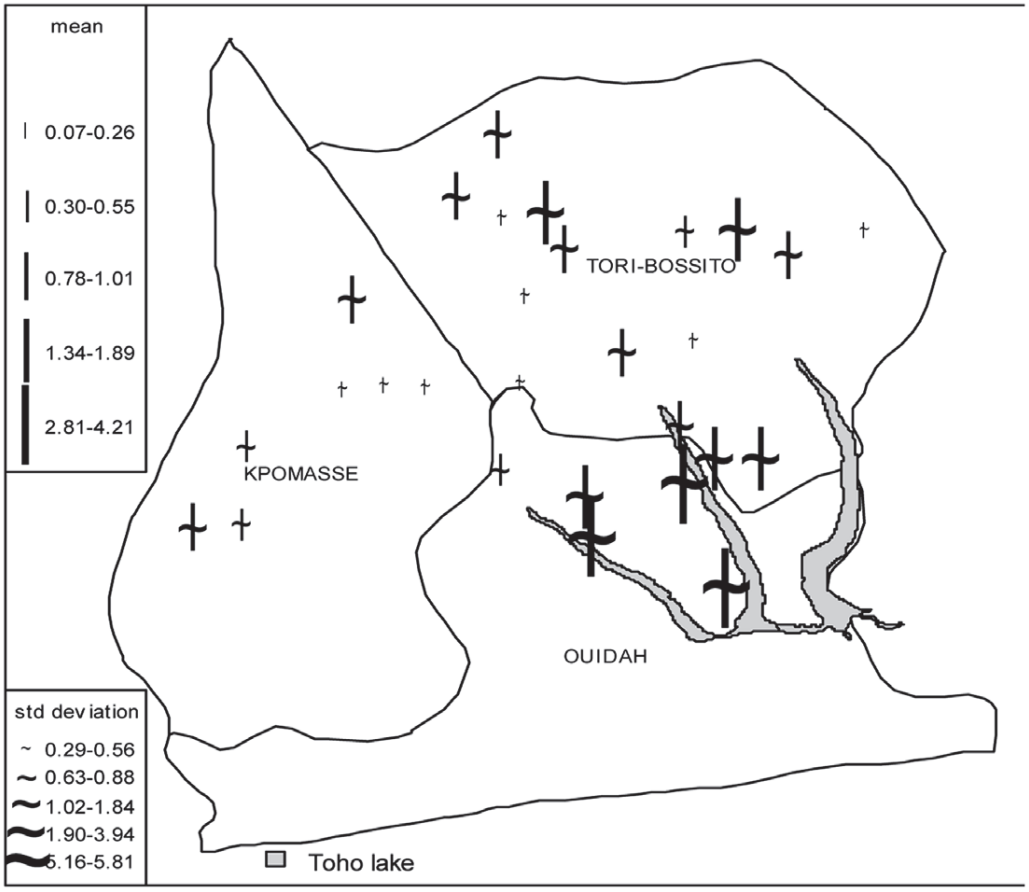


Figure 1. Means and standard deviations of the number of mosquitoes collected per site and per night at each of the 28 villages of the study.
doi:10.1371/journal.pone.0050452.g001

tal and climatic covariates with the "village" as a random effect. It is thus a "conditional" model (on "village"). The latter model allows for the following variables: average distance to Toho lake (in km), water conveyance (0 = absence, 1 = presence), market gardening (0 = absence, 1 = presence), cattle farms inside the village (0 = absence, 1 = presence), the layout of the village (0 = multi-cluster, 1 = single-cluster), population density (inhabitants per 100 m^2), both the mean cumulated rainfall over the 8 surveys (in mm) and the deviation from this mean at each survey, both the mean cumulated number of rainy days over the 8 surveys and the deviation from this mean at each survey, both the averaged NDVI over the 4 collection houses per village and the deviation from this average for each house and, finally, the specific collection site (0 = inside of the house, 1 = outside of the house).

According to the current recommendation for the use of hierarchical models, each covariate was centered on its mean before introduction into the model [25]. Variable "survey" was introduced into the model as a fixed effect. Mosquito collections made inside or outside each house of each village were considered as repeated measurements within that village.

In the NPMP conditional model, the number of malaria vectorsyig collected at a given site of a given village g during a given night i is supposed, conditionally to "the village", to follow a mixture of four Poisson distribution. Each Poisson distribution has a mean μig so that $\log(\mu ig) = \eta ig = xig\beta + \zeta ig$. Note that xig is the vector of values taken by the covariates, β the corresponding fixed effects, ζig the random intercept specific to each village so that $\zeta ig = ac$, with probabilities πc [25]. The values taken by ac are called "latent variables"; c indicating each latent class, here fixed to four$c \in \{1,2,3,4\}$. Hence, the density function of the model can be expressed as $f(yig) = \sum_{c=1}^{4} \pi cfc(yig)$ where fc is the density function of a Poisson distribution with mean μig. Thus, a non-parametric mixture model may also be called "latent class model". The four values exp(ac) (one value for each latent class) are the predicted mean numbers of malaria vectors collected whenever all the model covariates, centered on their means, are equal to zero.

Function ***allvc*** [16,23], a variant of ***alldist*** adapted for hierarchical data, was used for the latter model implementation in R software.

Assessing the impact of vector control strategies

For each village, a posteriori probability of belonging to each class after adjustment on all the covariates is estimated by the NPMP conditional model. Here, "a posteriori probability" means the conditional probability for a village to belong to a given class, given the data. For a village g and a latent class$k$, this probability can be expressed: $pkg = \frac{\pi kP(yg|k)}{\sum_{c=1}^{4} \pi cP(yg|c)}$. In this expression, $yg$is the

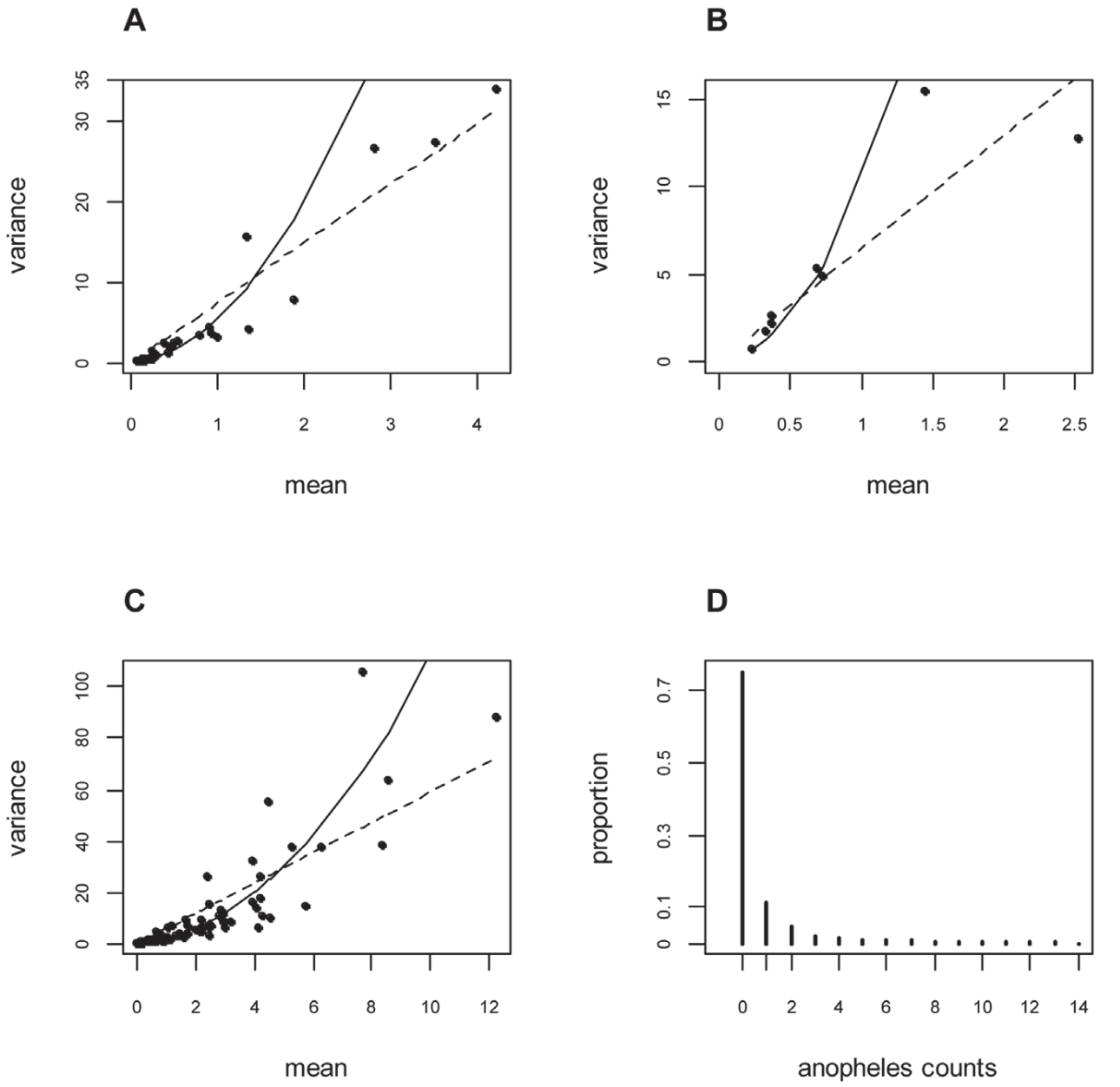


Figure 2. Mean-variance diagrams of the number of malaria vectors collected per village (Panel A), per mission (Panel B), and per village-mission (Panel C). Panel D shows a bar diagram of the distribution of mosquito counts at each collection site (the scale of the X-axis was limited to 14). On panels A, B and C: the dotted lines represent a linear link between the means and the variance ($variance = \alpha \times mean$ with $\alpha = 7.4$, 6.48 and 5.9 respectively); the curves represent a quadratic link between mean and variance ($variance = mean + \beta \times mean^2$ with $\beta = 4.4$, 8.9 and 1.1 respectively).
doi:10.1371/journal.pone.0050452.g002

vector of observations in the village g and $P(yg|c)$ the probability of yg assuming membership in class c.

Hence, each village is assigned to one of the classes based on the maximum of the a posteriori probabilities (MAP). This provides a classification of the villages according to the average number of malaria vectors collected at a given site over a given night after adjustment on all the covariates.

In order to assess the impact of TLLIN, ULLIN, ULLIN+CTPS and TLLIN+IRS vector control strategies, the village grouping for implementation of these vector control strategies and the classification resulting from the NPMP conditional model were compared using a Kruskall-Wallis test.

About the number of the latent classes

The relevance of a NPMP model also called Poisson latent classes model, be it marginal or conditional, depends jointly on its ability to provide a close distribution to that of the observed counts and on its ability to assign each count one of the classes. Essentially, two criteria contributed to the choice of the number of classes: the closeness of the predicted values to the observed ones,

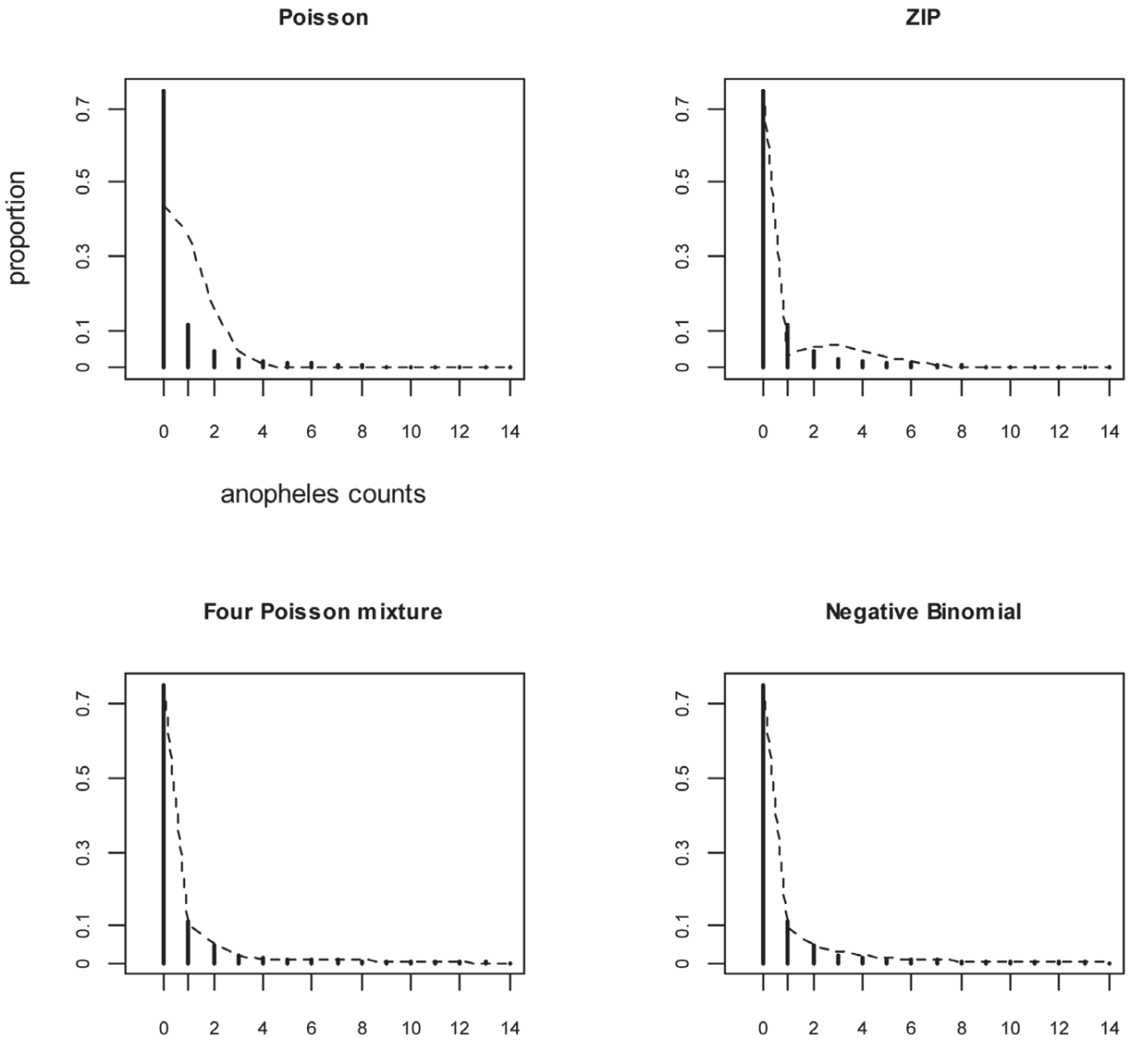


Figure 3. Observed and expected proportions of mosquito counts according to Poisson, ZIP, NPMP and NB distributions.
doi:10.1371/journal.pone.0050452.g003

which is the deviance expressed under the form of a Bayesian Information Criterion [26]; and, the ability of the model to assign each count one of the classes, which is expressed by the Entropy [27]. The Integrated Complete-data Likelihood (ICL-BIC) [28] is a combination of these two criteria; precisely, the BIC plus two times the entropy. Hence, the number of classes chosen for a latent class model is the one that maximizes the likelihood with low entropy equivalent to a minimum ICL-BIC.

Results

Entomological Data

Total of 2,994 malaria vectors were collected during 3,584 human-nights of mosquito collection. This corresponded to an average HBR of 0.835 bites per human per night. Among these vectors, 1,872 belonged to the *An. gambiae* complex and 1,122 belonged to the *funestus* Group. The density of anopheline collected changed over space and time (Figure 1). Indeed, the mean HBR in OKT ranged from 0.070 to 4.219 bites per human per night when the standard deviation ranged from 0.286 to 5.812. In most villages, high standard deviations corresponded to high means and between surveys the number of malaria vectors collected varied.

Study of the dispersion

Village, survey, and village-survey mean numbers of malaria vector collected per night on humans were plotted with their corresponding variances in Figures 2A, 2B and 2C respectively. The assumption of mean-variance equality of the Poisson distribution was not met. Indeed, the variances were much higher than the means and the slopes of the linear relationships were >1 showing even quadratic relationships. This indicates overdispersion of the data.

Table 1. Parameters and deviance as estimated by the Poisson, ZIP, NPMP, NB and ZINB models.

	Parameters			
Distribution	**Mean (SE)**	**Proportion (SE)**	**Dispersion parameter**	**−2logL**
Standard Poisson	0.835 (0.015)	1 (-)		13492.470
Zero-inflated Poisson (ZIP)				9229.370
Zero-class	0 (-)	0.736 (0.008)		
Poisson	3.169 (0.062)	0.264 (0.008)		
Poisson mixture model with 4 latent classes (NPMP)				7591.700
Low	0 (7×10^{-6})	0.630 (-)		
Median-low	0.923 (0.029)	0.296 (-)		
Median-high	6.555 (0.161)	0.070 (-)		
High class	24.480 (1.281)	0.004 (-)		
Negative Binomial (NB)	0.835 (0.038)	1 (-)	0.156 (0.007)	7581.856
Zero-inflated negative binomial (ZINB)				7581.856
Zero-class	0 (-)	3.6×10^{-6}(1.7×10^{-5})		
NB	0.835 (0.038)	0.999 (1.7×10^{-5})	0.156 (0.007)	

−2logL: −2 times the log-likelihood
doi:10.1371/journal.pone.0050452.t001

Distribution analysis

Frequency plot of the collected malaria vectors is shown in Figure 2D. The cases for which zero malaria vectors were collected represented 74.7% of the total. Table 1 shows the parameters for Poisson, ZIP, NPMP (the marginal model), NB and ZINB distributions. Based on the Poisson distribution with a mean

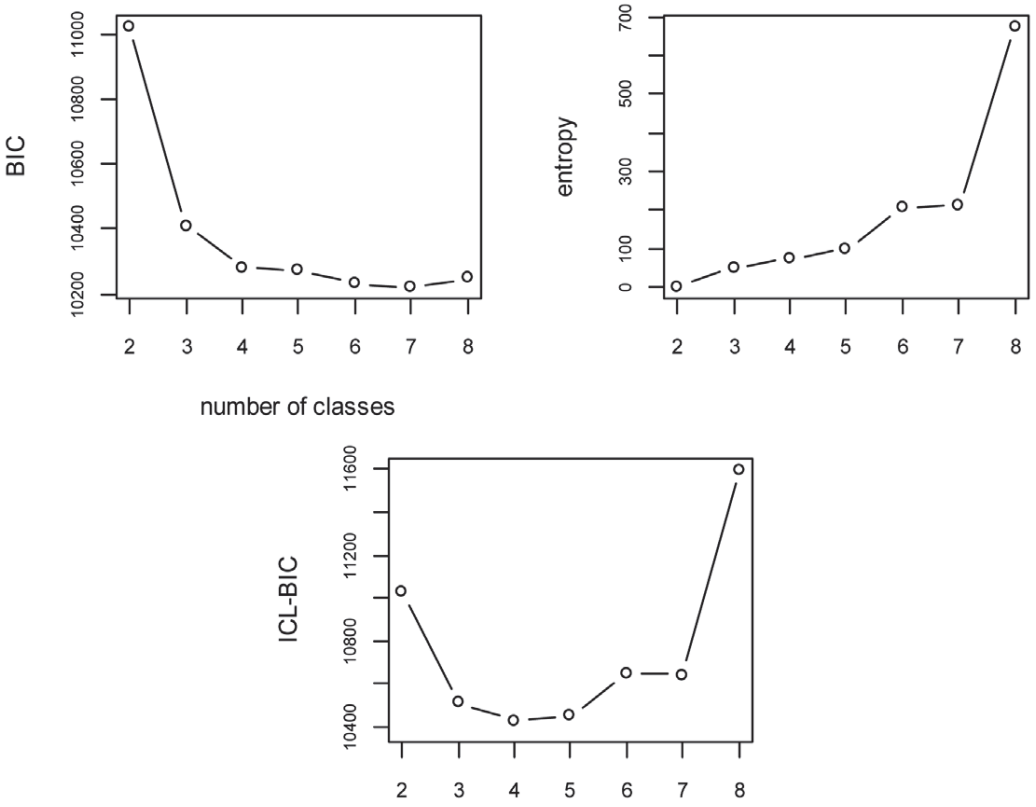


Figure 4. Changes in the values of the Bayesian Information Criterion (BIC), the entropy, and the Integrated Complete-data Likelihood (ICL-BIC) according to the number of latent classes.
doi:10.1371/journal.pone.0050452.g004

Table 2. Estimations of the relationships between mosquito density and various geographical and environmental factors in OKT region according to the conditional NPMP model.

Level and covariate	Relative Risk (95% CI)
Village	
Distance to a freshwater body (per additional km)	0.885 (0.871–0.899)
Presence of water conveyance (Yes vs. No)	0.411 (0.348–0.485)
Presence of market gardening (Yes vs. No)	1.146 (1.016–1.292)
Presence of cattle (Yes vs. No)	0.817 (0.700–0.954)
Layout of the village (single- vs. multi-cluster)	0.466 (0.377–0.574)
Population density (per additional inhabitant/100 m^2)	1.335 (1.079–1.651)
Mean rain quantity over all surveys (per additional mm)	1.325 (1.292–1.359)
Mean number of rainy days over all surveys (per additional day)	2.148 (1.675–2.754)
Mean NDVI (per additional grade)	0.849 (0.827–0.872)
House	
Deviation from the mean NDVI of the village (per additional grade)	0.990 (0.978–1.003)
Collection site (outside vs. inside)	1.182 (1.100–1.270)
Mission	
Deviation* from the mean rain quantity (per additional mm)	0.993 (0.989–0.997)
Deviation* from the mean number of rainy days (per additional day)	0.902 (0.827–0.984)

*Difference between the mean value over all surveys and the value at a given survey

doi:10.1371/journal.pone.0050452.t002

of 0.835, we would expect 43.4% of zero (Figure 3A) which is significantly lower than that observed in the dataset. In contrast, ZIP, NB and NPMP models well predicted the proportion of zero (respectively 74.7%, 74.9%, and 74.7%; Figure 3). Excluding the sites where no anopheline were collected, the ZIP model estimated the mean number of collected anopheline per site per night at 3 but this does not solve the problem of data overdispersion. The NPMP model suggested that, in the study area, the sites where the anopheline counts over a single night would be generally high (mean 24.48) were rather rare (nearly 0.4%). However, the model estimated at 63% and 30%, respectively, the proportions of counts with 0 and 1 as the mean number of collected anopheline per night. The dispersion parameter estimated by the NB model was 0.156 indicating high overdispersion. Since the NB model allows for the excess of zeros, the proportion of zeros estimated by the ZINB model is nearly null. The NB and the ZINB models are therefore equivalent.

There were a significant decrease of the model deviance when a ZIP model was used instead of the standard Poisson model and also when the NPMP was used instead of the ZIP model (Table 1). The NB and ZINB fitted the data as well as the NPMP (deviance were not significantly different). Figure 3 shows the bar diagrams and the expected density probability curves of the counts with the Poisson, ZIP, NPMP and NB models. The curves relative to the NB and to the NPMP models are very similar and fit well the observed data distribution. Conversely, the curve relative to the standard Poisson model does not fit the observed frequency of counts between 0 and 10. Aside from the proportion of zeros, the ZIP model was not able to reproduce the observed proportions of counts ranging from 1 to 10.

Choosing the number of latent classes

Figure 4 shows for the NPMP model conditional on "village", the progress of the BIC, the entropy, and the ICL-BIC according to the number of classes. Starting from 4 classes, the BIC became very low. The entropy augmented together with the number of classes. Their combination ICL-BIC was at its minimum with 4 classes. Therefore, the model with 4 latent classes was used to assess the number of malaria vector caught on humans according to climatic and environmental factors.

Multivariate Poisson mixture analysis

Table 2 shows the relative risk of the fixed effects as estimated by the model. Presence of market gardening, population density, mean cumulated rainfall over the 8 surveys, mean cumulated number of rainy days over the 8 surveys and outdoor position were positively associated with the number of malaria vectors caught on human. On the other hand, distance to a freshwater body, presence of water conveyance, presence of cattle, single-cluster village houses, mean NDVI and deviation at each survey from mean cumulated number of rainy days over the 8 surveys were negatively associated with the number of malaria vectors caught on human.

Table 3 shows the random effects of the model. These are the predicted mean number of collected malaria vectors per night for each of the four latent classes when all the covariates are at their mean values. This table shows also the final classification of the villages according to their respective MAP. The mean number of malaria vectors collected ranged from 0.050 vectors per human per night in the 1st class (with only one village: Hekandji) to 0.713 in the 4th class (with 8 villages).

A Kruskal-Wallis test did not show a significant association between villages classification obtained from the model and the villages grouping for vector control strategies (Chi2 = 2.029, p-value = 0.566). Thus, according to HBR, a significant difference in term of impact of vector control strategies (TLLIN, ULLIN, ULLIN+CTPS and TLLIN+IRS) is not showed.

Discussion

Knowledge of malaria vector density in a given area is often needed for implementing and evaluating vector control interven-

Table 3. Classification of the 28 villages according to the maximum a posteriori probability (MAP) of belonging to each class after adjustment on all other covariates.

Village	Latent class	Mean number of mosquitoes	Proportion of villages	MAP
Hekandji	1	0.050	0.036	0.992
Aidjedo	2	0.137	0.218	0.997
Assogbenou				1
Ayidohoue				0.990
Dokanmey				0.998
Hounkponouhoue				1
Abenihoue				1
Adjame	3	0.324	0.466	1
Amoulehoue				1
Adjahassa				0.924
Kindjitokpa				1
Vidjinnagnimon				1
Guezohoue				1
Hla				1
Agokon				0.968
Dekponhoue				0.998
Lokohoue				1
Todo				1
Wanho				1
Zoume				0.994
Agouako	4	0.713	0.280	0.775
Hinmadou				0.925
Manguevie				0.925
Satre				0.925
Soko				0.925
Tanto				0.925
Tokoli				0.925
Agadon				0.925

doi:10.1371/journal.pone.0050452.t003

tions. This requires vector counts at several sites of the area and statistical analyses of these counts.

McCullagh and Nelder [29] asserted that whenever the variable of interest is a count, its distribution is often an overdispersed Poisson distribution. The present data are another illustration of this assertion. The first part of this work aimed at comparing which distribution among Poisson, NB, ZIP, ZINB and NPMP better fit on counts of malaria vectors recorded using the HLC technique. Both NB and NPMP models dealt with the excess of zero, with overdispersion and provided the best predictions of the distribution of the observed data. However, unlike NB model, the NPMP does not do any further assumption about the distribution of the means of malaria vectors counts. Besides, the hierarchical structure of the observed data was taken into account by a NPMP model conditional on ''village''. Based on a posterior probability criterion, the NPMP model allowed ranking the villages in four latent classes according to the mean of vector density after adjustment on environmental and climatic covariates. The optimal number of latent classes was established on conventional criteria. Furthermore, the part each covariate played in the variability of malaria vector density in the area was estimated by the fixed effects of the model. However, the present study could not take into account all the possible hierarchical levels of the data because of the limits of software R in dealing with latent classes. Indeed, function ***allvc*** of package ''npmlreg'' cannot deal with more than two levels. We considered thus the catches at all sites of the same village as repeated measurements of the same variable. Therefore, we were not able to take into account the possible correlation between the counts from houses within the same village [30]. ''Human bait'' is another level that could induce correlation in the data but there is no sufficient information about all mosquito collectors. Besides, the rotation of the collectors during data collection reduces considerably such a correlation. ''Season'' could be another possible level of correlation; it was taken into account through rainfall data which is the main seasonal factor in the context of malaria vector density.

Moreover, the numbers of collected vectors during the 8 surveys are assumed to be uncorrelated although one may speculate about a correlation structure along time. Nevertheless, the correlation between mosquito counts from successive surveys is deemed to be very low because the time span between two successive surveys is 6 weeks whereas the lifespan of the vectors is only 3 to 4 weeks. Studying the correlation between counts from two nights during the same survey may reveal interesting results.

In southern Benin, both spatial and temporal heterogeneities in vector densities were mentioned by Djènontin et al. [5]. This can be explained by some factors we found associated with the density of malaria vectors. Firstly, cumulated rainfalls during the 15 days preceding the catches were positively associated with vector density as previously reported in Benin [30]. Moreover, the mean number (over all surveys) of rainy days was positively associated with the vector density whereas the deviation at each survey from this mean was negatively associated with the vector density. This suggested that high frequency of rainy events might flush out vectors breeding sites [31]. The vector density was lower in villages with water supply; this could be due to the absence of water storages that could have provided breeding sites for malaria vectors [32,33]. Moreover, the presence of irrigated market gardening could have provided breeding sites [34,35] and then, increased the density of vectors in villages closed to this activity as previously observed in Benin [36]. Permanent freshwaters of the Toho Lake could also have provided breeding sites for both *An. funestus* and *An. gambiae* [37–39] that are both present in our study area [5]. This explains why the vector density decreased when moving away from freshwater bodies as showed by Amek et al. [40] in Western Kenya. The presence of cattle was negatively correlated with vector density suggested that a part of the vector population could have bite on cattle instead of human. More vectors were caught in multi-cluster villages than in single-cluster villages. This might indicate that a multi-cluster village layout might increase the attractiveness of the village for malaria vectors because of the extra vegetation surrounding houses. Thus, the attractiveness of a multi-cluster village may be higher than that of a single-cluster village of same size. Catches were also more abundant outside than inside the houses. This indicates an exophagic behavior of malaria vectors in the study area. As suggested by two studies in the OKT region [4,41], a part of the exophagic population of vectors could have avoided indoor residual insecticides.

One unexpected finding of the present study was that the NDVI was negatively correlated with the density of malaria vectors. This finding contrasts with several studies that used satellite imagery at a lower resolution [42,43] but agrees with a study carried out in Burkina Faso that used the same SPOT images than ours [44]. In this study, the authors found a negative relationship between the larval productivity in ponds and the NDVI calculated from high resolution SPOT images. Indeed, a high NDVI might reflect the presence of submerged vegetation or water covered with vegetation that are usually related to very high *Anopheles* larval densities [44–46]. Moreover, the NDVI usually decreases with freshwater and unvegetated surfaces likely to provide breeding sites for the malaria vectors [37,47]. Nevertheless, the discussion about the NDVI effect can be more complex because of the co-existence in the region of two major malaria vectors with different breeding-site requirements.

In this work, villages were ranked into four classes of increasing mean malaria vector density but we were not able to find any relationship between this grouping structure and the vector control intervention implemented in the village. This confirms the finding of Corbel et al. [4] who demonstrated with the same data, that vector density was not significantly different between treatment arms (TLLIN, ULLIN, TLLIN+IRS, and ULLIN+CTPS).

In conclusion, we found that the NPMP model was useful to assess the relationships between vectors density and villages or environmental characteristics. It might therefore be an efficient tool to compute risk maps of the host-vector contact. Moreover, the NPMP model provided a classification of the villages after taking into account some covariates. Such a classification could be used at a pre-study step to improve the study design of mosquito collection and adapt the sampling effort according to the village characteristics, especially in region with high spatial and temporal heterogeneities of mosquito density, like in the OKT region. Furthermore, NPMP model could help in the study design of RCT when a stratified sampling is needed. The same model may be adapted and used in other settings for the study of the distribution of vectors of other diseases.

Acknowledgments

We thank the populations and authorities of the OKT district for their kind support and collaboration. We also thank Pr. Jean-François Etard for his helpful contribution to the conception and design of the present study.

Author Contributions

Conceived and designed the experiments: RE VC NF OB. Analyzed the data: OB RE. Wrote the paper: OB RE JI NM VC. Collected the data: AD NM SBB. Revised the manuscript: RE NF VC NM JI AD OB.

References

1. WHO (2011) World Malaria Report 2011. Geneva: World Health Organization.
2. Trape JF, Tall A, Diagne N, Ndiath O, Ly AB, et al. (2011) Malaria morbidity and pyrethroid resistance after the introduction of insecticide-treated bednets and artemisinin-based combination therapies: a longitudinal study. The Lancet Infectious Diseases.
3. Silver JB SM (2008) Sampling adults by animal bait catches and by animal-baited traps. Mosquito ecology field sampling methods: Springer. pp. 493–675.
4. Corbel V, Akogbeto M, Damien GB, Djenontin A, Chandre F, et al. (2012) Combination of malaria vector control interventions in pyrethroid resistance area in Benin: a cluster randomised controlled trial. Lancet Infect Dis 12: 617–626.
5. Djenontin A, Bio-Bangana S, Moiroux N, Henry MC, Boussari O, et al. (2010) Culicidae diversity, malaria transmission and insecticide resistance alleles in malaria vectors in Ouidah-Kpomasse-Tori district from Benin (West Africa): A pre-intervention study. Parasit Vectors 3: 83.
6. Alain E, Brenac T (2001) Modèles linéaires généralisés appliqués a l'étude des nombres d'accidents sur des sites routiers. Le modèle de Poisson et ses extensions. Recherche Transports Sécurité: 3–16.
7. Bouche G, Lepage B, Migeot V, Ingrand P (2009) Application of detecting and taking overdispersion into account in Poisson regression model. Rev Epidemiol Sante Publique 57: 285–296.
8. Gardner W, Mulvey EP, Shaw EC (1995) Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. Psychol Bull 118: 392–404.
9. Breslow N (1990) Tests of hypotheses in overdispersed Poisson regression and other quasi-likelihood models. Journal of the American Statistical Association: 565–571.
10. Dean C, Lawless JF (1989) Tests for detecting overdispersion in Poisson regression models. Journal of the American Statistical Association: 467–472.
11. Hauer E, Ng JCN, Lovell J, Board NRCTR (1988) Estimation of safety at signalized intersections: National Research Council, Transportation Research Board.
12. Johnson NL, Kotz S (1969) Discrete distributions: Wiley Online Library.
13. Lambert D (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics: 1–14.
14. McLachlan GJ, Peel D (2000) Finite mixture models: Wiley-Interscience.
15. Aitkin M (1996) A general maximum likelihood analysis of overdispersion in generalized linear models. Statistics and computing 6: 251–262.
16. Aitkin M (1999) A general maximum likelihood analysis of variance components in generalized linear models. Biometrics 55: 117–128.
17. Gillies M, De Meillon B (1968) The Anophelinae of Africa south of the Sahara (Ethiopian Zoogeographical Region). Johannesburg: S. Afric Inst Med Res.
18. Akogbeto M (1995) Entomological study on the malaria transmission in coastal and lagoon areas: the case of a village built on a brackish lake. Ann Soc Belg Med Trop 75: 219–227.
19. Fontenille D, Simard F (2004) Unravelling complexities in human malaria transmission dynamics in Africa through a comprehensive knowledge of vector populations. Comp Immunol Microbiol Infect Dis 27: 357–375.

20. Kelly-Hope LA, McKenzie FE (2009) The multiplicity of malaria transmission: a review of entomological inoculation rate measurements and methods across sub-Saharan Africa. Malar J 8: 19.
21. Dennis J, Schnabel RB (1983) Numerical methods for nonlinear equations and unconstrained optimization. Classics in Applied Math 16.
22. R Development Core Team and contributors worldwide. The R Stats Package, version 2.14.1.
23. Einbeck J, Darnell R, Hinde J (2009) npmlreg: Nonparametric maximum likelihood estimation for random effect models. R-project org/package = npmlreg. 0.44 ed.
24. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society Series B (Methodological): 1–38.
25. Rabe-Hesketh S, Skrondal A (2004) Counts. Generalized latent variable modeling: multilevel, longitudinal, and structural equation models: Chapman & Hall. pp. 349–372.
26. Schwarz G (1978) Estimating the dimension of a model. The annals of statistics 6: 461–464.
27. Nagin DS, Odgers CL (2010) Group-Based Trajectory Modeling in Clinical Research. Annual Review of Clinical Psychology 6:109–138.
28. Biernacki C, Celeux G, Govaert G (2000) Assessing a mixture model for clustering with the integrated completed likelihood. Pattern Analysis and Machine Intelligence, IEEE Transactions on 22: 719–725.
29. McCullagh P, Nelder JA (1989) Generalized linear models: Chapman & Hall/CRC.
30. Cottrell G, Kouwaye B, Pierrat C, le Port A, Bouraima A, et al. (2012) Modeling the influence of local environmental factors on malaria transmission in Benin and its implications for cohort study. PLoS One 7: e28812.
31. Paaijmans KP, Wandago MO, Githeko AK, Takken W (2007) Unexpected high losses of Anopheles gambiae larvae due to rainfall. PLoS One 2: e1146.
32. Bio-Bangana S, Chandre F, Djènontin A, Chabi J, Ossè R, et al. (2009) Characterization of mosquito fauna in Ouidah, Kpomasse and Tori-Bossito Sanitary Zone in Benin (West Africa). Multilateral Initiative on Malaria. Nairobi, Kenya. Poster.
33. Holstein MH (1954) Biology of Anopheles gambiae: research in French West Africa: World Health Organization.
34. Matthys B, Vounatsou P, Raso G, Tschannen AB, Becket EG, et al. (2006) Urban farming and malaria risk factors in a medium-sized town in Cote d'Ivoire. Am J Trop Med Hyg 75: 1223–1231.
35. Klinkenberg E, McCall P, Wilson MD, Amerasinghe FP, Donnelly MJ (2008) Impact of urban agriculture on malaria vectors in Accra, Ghana. Malar J 7: 151.
36. Yadouleton A, N'Guessan R, Allagbe H, Asidi A, Boko M, et al. The impact of the expansion of urban vegetable farming on malaria transmission in major cities of Benin. Parasit Vectors 3: 118.
37. Hamon J (1955) Biologie d'Anopheles funestus. Biologie des anophèles d'AOF et d'AEF. Paris: ORSTOM. pp. 6 multigr.
38. Costantini C, Ayala D, Guelbeogo WM, Pombi M, Some CY, et al. (2009) Living at the edge: biogeographic patterns of habitat segregation conform to speciation by niche expansion in Anopheles gambiae. BMC Ecol 9: 16.
39. Simard F, Ayala D, Kamdem GC, Pombi M, Etouna J, et al. (2009) Ecological niche partitioning between Anopheles gambiae molecular forms in Cameroon: the ecological side of speciation. BMC Ecol 9: 17.
40. Amek N, Bayoh N, Hamel M, Lindblade KA, Gimnig J, et al. (2012) Spatial and temporal dynamics of malaria transmission in rural Western Kenya. Parasit Vectors 5: 86.
41. Moiroux N, Bustamante GM, Pennetier C, Elanga E, Djègbé I, et al. (2012) Changes in Anopheles funestus biting behaviour following universal coverage of long-lasting insecticidal nets in Benin. J Infect Dis: In press.
42. Thomson MC, Connor SJ (2000) Environmental information systems for the control of arthropod vectors of disease. Med Vet Entomol 14: 227–244.
43. Hay SI, Omumbo JA, Craig MH, Snow RW (2000) Earth observation, geographic information systems and Plasmodium falciparum malaria in sub-Saharan Africa. Adv Parasitol 47: 173–215.
44. Dambach P, Sie A, Lacaux JP, Vignolles C, Machault V, et al. (2009) Using high spatial resolution remote sensing for risk mapping of malaria occurrence in the Nouna district, Burkina Faso. Glob Health Action 2.
45. Gillies M, Coetzee M (1987) A supplement to the Anophelinae of Africa South of the Sahara (Afrotropical Region). Johannesburg, South Africa: South African Institute for Medical Research. 143 p.
46. Gimnig JE, Ombok M, Kamau L, Hawley WA (2001) Characteristics of larval anopheline (Diptera: Culicidae) habitats in Western Kenya. Journal of Medical Entomology 38: 282–288.
47. Edillo FE, Toure YT, Lanzaro GC, Dolo G, Taylor CE (2002) Spatial and habitat distribution of Anopheles gambiae and Anopheles arabiensis (Diptera: Culicidae) in Banambani village, Mali. J Med Entomol 39: 70–77.

Chapitre 5

Profils annuels du contact homme-vecteur dans le cadre du paludisme

Dans le précédent chapitre, nous avons mis en évidence l'hétérogénéité spatiale de l'intensité moyenne (sur une année) du contact homme-vecteur dans un ensemble de 28 villages du Sud Bénin dans lesquels des captures de moustiques ont été effectuées. Le présent chapitre s'intéresse non pas à l'intensité moyenne, mais au profil évolutif du contact homme-vecteur au cours d'une année dans la même zone géographique et donc s'appuie sur le caractère longitudinal des données. Nous considérons la même période de collecte qu'au chapitre précédent, c'est-à-dire de Janvier à Décembre 2009. Les vecteurs considérés sont *An. gambiae* et *An. coluzzii*. La quasi-absence ou l'absence, dans plusieurs des 28 villages, d'*An. funestus* n'a pas permis de considérer ce vecteur dans le travail objet du présent chapitre. En effet 21 villages sur les 28 comptent chacun au moins 6 missions (sur les huit considérées) au cours desquelles aucun moustique *An. funestus* n'a été capturé.

5.1 Etude de l'hétérogénéité spatio-temporelle du contact homme-vecteur

Pour ne s'intéresser qu'aux profils annuels d'agressivité c'est-à-dire qu'à la forme des trajectoires de contacts homme-vecteur au cours de l'année, nous avons stan-

dardisé au niveau de chaque village, les nombres d'*An. gambiae* et *An. coluzzii* capturés par site (maison de capture) et par nuit. Ceci à consister à considérer au niveau de chaque village, 64 unités statistiques conformément au plan d'échantillonnage décrit dans la section 1.6 du Chapitre 1 : 4 maisons de capture par village, 2 nuits de capture par mission de capture, 8 missions de capture pendant une année (de Janvier à Décembre 2009). Le nombre de vecteurs capturés au niveau de chaque unité statistique a été centré et réduit en utilisant la moyenne et la variance calculées sur les 64 valeurs. La distribution de la variable standardisée ainsi obtenue au niveau de chaque village est alors de moyenne nulle et de variance 1. Ceci garantit que seules les trajectoires annuelles de contact homme-vecteur différencient les villages et non l'intensité du contact homme-vecteur. Par la suite nous faisons l'hypothèse d'une distribution normale des données standardisées.

Comme dans le chapitre précédent, nous avons opté pour les modèles à classes latentes dans la modélisation du profil évolutif du contact homme-vecteur. Dans le cas présent, nous parlons plutôt de trajectoires latentes en raison du caractère longitudinal de la variable réponse. Le temps est pris en compte dans le modèle à travers la variable «mission de capture» (variable catégorielle à 8 modalités). Le nombre de trajectoires latentes typiques est fixé à deux sur la base du critère ICL et la classification des villages est faite suivant la méthode du MAP. Nous aboutissons ainsi à deux profils typiques de trajectoires de contact homme-vecteur : un profil dit «continu» regroupant 18 villages et un profil dit «intermittent» constitué des 10 villages restants.

L'on pourrait penser que cette dichotomisation des profils de contact homme-vecteur traduit le comportement des deux types de vecteurs considérés (*An. gambiae* et *An. coluzzii*). A chaque vecteur correspondrait alors un profil au cours de l'année. Mais des modélisations utilisant séparément les deux types de vecteurs conduisent à des classifications des villages très proches de celle obtenue en cumulant les deux types de vecteurs.

La carte de répartition spatiale des villages suivant le profil du contact homme-

vecteur ne révèle pas la présence de clusters géographiques en termes de profil du contact homme-vecteur. Ceci pourrait s'expliquer par le fait que le profil du contact homme-vecteur soit assez sensible à des conditions suffisamment localisées. Aussi, plusieurs facteurs qui se sont révélés significativement associés à l'intensité moyenne du contact homme-vecteur, ne semblent pas avoir une influence significative sur le profil annuel du contact homme-vecteur. Ce travail, à l'instar de celui présenté dans le chapitre précédent sur l'hétérogénéité spatial de l'intensité du contact homme-vecteur, ne met pas en évidence un lien significatif entre les méthodes de LAV présentes dans les 28 villages et les profils typiques de contact homme-vecteur.

Nous présentons dans la section suivante, sous la forme d'un article, les détails du travail de modélisation des profils annuels du contact homme-vecteur.

5.2 Article 2 : Modeling the seasonality of Anopheles gambiae s.s. biting rates in a South Benin sanitary zone

Trans R Soc Trop Med Hyg 2014; **108**: 237–243
doi:10.1093/trstmh/tru027 Advance Access publication 21 February 2014

ORIGINAL ARTICLE

Modeling the seasonality of *Anopheles gambiae s.s.* biting rates in a South Benin sanitary zone

Olayidé Boussari[a,b,*], Fabien Subtil[b], Nicolas Moiroux[c,d], Armel Djènontin[c,e], Jean Iwaz[b], Vincent Corbel[c,e], Noël Fonton[a], André Garcia[f], Jean-François Etard[g,h] and René Ecochard[b]

[a]International Chair in Mathematical Physics and Applications, Laboratoire d'Etude et de Recherche en Statistique Appliquée et Modélisation, Université d'Abomey-Calavi, Abomey-Calavi, Bénin; [b]Université de Lyon, F-69000, Lyon, France; Université Lyon 1, F-69100, Villeurbanne, France; Laboratoire de Biométrie et Biologie Evolutive, CNRS, UMR5558, F-69100, Villeurbanne, France; Service de Biostatistique, Hospices Civils de Lyon, F-69003, Lyon, France; [c]Institut de Recherche pour le Développement, MIVEGEC (IRD 224-CNRS 5290-UM1-UM2), Cotonou, Benin; [d]Institut de Recherche pour le Développement, MIVEGEC (IRD 224-CNRS 5290-UM1-UM2), Montpellier, France; [e]Centre de Recherche en Entomologie de Cotonou, Ministère de la Santé, Cotonou, Bénin; [f]Institut de Recherche pour le Développement, UMR 216 Mère et Enfant Face aux Infections Tropicales, Paris, France; [g]Epicentre, Paris, France; [h]Institut de Recherche pour le Développement, UMI 233 TransVIHMI, Université Montpellier 1, Montpellier, France

*Corresponding author: Tel: +33 4 72 11 57 58; Fax: +33 4 72 11 51 41; E-mail: olayideb@yahoo.fr

Received 28 October 2013; revised 6 January 2014; accepted 27 January 2014

Background: Efficient malaria vector control requires knowledge of spatio-temporal vector dynamics. We have classified village groups according to the biting rate profiles of both *Anopheles coluzzii* and *An. gambiae*, the major malaria vectors in these villages.

Methods: Mosquitoes were captured by human bait in 28 South Benin villages during 2009. Both *An. coluzzii* and *An. gambiae* counts in each village were standardized to focus on changes in the vector biting rate over time. Latent class trajectory modeling, allowing for random intercept at the 'village' level, was adjusted to standardized values.

Results: The villages could be classified into two groups with distinct vector biting rate profiles (continuous/transient). This classification helped creating a map of vector biting rates in the area. The biting rate profiles were found to be significantly correlated with mean rainfall, altitude, average number of larval sites, and average normalized difference vegetation index.

Conclusions: In highly malaria-prone regions, knowledge of vector biting rate profiles is important to improve vector control interventions. A similar methodology may be applied to study the biting rate profiles of other vector-borne infections.

Keywords: Anopheles gambiae s.s., Biting rate, Classification, Latent trajectory modeling, Malaria vectors, South Benin

Introduction

Malaria is a disease with various epidemiological features,[1–3] not only within large regions but also within relatively restricted areas. This mosaic of features may be explained by a vector distribution that depends highly on environmental factors[4–6] and on changes in vector behaviour.[7–10]

Special attention is paid to vectors in the fight against malaria. Several studies have made substantial progress in generating knowledge on the relationship between vector population dynamics and many area-specific factors (often seasonal).[11] Adapting vector-control measures to specific areas and seasons make them more efficient. It therefore seems important to obtain reliable data on mean vector biting rate (human–vector contact) intensity and changes in this parameter across seasons to identify areas with similar vector biting rates. For example, many studies have shown that mosquito-net use varies considerably during the year and from one area to another, without reliable information on vector biting rate dynamics.

In tropical Africa, the seasonal biting rates of both *Anopheles coluzzii* and *An. gambiae* (formerly *An. gambiae* M and S molecular forms, respectively)[12] are mainly linked to annual rainfall.[13] However, over a whole year, different biting rate profiles may be seen in neighbouring villages that receive similar rainfall levels. These micro-geographic changes may also be linked to other factors, such as the nature of the soil, the presence of freshwater bodies, and livestock breeding.[14,15]

© *The Author 2014. Published by Oxford University Press on behalf of Royal Society of Tropical Medicine and Hygiene. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.*

Downloaded from http://trstmh.oxfordjournals.org/ by guest on April 17, 2014

A previous study[16] demonstrated spatial variability in mean biting rate intensity by major malaria vectors (i.e., *An. coluzzii, An. gambiae* and *An. funestus*) in a specific area of South Benin where four vector-control strategies were implemented. It described the relationship between this variability and several environmental factors.

The present work investigates changes in both *An. coluzzii* and *An. gambiae* biting rate profiles over a whole year in a geographical area measuring nearly 800 km^2. It implemented a geographical classification method based exclusively on annual changes in both *An. coluzzii* and *An. gambiae* biting rate, and not on its mean intensity. The results may help design area-specific, season-specific and, thus, more efficient anti-vector campaigns.

Materials and methods

Data collection and ethics statement

The data analyzed in the present work stem from mosquito captures in the sanitary zone of Ouidah-Kpomassè-Tori (South Benin) from January to December 2009.

Vectors were captured by the human landing catches technique in 28 villages, in four randomly-chosen houses per village and at two sites per house (inside and outside). In each village, eight capture surveys were organized. Each survey consisted of two consecutive capture nights. Successive surveys were conducted 6 weeks apart in each village. These surveys were part of a study that compared four vector-control strategies in the same area (seven villages per strategy)[17]: targeted, long-lasting insecticidal net (LLIN) coverage of pregnant women and children younger than 6 years; universal LLIN coverage of sleep units; targeted LLIN coverage plus full coverage by carbamate indoor residual spraying; and universal LLIN coverage plus full coverage of carbamate-treated plastic sheeting lined up to household walls.

Caught mosquitoes were identified[17,18] by morphological characteristics.[19] All mosquitoes belonging to *An. gambiae* complex or *An. funestus* group were categorized by species with PCR.[17,18,20,21] In addition, *An. coluzzii* and *An. gambiae* were identified[18] by the methodology described in Favia et al.[22] The present work focuses on both *An. coluzzii* and *An. gambiae*, the main malaria vectors in the study area.

The Comité National d'Ethique pour la Recherche en Santé (Reference No. IRB00006860, Benin) And the Comité Consultatif de Déontologie et d'Ethique (Institut de Recherche pour le Développement) approved the study. Mosquitoes were only collected with the approval of village chiefs and all household dwellers. Mosquito collectors gave their written informed consent and were treated free of charge against presumed malaria illness throughout the study.

Data standardization

To focus on biting rate profiles, counts of both *An. coluzzii* and *An. gambiae* mosquitoes caught per site and per night were standardized, i.e., centred and reduced for each village. The distribution of the standardized-count variable has thus, mean zero (0) and variance 1, ensuring that only year-round biting rate profiles differed between villages and not biting rate intensities or the amplitudes of their variability.

Village classification

A latent class model with variable 'village' as random effect was built with standardized data.[23–25] This model took the sequence of observations into account through 'surveys' as a categorical variable. The likelihood function of the model was obtained as follows:

Let $Y_i=(y_{i1},y_{i2},\ldots,y_{i8},)$ denote a longitudinal sequence of standardized numbers of malaria vectors collected in a given village i over eight capture surveys, and $L(Y_i)$ denote the likelihood function of Y_i. Hence, $L(Y_i)=\sum_{j=1}^{J}\pi_j L^j(Y_i)$ where π_j is the probability of membership to group or class j and was linked to a set of parameters θ_j by $\pi_j=\exp(\theta_j)/\left(\sum_{j=1}^{J}\exp(\theta_j)\right)$, whose expression can allow for covariables taken as predictors of belonging to biting rate classes. Let $L^j(Y_i)$ denote the likelihood function of Y_i, given the membership of each village to group j so that $L^j(Y_i)=\prod_{t=1}^{8} f^j(y_{it})$, with f^j being the probability distribution function (here Gaussian) of Y_{it}, given the membership of village i to group j.

The likelihood of all the data was simply the product of values $L(Y_i)$. The model's estimated parameters were the result of maximum likelihood estimation through an Expectation-Maximization (EM) algorithm.[26] The 'mmlcr' function from the mmlcr library[27] in R software was used for this purpose. Based on Integrated Complete-data Likelihood criteria (ICL-BIC),[28] the number of groups was set at 2 ($J=2$) . Note that the number of groups chosen minimizes ICL-BIC, which is equivalent to low entropy[29] with maximum likelihood.

Spatial distribution of biting rate profiles

To illustrate the spatial distribution of both *An. coluzzii* and *An. gambiae* biting rate profiles over a year, a map of the area villages was created. Classification of villages was based on the maximum of a posteriori probabilities (MAP) of each village to belong to a given latent class. It allowed us to: investigate whether there are geographic zones whose villages share similar biting rate profiles; and to develop and check hypotheses about correlations between biting rate profiles and specific environmental factors.

Linkage between vector species and biting rate profiles of study villages

Fluctuations of *An. coluzzii* and *An. gambiae* ratios among villages were summarized based on groups of biting rate profiles. A mixed logistic regression model,[30] allowing for random intercept at the 'village' level and 'biting rate profiles' as fixed effect, was produced to check correlations between *An. coluzzii* proportions and biting rate profiles. In addition, the same model that provided village classification with the entire dataset was undertaken with separate *An. coluzzii* and *An. gambiae* data subsets. Village classifications

Downloaded from http://trstmh.oxfordjournals.org/ by guest on April 17, 2014

with these two models were respectively compared to model classification with the entire dataset.

Effects of environmental factors on the spatial distribution of biting rate profiles

As environmental factors are important in determining malaria vector biting rates,[1] we considered it worthwhile to check for significant linkage between them and the spatial distribution of biting rate profiles in the study area over a year. The village-linked factors considered were: mean annual rainfall (in mm) and number of rainy days: daily rainfall data from eight weather stations were interpolated spatially to compute cumulated rainfall and number of rainy days in each village during a 2-week period preceding each survey, to ascertain mean annual values; altitude (in meters); hydromorphous soil surface area (in km^2); distance to the nearest freshwater body (in km); and annual mean of the normalized difference vegetation index (NDVI): for each survey and each village, the average of 16-day NDVI (2-week period preceding survey plus 2-day survey) was extracted, to compute mean annual values. The NDVI was measured at a spatial resolution of 250 meters by the Moderate Resolution Imaging Spectroradiometer (MODIS) sensors.

Statistical calculations

Linkage between village classification per biting rate profile and each of the above-cited factors was assessed by non-parametric Wilcoxon–Mann–Whitney test.

Fisher's exact test ascertained linkage between village classification per biting rate profile and three factors linked to human activities: water conveyance (presence/absence); market gardening (presence/absence); and livestock breeding (yes/no).

The χ^2 test with Monte Carlo method for p-value estimation studied the relationship between biting rate profile and four vector-control strategies in the area.

The relationship between biting rate profile classes and average annual number of positive peri-domestic larval sites was evaluated by non-parametric Wilcoxon–Mann–Whitney test. $p<0.05$ values were considered to be statistically significant.

Factors found to be significantly correlated with biting rate profiles were introduced into the 2-latent-class model as predictors of belonging to latent classes, making it possible to measure their effects on village classification per biting rate profile. This was preceded by correlation testing between these factors to avoid the introduction of factors carrying the same information in the model.

Results

Entomology data

Over 3584 individual captures, 1872 *An. Coluzzii* and *An. gambiae* were caught (Figure 1), corresponding to a mean rate of 0.522 bites per person per night. Among these vectors, about 82% were *An. coluzzii,* with an average rate of 0.429 bites per person per night, while the corresponding rate for *An. gambiae* was 0.093 bites per person per night. The biting rate of each species was variable between villages and between surveys within each village (Supplementary Table 1).

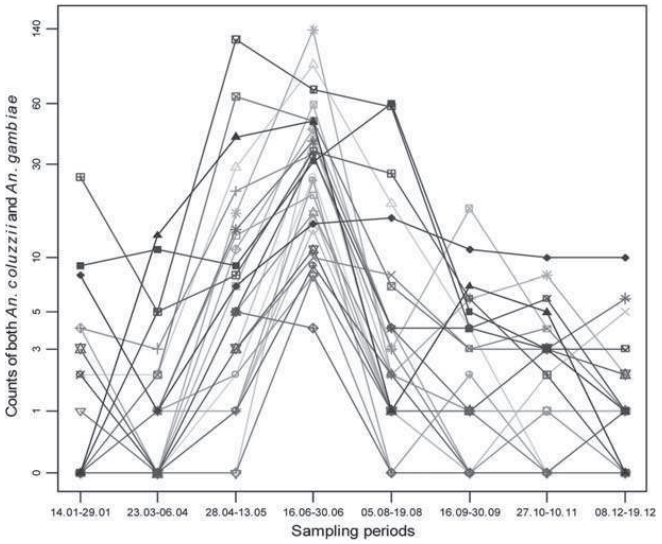


Figure 1. Changes in counts of both *Anopheles coluzzii* and *An. gambiae* caught over all eight surveys in each of 28 villages of the study area. Surveys took place according to sampling periods expressed as start-end dates (day.month) on the x-axis and logarithmic scale on the y-axis.

Seasonal changes in vector biting rate

In terms of *An. coluzzii* and *An. gambiae* biting rate, the latent class trajectory model revealed the existence of two village groups: Group A with 36% of villages, and Group B with 64% of villages. Figures 2A and 2B display the changes in vector biting rate in each village group over the study period.

The curves in Figure 2A correspond to Group A villages where vector activity seemed to be a bouncing or transient (seasonal) phenomenon whose maximum was reached in June, a period covering the onset of the rainy season. This activity included quiescent periods in most villages during much of the year.

Figure 2B shows the biting rate profiles in Group B villages. In the majority of villages in this group, the biting rate appeared to be continuous over much of the year: vectors were present and renewed throughout the year, with a maximum reached in June.

Spatial heterogeneity of vector biting rate profiles in the study area

Figure 3 depicts the distribution of both *An. coluzzii* and *An. gambiae* biting rate profiles in the 28 study villages. It reveals spatial heterogeneity of the vectors biting rate in the study area. Although most villages in the south seemed to be suitable for a continuous biting rate throughout the year, clustering was not clearly apparent.

Correlation between vector species and biting rate profiles of villages

Proportions at the village level (see Supplementary Table 1) of *An. coluzzii* in Group A (Group B respectively) ranged from 58 to 95% (from 33 to 100% respectively) with an average of 76% (85% respectively). Proportions of *An. gambiae* were obtained simply by 100 minus *An. coluzzii* proportions.

Downloaded from http://trstmh.oxfordjournals.org/ by guest on April 17, 2014

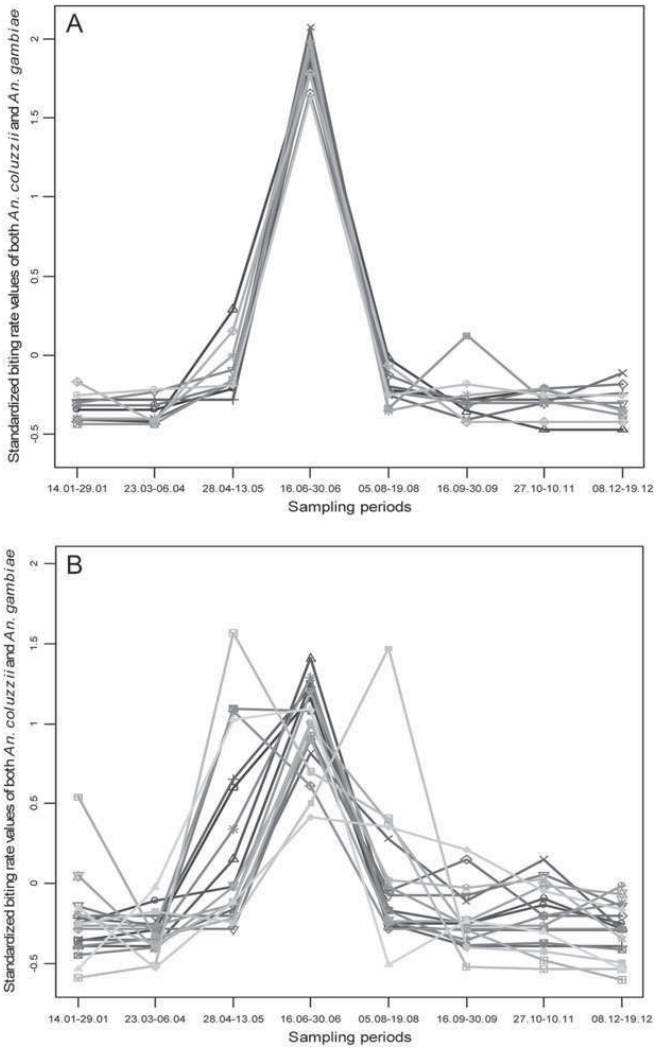


Figure 2. Biting rate profiles of both *Anopheles coluzzii* and *An. gambiae* in 28 villages of the study area. (A) Transient biting rate profile; (B) Continuous biting rate profile.

The mixed logistic regression model did not show a significant linkage between proportions of *An. coluzzii* and groups of biting rate profiles. Odds ratio (OR) (Group B versus Group A) was 1.40 (95% CI: 0.52, 3.72). Thus, correlation between vector species and the biting rate profiles of villages was not apparent.

Indeed, in terms of biting rate profiles, classifications were identical for villages with data combining the two vectors or *An. coluzzii* data alone (Supplementary Table 2 and Supplementary Figure 1). The biting rate profiles of three villages changed (two villages from 'transient' to 'continuous') when only *An. gambiae* data were analyzed (Supplementary Table 2 and Supplementary Figure 2).

Correlation between biting rate profiles and environmental factors

The Wilcoxon–Mann–Whitney test detected no significant linkage between biting rate profiles and distance to the nearest

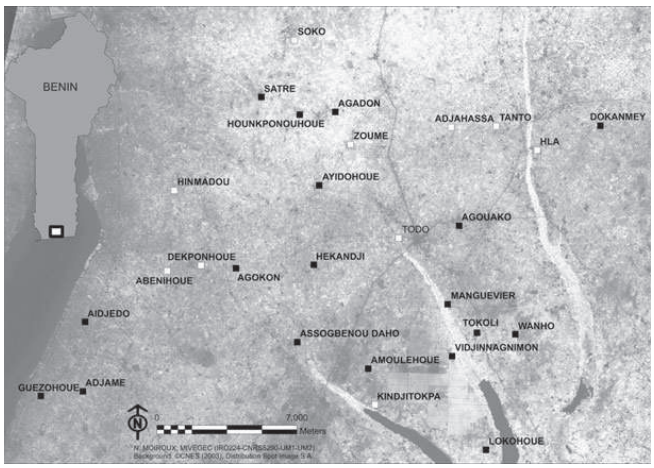


Figure 3. Spatial distribution of both *Anopheles coluzzii* and *An. gambiae* biting rate profiles in 28 study villages. Black squares: villages with a high posterior probability of belonging to the continuous biting rate group; Blank squares: villages with a high posterior probability of belonging to the transient biting rate group. The background of the map represents the Normalised Difference Vegetation Index (NDVI). Freshwaters and bare soils appear in dark grey; healthy vegetation appears in light grey or white.

freshwater body (p=0.12) or between biting rate profiles and the mean annual number of rainy days (p=0.58). Linkage between biting rate profiles and hydromorphous soil was not significant (p=0.35), although the surface areas of hydromorphous soil seemed to be more extensive in villages with transient biting rate profiles.

Linkage between biting rate profiles and altitude and linkage between biting rate profile and the NDVI were of borderline significance (p=0.06 in both cases). The vectors biting rate thus seemed to be continuous in low-altitude villages and transient in high-NDVI villages. Linkage between biting rate profile and mean annual rainfall was significant (p=0.007). As expected, villages with continuous vector biting rates had high annual rainfall. The association between the biting rate profile and the average annual number of positive peri-domestic larval sites was also significant (p=0.01). Thus, over the year, positive peri-domestic larval sites seemed to be more abundant in villages with continuous biting rate profile. Figure 4 depicts the distribution of mean annual rainfall, altitude, number of positive larval sites, and the NDVI according to biting rate profile. Fisher's exact test did not show significant linkage between biting rate profiles and either water conveyance, market gardening, or livestock breeding (p=0.12, 0.21 and 0.63, respectively).

No significant linkage between biting rate profile and vector-control strategies in the area was evident (χ^2=3.11, p=0.56).

Prediction of membership in each latent class of biting rate profile

The average amount of annual rainfall and average number of positive domestic breeding sites were positively correlated (p=0.008), as were altitude and vegetation index (p=0.003). Hence, only average rainfall and altitude were introduced into the model as predictor of membership to a latent class of biting

Downloaded from http://trstmh.oxfordjournals.org/ by guest on April 17, 2014

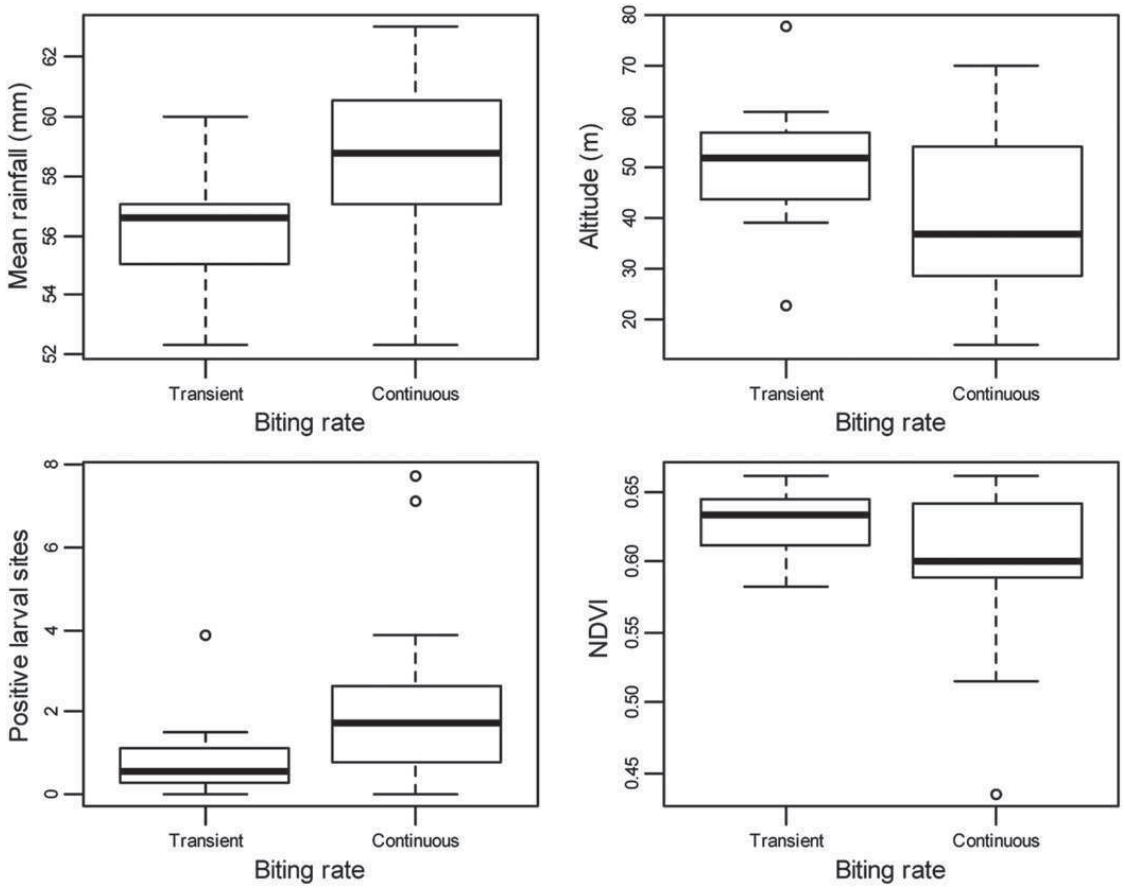


Figure 4. Relationships between both *Anopheles coluzzii* and *An. gambiae* biting rate profiles and four environmental factors.

rate profile. OR ('transient' class as reference) were 1.32 (95% CI: 0.96, 1.83) for each millimeter increase in rainfall and 1.58 (95% CI: 0.86, 2.90) for each 10-meter decrease in altitude. Despite these non-significant results, a millimeter increase in average annual rainfall would increase by 32% the OR of a village belonging to the 'continuous biting rate' class, whereas a 10-meter decrease in altitude would increase this OR by 58%.

Discussion

Knowledge of the spatio-temporal dynamics of malaria vector populations is an important element in malaria control and has been the subject of several studies.[4,14,31,32]

The present work focused on the *An. coluzzii* and *An. gambiae* biting rate profiles over a year. A 2-latent trajectory model was built. Villages could be classified into a group with 'transient' biting rate profile or another with 'continuous' biting rate profile.

Our results are close to those reported by Finkenstädt et al.[33] in a study on the spatio-temporal dynamics of measles in the UK. These authors disclosed the co-existence of two epidemiological settings: endemic episodes without obvious regularity separated by periods of disease extinction in small towns, which represents the transient vector biting rate in some villages; and epidemic episodes with some regularity and without disease extinction between episodes in large towns, which represents the continuous vector biting rate in other villages. In their study on measles, persistence of the epidemic in large towns was due to high birth rates that regularly supplied a susceptible population. Here, note that rainfall ensures the persistence of peri-domestic larval sites in villages with continuous vector biting rates.

Though MAPs stemming from the latent trajectory model were very high (Supplementary Table 2), representation of biting rate profiles indicates that certain villages were misclassified. This could explain why the classification of three villages changed (two villages from 'transient' to 'continuous') when analyzing only *An. gambiae* data instead of data combining the two species, as no significant correlation was found between vector species and biting rate profiles. The results of the model on only *An. gambiae* data could also be influenced by the small number of *An. gambiae* captured in the villages.

A previous study of the same area (Ouidah-Kpomassè-Tori, South Benin) was able to classify 28 villages according to mean vector biting rate intensity, given various environmental factors.[16] The present work presents another classification based on changes in vector biting rate profile around the year. The first classification was found to be linked to water conveyance, market gardening, and livestock breeding. The present changes in biting rate profile were not found to be linked to the same factors. Because

Downloaded from http://trstmh.oxfordjournals.org/ by guest on April 17, 2014

the changes in biting rate profile were not even over the area, one may infer that specific geological or geographical differences favour either continuous biting rates throughout the year or transient biting rates over a limited period. Among these differences, the present study pointed out only altitude and the NDVI. The lack of correlation between biting rate profile and the other factors considered in this work may be potentially explained by the small number (28) of villages.

A study of the same area by Cottrell et al.[14] demonstrated significant linkage between the NDVI and the spatio-temporal dynamics of Anopheles biting rates. Moiroux et al.[34] reported that the presence of both *An. coluzzii* and *An. gambiae* over the year was positively correlated with hydromorphous soil, whereas Cottrell et al.[14] obtained opposite results. The present work seems to be in agreement with the latter group; however, it did not find a significant linkage between hydromorphous soil and the vectors biting rate profile.

As expected, both *An. coluzzii* and *An. gambiae* biting rates fluctuate over the year[4] with rainfall: villages with high annual average rainfall were those showing a continuous biting rate profile. This result agrees with that of Moiroux et al.[34] who reported a positive correlation between the presence of malaria vectors and rainfall in the same area. The present work found that villages that had positive domestic breeding sites over the year were suitable for a continuous biting rate profile. This finding was reported in another way by Moiroux et al.[34] who noted a positive correlation between positive domestic breeding sites and the continuous presence of both *An. coluzzii* and *An. gambiae* over the year.

The present study undertook trajectory modeling to investigate the biting rate profiles of both *An. coluzzii* and *An. gambiae*. The same approach may be adopted to examine the spatio-temporal dynamics of other disease vectors. In the context of malaria, it allowed the classification of specific zones according to vector biting rate profiles. This and previous classifications according to mean biting rate intensity[16] are complementary indicators that may be targeted by healthcare organizations or epidemiological surveillance teams in regions where malaria is still a major public health problem. Conjointly, these indicators allow the mobilization of area- and season-specific malaria control measures.

Supplementary data

Supplementary data are available at Transactions Online (http://trstmh.oxfordjournals.org/).

Authors' contributions: RE, VC, NF and OB conceived and designed the study; AD and NM collected the data; OB, FS and RE analyzed and interpreted the data; OB, JI, FS and RE drafted the manuscript. OB, JI, FS, AD, NM, AG, JFE and RE critically revised the manuscript for intellectual content. All authors read and approved the final manuscript. OB and RE are guarantors of the paper.

Acknowledgements: We thank the populations and authorities of the Ouidah-Kpomassè-Tori district for their kind support and collaboration during data collection.

Funding: We thank the Ministère Français des Affaires Etrangères and Hospices Civils de Lyon for their financial support to Olayidé Boussari.

Competing interests: None declared.

Ethical approval: Not required.

References

1 Carnevale P, Robert V, Manguin S et al. Anopheles: biology, Plasmodium transmission and vector control [in French]. Marseille: IRD Editions; 2009.

2 Dossou-Yovo J, Doannio JMC, Diarrassouba S, Chauvancy G. The impact of rice fields on malaria transmission in the city of Bouaké, Côte d'Ivoire [in French]. Bull Soc Path Exot 1998;91:327–33.

3 Manguin S, Carneval P, Mouchet J. Biodiversity of malaria in the world. Montrouge, France: John Libbey Eurotext; 2008.

4 Amek N, Bayoh N, Hamel M et al. Spatial and temporal dynamics of malaria transmission in rural Western Kenya. Parasit Vectors 2012;5:86.

5 Gosoniu L, Vounatsou P, Sogoba N, Smith T. Bayesian modelling of geostatistical malaria risk data. Geospatial Health 2006;1:127–39.

6 Matthys B, Vounatsou P, Raso G et al. Urban farming and malaria risk factors in a medium-sized town in Côte d'Ivoire. Am J Trop Med Hyg 2006;75:1223–31.

7 Coluzzi M. Heterogeneities of the malaria vectorial system in tropical Africa and their significance in malaria epidemiology and control. Bull World Health Organ 1984;62(Suppl):107–13.

8 Coluzzi M. Malaria and the Afrotropical ecosystems: impact of man-made environmental changes. Parassitologia 1994;36:223–7.

9 Moiroux N, Gomez M, Pennetier C et al. Changes in *Anopheles funestus* biting behavior following universal coverage of long-lasting insecticidal nets in Benin. J Infect Dis 2012;206:1622–9.

10 White GB. *Anopheles gambiae* complex and disease transmission in Africa. Trans R Soc Trop Med Hyg 1974;68:278–98.

11 Hay SI, Snow RW, Rogers DJ. From predicting mosquito habitat to malaria seasons using remotely sensed data: practice, problems and perspectives. Parasitol Today 1998;14:306–13.

12 Coetzee M, Hunt RH, Wilkerson R et al. *Anopheles coluzzii* and *Anopheles amharicus*, new members of the *Anopheles gambiae* complex. Zootaxa 2013;3619:246–74.

13 Koenraadt CJM, Githeko AK, Takken W. The effects of rainfall and evapotranspiration on the temporal dynamics of *Anopheles gambiae* s.s. and *Anopheles arabiensis* in a Kenyan village. Acta Trop 2004;90:141–53.

14 Cottrell G, Kouwaye B, Pierrat C et al. Modeling the influence of local environmental factors on malaria transmission in Benin and its implications for cohort study. PloS One 2012;7: e28812.

15 Trape JF, Lefebvre-Zante E, Legros F et al. Vector density gradients and the epidemiology of urban malaria in Dakar, Senegal. Am J Trop Med Hyg 1992;47:181–9.

16 Boussari O, Moiroux N, Iwaz J et al. Use of a mixture statistical model in studying malaria vector density. PloS One 2012;7:e50452.

17 Corbel V, Akogbeto M, Damien GB et al. Combination of malaria vector control interventions in pyrethroid resistance area in Benin: a cluster randomised controlled trial. Lancet Infect Dis 2012;12:617–26.

18 Djenontin A, Bio-Bangana S, Moiroux N et al. Culicidae diversity, malaria transmission and insecticide resistance alleles in malaria vectors in Ouidah-Kpomasse-Tori district from Benin (West Africa): a pre-intervention study. Parasit Vectors 2010;3:83.

19 Gillies MT, De Meillon B, The Anophelinae of Africa south of the Sahara (Ethiopian zoogeographical region). Johannesburg: South African Institute for Medical Research; 1968.

Downloaded from http://trstmh.oxfordjournals.org/ by guest on April 17, 2014

20 Scott JA, Brogdon WG, Collins FH. Identification of single specimens of the Anopheles gambiae complex by the polymerase chain reaction. Am J Trop Med Hyg 1993;49:520–9.

21 Koekemoer LL, Kamau L, Hunt RH, Coetzee M. A cocktail polymerase chain reaction assay to identify members of the *Anopheles funestus* (Diptera: Culicidae) group. Am J Trop Med Hyg 2002;66:804–11.

22 Favia G, Lanfrancotti A, Spanos L et al. Molecular characterization of ribosomal DNA polymorphisms discriminating among chromosomal forms of *Anopheles gambiae s.s.* Insect Mol Biol 2001;10: 19–23.

23 McLachlan G, Peel D. Finite Mixture Models. Wiley: New York; 2000.

24 Nagin D. Group-based Modeling of Development. Cambridge, MA: Harvard University Press; 2005.

25 Skrondal A, Rabe-Hesketh S. Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models. Boca Raton, FL: CRC Press; 2004.

26 Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Series B Stat Methodol 1977;39:1–38.

27 Buyske S. The mmlcr Package. April 19, 2006. http://ftp.auckland.ac.nz/software/CRAN/doc/packages/mmlcr.pdf [accessed 5 September 2013].

28 Biernacki C, Celeux G, Govaert G. Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Trans Pattern Anal Mach Intell 2000;22:719–25.

29 Nagin DS, Odgers CL. Group-based trajectory modeling in clinical research. Annu Rev Clin Psychol 2010;6:109–38.

30 Bates D, Maechler M, Bolker B, Walker S. lme4: Linear mixed-effects models using Eigen and S4. R package version 1.0-5. 2013. http://cran.r-project.org/package=lme4 [accessed 26 September 2013].

31 Parham PE, Pople D, Christiansen-Jucht C et al. Modeling the role of environmental variables on the population dynamics of the malaria vector *Anopheles gambiae sensu stricto*. Malar J 2012;11:271.

32 Ruiz D, Poveda G, Velez ID et al. Modelling entomological-climatic interactions of Plasmodium falciparum malaria transmission in two Colombian endemic-regions: contributions to a National Malaria Early Warning System. Malar J 2006;5:66.

33 Finkenstädt BF, Bjørnstad ON, Grenfell BT. A stochastic model for extinction recurrence of epidemics: estimation, inference for measles outbreaks. Biostatistics 2002;3:493–510.

34 Moiroux N, Bio-Bangana AS, Djenontin A et al. Modelling the risk of being bitten by malaria vectors in a vector control area in southern Benin, west Africa. Parasit Vectors 2013;6:71.

Downloaded from http://trstmh.oxfordjournals.org/ by guest on April 17, 2014

Supplementary files

Supplementary Table 1. Proportion of *Anopheles coluzzii* caught per village and per village survey. (Proportions of *An. gambiae* can be derived by 100 minus proportions of *An. coluzzii*.)

Village	Biting rate profile	Number of *An. gambiae* and *An. coluzzii* collected	Proportion of *An. coluzzii* (%)	Proportion of *An. coluzzii* (%) per sampling period							
				14.01 - 29.01	23.03 - 06.04	28.04 - 13.05	16.06 - 30.06	05.08 - 19.08	16.09 - 30.09	27.10 - 10.11	08.12 - 19.12
Adjame	continuous	39	89.74	NA	100	81.82	90.48	100	100	100	100
Agouako	continuous	12	58.33	NA	0	50	75	NA	NA	0	NA
Aidjedo	continuous	29	93.10	NA	NA	100	100	50	100	66.67	100
Amoulehoue	continuous	66	86.36	75	100	100	78.12	100	NA	NA	NA
Assogbenou	continuous	29	79.31	NA	NA	100	70	100	100	25	100
Ayidohoue	continuous	10	100	NA	NA	NA	100	100	100	NA	NA
Adjahassa	transient	31	87.10	NA	NA	50	88.46	100	100	100	NA
Dokanmey	continuous	15	66.67	0	NA	NA	100	100	100	0	0
Kindjitokpa	transient	149	95.30	50	100	100	94.57	100	100	NA	NA
Vidjinnagnimon	continuous	130	89.23	NA	50	92.59	95.92	83.33	33.33	33.33	100
Guezohoue	continuous	68	100	NA	100	100	100	100	100	100	100
Hekandji	continuous	9	33.33	NA	NA	0	75	NA	NA	NA	NA
Hinmadou	transient	47	80.85	NA	NA	NA	80	100	NA	NA	100
Hla	transient	52	84.62	NA	NA	50	89.19	50	NA	100	80
Agokon	continuous	13	92.31	NA	NA	NA	100	100	NA	NA	0
Hounkponouhoue	continuous	24	41.67	0	NA	0	77.78	100	0	0	0
Abenihoue	transient	18	77.78	NA	NA	100	83.33	100	NA	0	0
Dekponhoue	transient	20	85	NA	100	NA	82.35	100	NA	NA	NA
Manguevie	continuous	109	96.33	88	100	100	100	96.15	100	100	NA
Satre	continuous	55	60	0	NA	20	78.79	25	100	0	NA
Soko	transient	92	59.78	NA	NA	33.33	78.57	100	0	25	100
Tanto	transient	172	58.14	NA	NA	26.67	57.63	50	100	100	33.33
Lokohoue	continuous	267	94.76	NA	80	100	95.59	88.89	100	100	100
Todo	transient	64	90.63	75	NA	81.82	93.18	100	NA	NA	NA
Tokoli	continuous	129	82.17	27.27	100	62.5	87.50	95	60	33.33	0
Wanho	continuous	78	100	100	100	100	100	100	100	100	100
Agadon	continuous	116	61.21	NA	100	13.51	100	100	14.29	0	NA
Zoume	transient	29	68.97	NA	100	NA	70.83	NA	50	NA	NA

Villages highlighted in yellow: no *An. gambiae* caught; green: changed biting rate profile (continuous to transient) when only *An. gambiae* data were analyzed; cyan: changed biting rate profile (transient to continuous) when only *An. gambiae* data were analyzed.

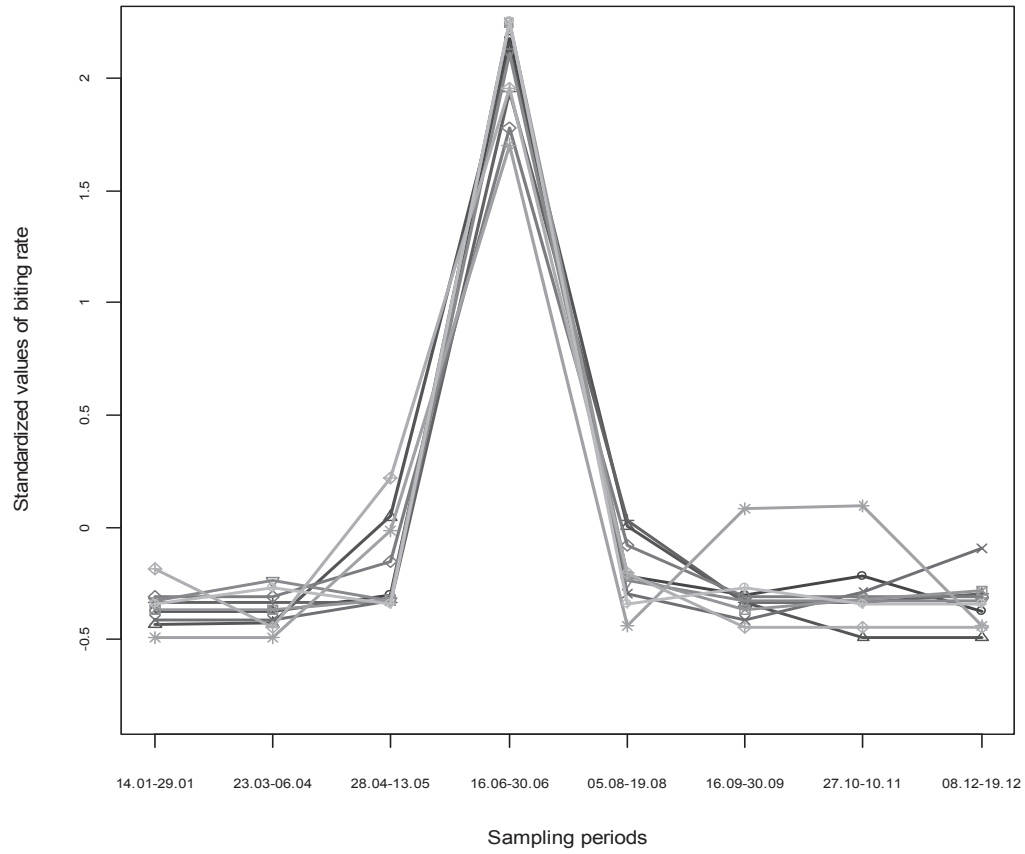
Supplementary Table 2. Villages biting rate profiles and posterior probabilities of belonging to each biting rate profile when were analyzed respectively the entire dataset, *Anopheles coluzzii* data subset and *An. gambiae* data subset.

Village	Both *An. coluzzii* and *An. gambiae* data			*An. coluzzii* data			*An. gambiae* data		
	Posterior probability of belonging to		Biting rate profile	Posterior probability of belonging to		Biting Rate profile	Posterior probability of belonging to		Biting rate profile
	Transient class	Continuous class		Transient class	Continuous class		Transient class	Continuous class	
Adjame	0	1	continuous	0	1	continuous	0.209	0.791	continuous
Agouako	0	1	continuous	0	1	continuous	0.011	0.989	continuous
Aidjedo	0.005	0.995	continuous	0.005	0.995	continuous	0	1	continuous
Amoulehoue	0	1	continuous	0	1	continuous	0.810	0.190	transient
Assogbenou	0	1	continuous	0	1	continuous	0.002	0.998	continuous
Ayidohoue	0	1	continuous	0	1	continuous	NA	NA	NA
Adjahassa	1	0	transient	1	0	transient	0.994	0.006	transient
Dokanmey	0	1	continuous	0	1	continuous	0	1	continuous
Kindjitokpa	1	0	transient	1	0	transient	0.975	0.025	transient
Vidjinnagnimon	0	1	continuous	0	1	continuous	0	1	continuous
Guezohoue	0	1	continuous	0	1	continuous	NA	NA	NA
Hekandji	0	1	continuous	0	1	continuous	0	1	continuous
Hinmadou	1	0	transient	1	0	transient	1	0	transient
Hla	1	0	transient	1	0	transient	0.962	0.038	transient
Agokon	0	1	continuous	0	1	continuous	0.005	0.995	continuous
Hounkponouhoue	0	1	continuous	0	1	continuous	0	1	continuous
Abenihoue	0.999	0.001	transient	0.999	0.001	transient	0.215	0.785	continuous
Dekponhoue	1	0	transient	1	0	transient	0.999	0.001	transient
Manguevie	0	1	continuous	0	1	continuous	0	1	continuous
Satre	0	1	continuous	0	1	continuous	0.001	0.999	continuous
Soko	1	0	transient	1	0	transient	0.774	0.226	transient
Tanto	1	0	transient	1	0	transient	1	0	transient
Lokohoue	0	1	continuous	0	1	continuous	0	1	continuous
Todo	1	0	transient	1	0	transient	0.194	0.806	continuous
Tokoli	0	1	continuous	0	1	continuous	0	1	continuous
Wanho	0	1	continuous	0	1	continuous	NA	NA	NA
Agadon	0	1	continuous	0	1	continuous	0	1	continuous
Zoume	0.998	0.002	transient	0.998	0.002	transient	1	0	transient

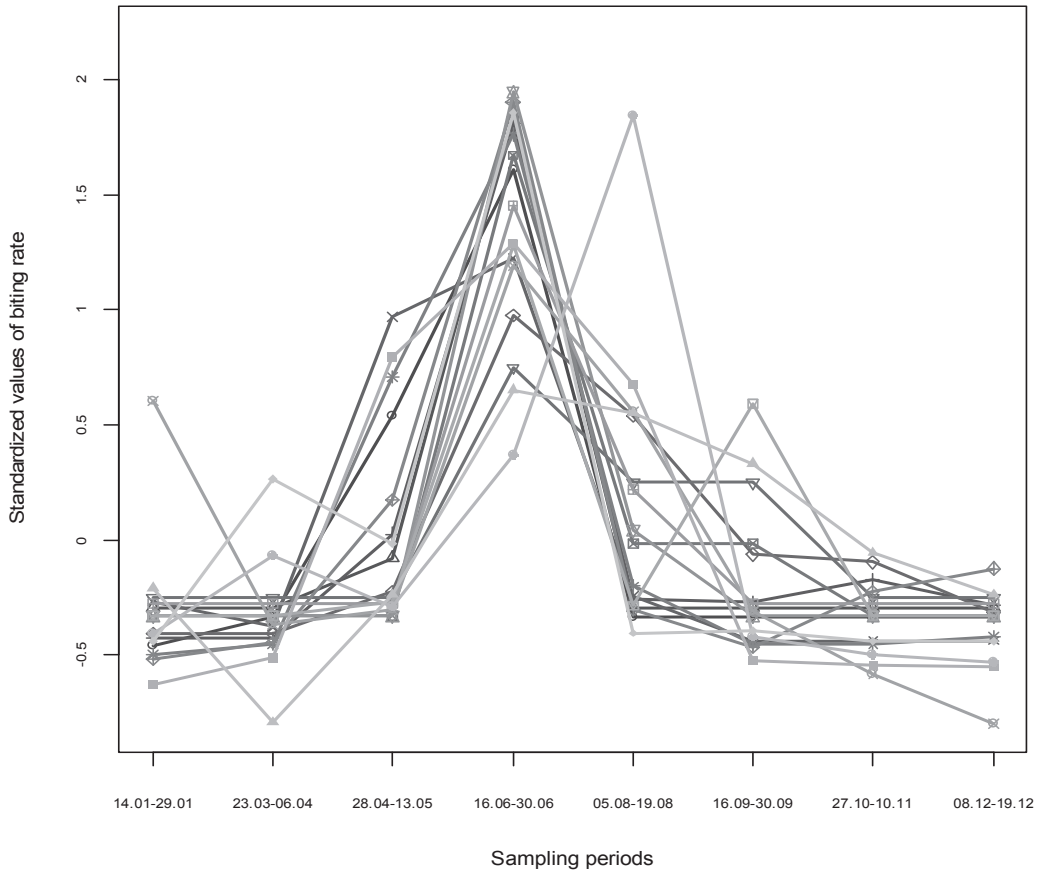
Villages highlighted in yellow: no *An. gambiae* caught; green: changed biting rate profile (continuous to transient) when only *An. gambiae* data were analyzed; cyan: changed biting rate profile (transient to continuous) when only *An. gambiae* data were analyzed.

Supplementary Figure1: *An. coluzzii* biting rate profile in the two groups of villages

1A Transient group

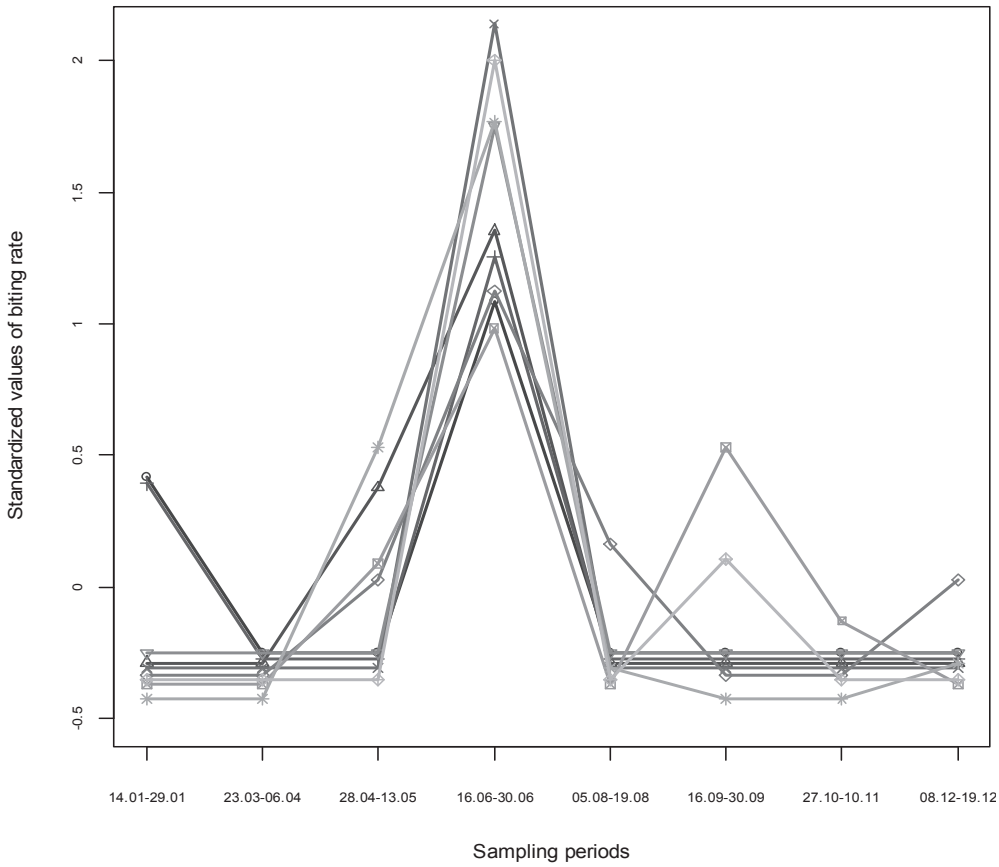


1B Continuous group

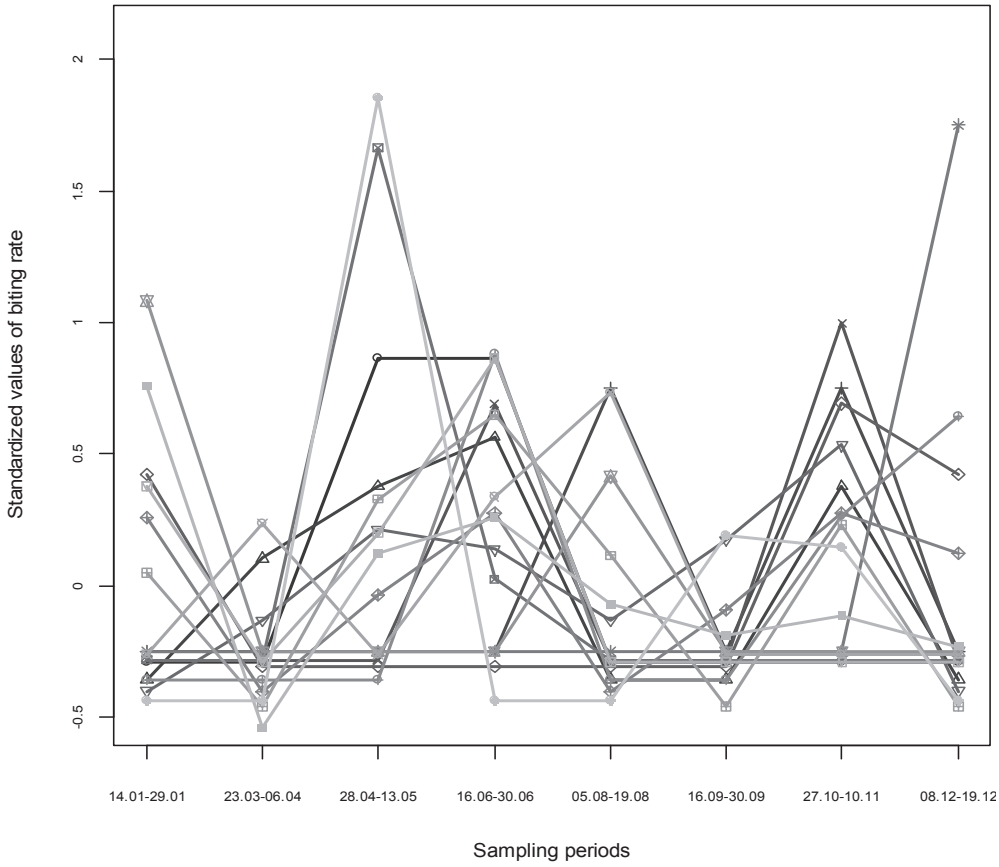


Supplementary Figure2: *An. gambiae* biting rate profile in the two groups of villages

2A Transient group



2B Continous group



Chapitre 6

Variabilité de l'observance aux traitements antirétroviraux

Ce chapitre quitte le cadre de la dynamique spatiale et temporelle des vecteurs de paludisme au cœur des deux précédents chapitres pour s'intéresser à un problème lié à l'observance aux traitements antirétroviraux dans le cadre du VIH/SIDA. De ce point de vue, le chapitre 6 peut sembler en rupture avec les chapitres 4 et 5. Cependant, nous verrons que la problématique de santé à laquelle il répond dépasse le cadre du VIH/SIDA et peut aisément se poser aussi bien dans le cadre du paludisme que pour d'autres maladies. Aussi, la base du développement méthodologique adopté pour le traitement des données objet de ce chapitre, reste les modèles à classes latentes.

6.1 Retour au problème posé sur la base d'un exemple

Considérons deux patients A et B soufrant d'une même pathologie et soumis à un même traitement. Admettons que le traitement consiste en une prise journalière d'un certain nombre de comprimés sur une période de 30 jours et définissons l'observance journalière au traitement comme étant égale au quotient du nombre de comprimés pris sur le nombre total de comprimés prescrits pour une journée. Les observances des patients A et B sur les 30 jours sont rapportées en pourcentages sur le graphique de la Figure 6.1.

Les observances moyennes mensuelles des patients A et B sont presque identiques (59.47% et 58.34% respectivement). En revanche, le patient B est caractérisé par une grande instabilité (ou variabilité) en termes d'observance comparativement au patient A.

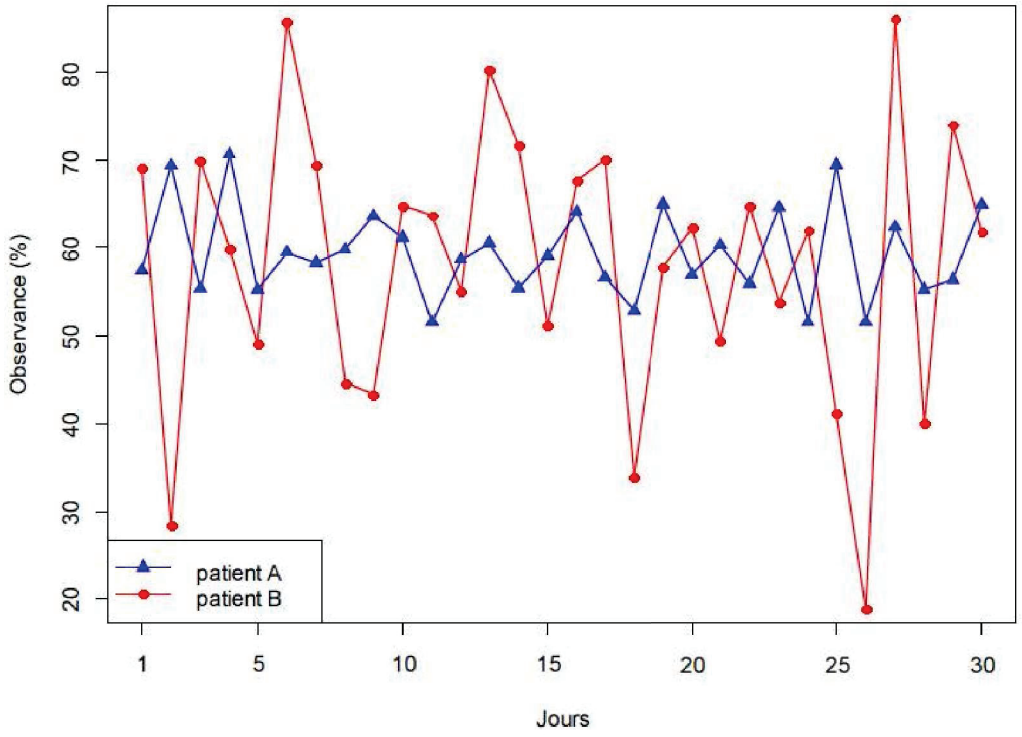


FIGURE 6.1 – Observances journalières au même traitement par deux patients

La variance des observances du patient B est presque dix fois plus importante que celle des observances du patient A soit un écart type de 16.15% pour le patient B contre 5.28% pour le patient A. Une question intéressante pour le praticien serait de savoir si les processus de rétablissement des patients A et B sont identiques. Autrement dit, à observance moyenne égale, quel est l'impact de la variabilité dans l'observance sur l'efficacité du traitement reçu par les patients A et B ? Un questionnement sous jacent au premier serait de savoir si cet impact de la variabilité aurait la même ampleur dans le cas où les moyennes d'observance chez les deux patients étaient beaucoup plus faibles par exemple.

Dans le cadre du traitement contre le VIH, plusieurs études ont montré le lien

entre le niveau moyen de l'observance au traitement antirétroviral et la réponse immuno-virologique ou la mortalité. Nous investiguons dans ce chapitre l'impact de la variabilité dans l'observance sur les mêmes indicateurs tout en prenant en compte celui de l'observance moyenne. Les données présentées dans la section 4 du chapitre 2 ont servi de cadre d'applications. L'observance a été mesurée chez plusieurs patients VIH de façon mensuelle ou bimensuelle sur plus de neuf ans. Sur le plan méthodologique, il s'agit comme au chapitre précédent, d'un problème de modélisation de données longitudinales. Mais ici, il a été question de définir d'abord une mesure de la moyenne et de la variance d'observance. Pour chaque patient, elles correspondent, pour chaque mois de mesure (à partir du douzième mois), à la moyenne et à la variance des pourcentages d'observance mesurés sur les douze mois précédents. Il s'agit donc d'une moyenne et d'une variance mobiles. L'idée de considérer des valeurs des mois précédents pour calculer la moyenne et la variance d'observance du mois en cours vient du fait que l'état actuel du patient est une résultante de ces comportements (en termes d'observance) au cours des mois précédents. Cependant il faut remarquer que les deux statistiques ainsi obtenues sont très liées. Par exemple à une bonne observance moyenne est associée en général une stabilité de cette observance. Il a donc été nécessaire de définir une métrique de la variance d'observance qui la dissocie de la moyenne d'observance.

Cette métrique est construite comme suit : Les moyennes mobiles obtenues ont été ordonnées de manière croissante et groupées en des classes d'amplitudes relativement faibles et égales. Les variances mobiles correspondantes ont ensuite été ordonnées de manière croissante au sein de chaque classe de moyennes mobiles. La fonction de répartition empirique des variances à l'intérieur de chaque classe de moyennes est obtenue en divisant les variances ordonnées par la taille de la classe augmentée de 1. Les quantiles de cette fonction de répartition sont ensuite assimilés à ceux d'une distribution normale. On obtient ainsi une métrique de la variance d'observance (variance standardisée) indépendante de la moyenne d'observance. La Figure 6.2 illustre bien que le lien observé entre la moyenne et la variance avant

standardisation disparaît après standardisation de la variance.

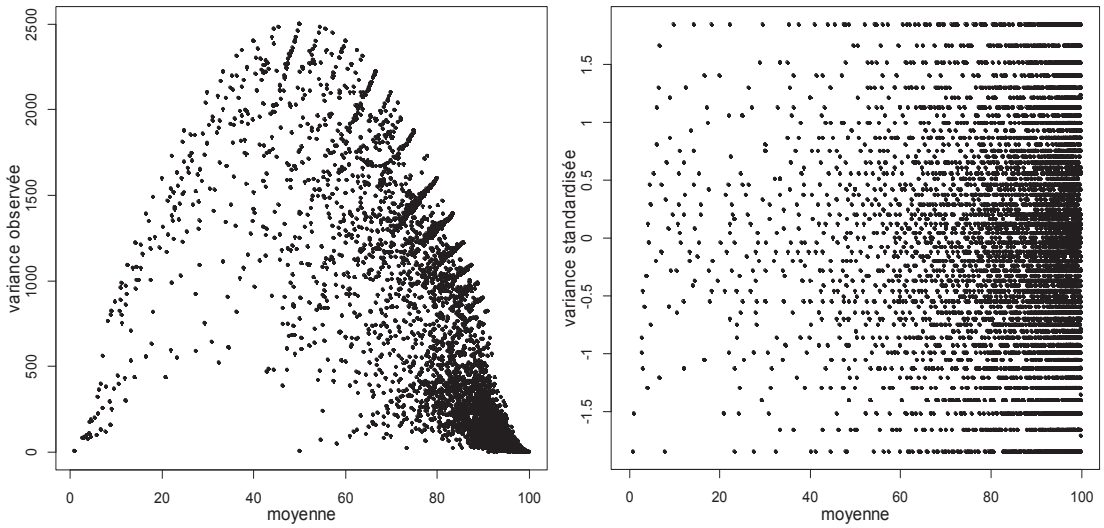


FIGURE 6.2 – Variances mobiles avant et après standardisation en fonction des moyennes mobiles

D'un point de vue pratique, il semble intéressant de travailler en termes de trajectoires typiques de moyenne et de variance d'observance afin de pouvoir catégoriser les patients. Ceci rendrait en effet une flexibilité aux résultats qui seraient alors facilement exploitables par les praticiens.

Un modèle à classes latentes de trajectoires de moyenne d'observance et un autre de la variance d'observance ont alors été construits. Les patients ont été ensuite classifiés (sur la base des MAP) suivant les deux caractéristiques de l'observance (moyenne et variance) avant d'établir le lien de ces caractéristiques avec la reconstitution immunitaire (variation du taux de CD4 dans le sang), l'élimination des virus et la survie des patients.

Il est important de souligner que le nombre de classes latentes de trajectoires

a été arbritrairement fixé à 3 (trois) dans le modèle relatif à la moyenne d'observance comme dans celui relatif à la variance d'observance. Ce choix ne tient donc pas compte des critères classiques (AIC, BIC, ICL) mais est fortement motivé à la fois par le souci de limiter la complexité du modèle qui induirait une complexité d'interprétabilité chez le praticien et par le souci de montrer une tendance dans la moyenne et la variance d'observance en évitant une dichotomisation de l'ensemble des trajectoires observées dans chaque cas. Toutefois, ce nombre de classes (trois) a permi d'obtenir des résultats assez satisfaisants en termes d'adéquation du modèle aux données. Par exemple, dans les deux cas (trajectoires de moyennes/variances d'observance), les MAP sont assez proches de 1 pour la quasi-totalité des individus et les proportions des groupes estimées par le modèle sont presqu'identiques à celles des groupes obtenues sur la base des MAP.

Les détails de ce travail sont présentés dans la section suivante sous forme d'un article soumis à la revue JAIDS pour publication.

6.2 Article 3 : Impact of the variability in adherence to antiretroviral treatment on the immunovirological response and mortality

Title:

Impact of the variability in adherence to antiretroviral treatment on the immunovirological response and mortality

Short Title:

Impacts of changes in adherence to HAART

Olayidé BOUSSARI [a-e], Fabien SUBTIL [b-e], Christophe GENOLINI [f,g], Mathieu BASTARD [h], Jean IWAZ [b-e], Noël FONTON [a], Jean-François ETARD [i] and René ECOCHARD [b-e] for the ANRS 1215 study group*

[a] International Chair in Mathematical Physics and Applications, Laboratoire d'Etude et de Recherche en Statistique Appliquée et Modélisation, Université d'Abomey-Calavi, Abomey-Calavi, Bénin.

[b] Hospices Civils de Lyon, Service de Biostatistique, F-69003 Lyon, France.

[c] Université de Lyon, F-69000, Lyon, France.

[d] Université Lyon 1, F-69100, Villeurbanne, France.

[e] CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, Equipe Biostatistique-Santé, F-69100, Villeurbanne, France.

[f] INSERM, UMR 1027, Research Unit on Perinatal Epidemiology and Childhood Disabilities, Adolescent Health, Toulouse F-31062, France; Université Paul Sabatier, UMR 1027, Toulouse F-31062, France.

[g] CeRSM (EA 2931), UFR STAPS, Université de Paris Ouest-Nanterre-La Défense, France.

[h] Epicentre, Paris, France.

[i] UMI 233 TransVIHMI, Institut de Recherche pour le Développement, Université Montpellier 1, Montpellier, France.

Corresponding author:

Telephone: (+33) 4 72 11 57 58; Fax: (+33) 4 72 11 51 41;

E-mail address: olayideb@yahoo.fr

Abstract

Objective: Investigate the impact of variability in adherence to HIV antiretroviral treatment on the viral load, the CD4 cell count, and mortality.

Methods: The study concerned HIV-1 infected patients enrolled in the Senegalese ART Initiative (August 1998 to April 2002) and analyzed data on monthly adherence data in 317 among them (November 1999 to April 2009). Latent-class trajectory models were used to build typical trajectories for the average adherence and the standardized variance of adherence. The relationships between the standardized variance of adherence and each of the change in CD4 cell count, the change in viral load, and mortality were established using, respectively, a mixed linear regression, a mixed logistic regression, and a Cox model with time-dependent covariates. All the models were adjusted on the average adherence.

Results: Three trajectories for the average adherence and three trajectories for the standardized variance of adherence were identified. Increases in the CD4 cell count and in the percentage of undetectable viral loads were negatively associated with the standardized variance of adherence but positively associated with the average adherence. The risk of death decreased significantly with the increase in the average adherence and increased significantly with the increase of the standardized variance of adherence.

Conclusion: These impacts of the variability in adherence on the immunovirological response and on mortality can be used, first in the assessment of various HAART regimen, then in the therapeutic education of people living with HIV.

Key words: antiretroviral therapy; adherence variability; latent trajectory modeling; classification

Introduction

The advent of highly active antiretroviral therapy (HAART) two decades ago has improved the health status of many people living with HIV and has significantly reduced HIV-linked death rates [1,2]. To date, several studies have shown the relationship between adherence to HAART and each of plasma viral load, CD4 recovery, the progress toward AIDS, or mortality [3-8] but the impact of the variability in the adherence to HAART on the immunovirological response does not seem to have been studied.

For a number of various reasons, contrarily to high-income countries, the rate of access to HAART in most Sub-Saharan African countries has been low over a very long period [9,10]. Many efforts have been recently devoted in these countries to substantially improve that access.

Senegal has been one of the first Sub-Saharan African countries to have a public policy of access to HAART. Indeed, the Senegalese initiative of access to antiretroviral therapy (Initiative Sénégalaise d'Accès aux Antirétroviraux, ISAARV) was launched in 1998 [11]. Not long after, an operational research project was designed to follow-up the patients, evaluate the level of adherence to HAART, and find the reasons of non-adherence [12]. ISAARV and the follow-up project have been the object of several studies on various time periods. These studies analyzed the determinants of adherence, the levels of adherence, or the link between the level of adherence and the immunovirological response or mortality [13-15].

The present work examines first not only the progress of the level of adherence to HAART (i.e., the average adherence) but also the variability of this adherence over time. It examines then the impact of this variability on the viral load, the CD4 cell count, and mortality with adjustment on the average adherence.

Material and Methods

The data source

The original dataset concerned the ANRS 1215 cohort that included 404 patients with HIV-1 infection and prescribed HAART within the context of ISAARV [12,16]. These patients were included between August 1998 and April 2002. The detailed criteria for inclusion may be found in previous studies on ISAARV cohort [12,15,16]. In short, 80 patients were enrolled between January 2000 and April 2001 if they were HAART naïve and had a CD4 cell count < 350 cells/mL and a plasma viral load > $3x10^4$ copies/mL. The 324 others were enrolled between August 1998 and April 2002 if they had < 350 CD4 cells/mL (< 200 CD4/mL after October 2000) and a plasma viral load > 10^5 copies/mL (asymptomatic patients) or > 10^4 copies/mL (paucisymptomatic patients); symptomatic patients free from major opportunistic infections were included whatever the CD4 cell count or the plasma viral load.

The examination of adherence to HAART started on November 1999 for the first 180 patients enrolled in ISAARV and on May 2004 for the 224 others. The data relative to adherence to HAART were complete for 330 patients; the others died before these data collection. Furthermore, for 13 patients, the data were irregular and unsuitable for the analysis. The final analysis concerned then 317 patients: 175 women and 142 men. The mean age was 37.5 years with 31 - 43 years as interquartile range (IQR). The time on HAART was censored at 108 months. The median time on HAART was 92 months (IQR: 84 -105 months).

The patients were seen two weeks, one month, and two months after HAART initiation then at least every two months. They had to obtain the drugs from the same centre (Fann Hospital, Dakar). At each drug delivery, the pharmacist estimated adherence by counting the number of remaining pills and by interviewing the patients about the reasons of non-adherence. The adherence to each drug was calculated as the number of pills taken divided by the number of pills prescribed over the last month. The overall adherence over the last 30 days was the arithmetic mean of the distinct drug adherences.

Besides, every six months, each patient had laboratory investigations that included a plasma viral load and a CD4 cell count. The detailed investigations can be found in a previous publication [16].

Ethics

The study was conducted with the approval of the Senegalese Ministry of Health (Conseil National de la Recherche en Santé No. 0017 MSP/DS/CNRS and Direction de la Santé No. 0760 MSP/DS/CNRS). All the patients gave written informed consents for participation in the study.

Statistical analyses

All data were censored at death or last visit before the end of the 108th month after HAART initiation.

Starting from the 12th, a monthly moving average and a monthly moving variance of adherence to HAART were calculated using at each time point the adherences of twelve previous months. These two values are highly linked: a high average adherence is usually associated with a low variance of adherence; thus, they carry the same information concerning the impact of adherence on the immunovirological response. A measure of variance that is independent of the average was then necessary and was conceived as follows: the averages were first sorted by increasing order then grouped in 604 small-amplitude classes of same-size (~30 averages). The corresponding variances were sorted in increasing order within each class. The empirical repartition function of the variances within each class of averages was obtained by dividing the ordered variances by 1+ the class size. The quantiles of this empirical repartition function were then assimilated to the quantiles of a normal distribution. This provided standardized variances of adherence to HAART.

A latent-class trajectory model was used to distinguish typical trajectories [17-19] of the moving averages of adherence to HAART. In the model, the number of latent-classes was arbitrarily fixed to three, which is sufficient to show the trends and limits the complexity of the model. This model considers time as a random effect and a random intercept at "patient" level. It may be specified as follows:

Let $\hat{y}_{it}$ be the average adherence predicted for patient i for month t (since HAART initiation). With $T \in \{12, 24, 36, 48, 60, 72, 84, 96\}$, let h_T be a function defined by $h_T(t) = \begin{cases} t - T \text{ when } t > T \\ 0 \text{ otherwise} \end{cases}$. Then $\hat{y}_{it} = \sum_{j=1}^{J} \left[p_{ij} \times f^j(t) \right]$ with p_{ij} as the posterior probability for patient i to belong to group j and $f^j(t) = \beta_{0j} + \beta_{1j} \times t + \sum_T \left[\beta_{Tj} \times h_T(t) \right]$ the mean of a Gaussian distribution conditionally on the fact that i belong to group j. The

value attributable to typical trajectory j at month t is thus the mean of $f^{j}(t)$ weighted by p_{ij} values.

A typical trajectory of average adherence was attributed to each patient according to the maximum a posteriori probability (MAP) rule [19]. The same model was also applied to the standardized variances of adherence to obtain typical trajectories of variance in adherence and then a classification of the patients according to these typical trajectories.

To investigate the link between adherence to HAART and the viral load, percentages of patients with undetectable viral loads (<1000 copies/mL of blood) were calculated per six-month periods, for each average-standardized variance pair. The link between these percentages and adherence was explored using a mixed logistic regression that considered the classes of adherence averages and the standardized variance as well as their interaction as fixed effects and the six-month period as random intercept.

To investigate the link between adherence to HAART and the CD4 cell count, a semestrial rate of change (increase or decrease) in CD4 cell count was calculated for each average-standardized variance pair. These rates of change were transformed into monthly values (by dividing by 6). The link between the latter values and adherence was explored using a mixed linear model that included the classes of adherence averages and the standardized variance as well as their interaction as fixed effects and "semester" as random intercept.

To investigate the link between adherence to HAART and mortality, the values of the moving averages and standardized variances of adherence were taken as explanatory variables in a time-dependent Cox model [20-22]. The change in patient adherence (i.e., the mean and the standardized variance) was not considered constant over time.

All statistical analyses were performed using packages from R (version 3.0.2) software [23].

Results

Adherence

Fig. 1 (left panel) shows the three typical trajectories of the average adherence to HAART: i) one trajectory of "constantly high" (cH) average that groups 69.47% of all the moving averages and whose average is close to 95%; ii) another trajectory of "high but slowly decreasing" (HsD) average that groups 17.36% of the moving averages and whose average ranges from 90% (up to the fourth year) to 65% (on the seventh and eighth year); iii) a third trajectory of "decreasing then rapidly increasing" (DrI) average that groups the remaining 13.17% of the moving averages and whose average falls sharply from 80% over the first year to less than 50% on the third year before increasing up to 90% over the seventh and eighth year.

Fig. 1 (right panel) shows the three typical trajectories of the standardized variance of adherence to HAART. Though irregular along time, these trajectories are somewhat ordered. They may be labelled "low" (22.59% of the moving variances), "moderate" (49.08% of the patients) and "high" (28.32% of the moving variances).

Table 1 shows the distribution of the 317 patients in the nine average / standardized variance groups according to the MAP rule. About 98% (respectively, 90%) of the MAPs are higher than 0.85 regarding average adherence (respectively, standardized variance of adherence) latent trajectories.

Among the 221 patients with cH average adherence, nearly six out of ten patients have a moderate standardized variance. The distribution of the patients of the two other groups of average adherence over each of the standardized variance groups was overall rather homogeneous.

Impact of the variability in adherence to HAART on the viral load

Fig. 2 (panel a) shows the mean percentages of undetectable viral loads according to the nine average / standardized variance groups. The calculation of these means did not include the percentages found at HAART initiation which were nearly 0 in all groups. The mean percentage increases together with the average adherence. The significance of this association is confirmed by the results of the mixed logistic regression (Table 2).

Table 2 shows clearly that the probability of undetectable viral load increases when the status shifts from DrI average adherence to HsD average adhrence (OR=2.88; 95% CI: 1.52–5.45) or to cH average adherence (OR=2.91; 95% CI: 1.74–4.86).

Within each group of average adherence, there seems to be an association between the mean percentage of undetectable viral load and the standardized variance in adherence. However, this association has not the same effect size within the three groups. Indeed, in the group of cH average adherence, the odds of the viral load increases by 76% (OR=1.76; 95% CI: 1.35–2.30) when the status shifts from "High" to "Moderate" variance and by 82% (OR=1.82; 95% CI: 1.25–2.66) when the status shifts from "High" to "Low" variance (Table 2, bottom rows). In the group DrI average, the associations are the same though not significant but the effect sizes are rather important.

The association between the standardized variance of adherence to HAART and viral load undetectability was less important within HsD average (Moderate vs. high variance OR = 0.69; 95% CI: 0.37–1.28; Low vs. High variance OR = 1.71; 95% CI: 0.85–3.45).

Impact of the variability in adherence to HAART on the CD4 cell count

Fig. 2 (panel b) shows the distributions of the rates of change (increase or decrease) of the CD4 cell count according to the nine average / standardized variance groups. Within the groups of cH average and HsD average adherence, more than 75% of the rates were positive; i.e., were increasing. Within the group DrI average, the IQRs of the rates include negative values (i.e., decreasing rates). In other words, the higher is average adherence, the higher is the rate of change of the CD4 cell count, which means that the higher is the average adherence, the faster is the immunologic recovery.

Besides, Fig. 2 (panel b) shows that, in any given average adherence group, the rate of change of the CD4 cell count tends to decrease with the increase of the standardized variance of adhesion to HAART. This trend is very clear in group DrI average, which could mean that the higher is the standardized variance of adherence, the slower is the immunologic recovery.

Table 3 shows the results of the mixed linear model regarding the trends of the monthly CD4 cell count. This trend remains positive on average whatever the average-standardized variance combination, which means an overall increase of the monthly mean of CD4 cell count. This increase is significantly more important when the status shifts from DrI average to cH average than to HsD average (the differences being 4.70 and 4.28 CD4 cells/mm^3 per month, respectively). There is a significant relationship between the increase in the monthly mean CD4 cell count and the standardized variance of adherence after adjustment on the average adherence to HAART. Indeed, within the group DrI average, the increase in the monthly cell count rises significantly when the standardized variance falls: the difference

is 2.79 CD4 cells/mm^3 per month (respectively, 2.43), when the status shifts from "High" to "Low" variance (respectively, "Moderate" variance).

Within the group HsD average, the monthly mean of CD4 cell count did not seem to change within High and Moderate variance; however, it was significantly higher in the High than in the Low variance group (Difference: 2.20; 95% CI: 1.20–3.21 CD4 cells/mm^3 per month).

Similarly, within the group cH average, there seems to be no significant difference between the increases in the monthly mean of the CD4 cell count in High and Moderate variance. This increase was significantly more important in the Low than in the High variance group (Difference: 0.76; 95% CI: 0.26–1.25 CD4 cells/mm^3 per month).

Impact of the variability in adherence to HAART on mortality

Over the whole follow-up period, the moving average of adherence ranged from 0.91 to 99.97 (IQR: 87.30–98.41) whereas the moving standardized variance ranged from -1.85 to 1.85 (IQR: -0.55–0.51). The total number of deaths over the follow-up period was 52 (16.40% of the participants).

The results of the use of the Cox model showed that the instantaneous risk of death was significantly associated with both the average and to the standardized variance of adherence to HAART.

In the univariate analysis, i) a 10% increase in the average adherence induced a 30% decrease of the relative risk of death (RR = 0.73; 95% CI: 0.66–0.81); ii) a unit increase in the standardized variance induced a 50% increase in the relative risk of death (RR = 1.49; 95% CI: 1.04–2.12).

In the multivariate analysis, all things being equal, i) a 10% increase in the overall adherence induced a 30% decrease of the relative risk of death (RR = 0.73; 95% CI: 0.66–0.81); ii) a unit increase in the standardized variance induced a 45% increase in the relative risk of death (RR = 1.45; 95% CI: 1.03–2.06).

In the multivariate analysis, a model that included a term of interaction between the average and the standardized variance of adherence did not show a better fit than a model without this interaction. However, it helped showing that the higher was the average adherence, the higher was the effect of the standardized variance on the risk of death.

Discussion

The ANRS 1215 cohort provided the advantage of a long follow-up of patients with HIV receiving HAART and allowed proposing latent-class trajectories for the average and the variance of adherence to HAART over time. The study was able to establish associations between adherence (through its average or variance) with the viral load, the CD4 cell count, and mortality. The present study showed that trajectory cH was the dominant. This may be explained by a bias of longer patient survival but also, past the first year of treatment, by the progress made in HAART (better tolerance, less pills to be taken, fixed-dose combinations, and once-a-pill).

A few previous studies have already used trajectories groups of average adherence to HAART [15,24]. Here, the study goes beyond the average and, independently from it, identifies trajectory groups of variance of adherence within each average adherence group.

The results of the present study are in agreement with others regarding the impact of the level of adherence to HAART on the immunovirological response or on mortality. Indeed, part of the present results confirm those of Haubrich et al. [4] and Mannheimer et al. [6] regarding the increase in the percentage of undetectable viral load (meaning viral disappearance from blood) and the increase in the CD4 cell count together with the increase in the average adherence. Another part of the present results confirm those of Palella et al. [3], Nachega et al. [7], and Abaasa et al. [8] regarding the link between the average adherence and patient survival. Moreover, in a recent study conducted in Burkina Faso [25], the authors expressed adherence with a score (range from 0 to 10 points) and considered it in a Cox model as a time-varying covariate to point out that the less adherent patients have a higher risk of death. By classifying adherence into two categories (optimal: 8-10 points and sub-optimal: 0-7 points), the authors showed also that patients with optimal adherence had the best CD4 cell count recovery.

Nevertheless, the study presents real novelties regarding the impact of the variance in the adherence to HAART on the immunovirological response and on survival. Indeed, whatever the level of the average adherence: i) the rate of CD4 recovery decreases when the variance increases; ii) the probability of achieving an undetectable viral load increases when the variance is low; and, iii) the risk of death increases along with the increase of the variance.

Various methods have been used to measure adherence [26-28] and each has its advantages and limits. Here, the approach is mixed; it combines a quantitative component

(pill count) and a qualitative component (patient interview). This approach is known to be well adapted to the Sub-Saharan context because of its simplicity and low cost [26,28].

The present study shows one way to dissociate the average adherence from the variance of adherence. This method provides a more efficient measurement of the variability in adherence that is independent of the average adherence. However, though this method allowed showing significant impacts of adherence variability on some elements of the immunovirological response, the interpretation of the size effects remained difficult.

The latent-class trajectory approach with the average and the variance of adherence offers the advantage of showing that the effects of the variance may change according to the level of the average and avoids enforcing a specific form to the link between average and variance. However, this approach may induce some loss of power regarding the significance of the effects.

In conclusion, the impacts of the two components of adherence variability on the immunovirological response and survival justify the inclusion of these aspects into the process of patient education: adherence should be both high and constant. An inconstantly high adherence is not favourable for an efficient HAART.

Acknowledgments

Financial support

We thank the Projet IDoL: ANR-12-BSV1-0036 for its financial support to Olayidé Boussari.

Authors Contributions

René ECOCHARD, Olayidé BOUSSARI, and Jean-François ETARD designed the study.

Olayidé BOUSSARI, René ECOCHARD, and Subtil FABIEN did the statistical analysis.

Olayidé BOUSSARI, René ECOCHARD, Jean-François ETARD interpreted the results.

Olayidé BOUSSARI, René ECOCHARD, Jean IWAZ wrote the article.

All the authors reviewed the article.

References

1. Mocroft A, Ledergerber B, Katlama C, Kirk O, Reiss P, d'Arminio Monforte A, *et al*. **Decline in the AIDS and death rates in the EuroSIDA study: an observational study**. *Lancet* 2003; **362**: 22–29.
2. Braitstein P, Brinkhof MW, Dabis F, Schechter M, Boulle A, Miotti P, *et al*. **Mortality of HIV-1-infected patients in the first year of antiretroviral therapy: comparison between low-income and high-income countries**. *Lancet* 2006; **367**: 817–824.
3. Palella FJ Jr, Delaney KM, Moorman AC, Loveless MO, Fuhrer J, Satten G, *et al*. **Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. HIV Outpatient Study Investigators.** *N Engl J Med* 1998; **338**: 853–860.
4. Haubrich RH, Little SJ, Currier JS, Forthal DN, Kemper CA, Beall GN, *et al*. **The value of patient-reported adherence to antiretroviral therapy in predicting virologic and immunologic response. California Collaborative Treatment Group.** *AIDS* 1999; **13**: 1099 –1107.
5. Bangsberg DR, Perry S, Charlebois ED, Clark RA, Roberston M, Zolopa AR, *et al*. **Non-adherence to highly active antiretroviral therapy predicts progression to AIDS** [Letter]. *AIDS* 2001; **15**: 1181–1183.
6. Mannheimer S, Friedland G, Matts J, Child C. **The Consistency of Adherence to Antiretroviral Therapy Predicts Biologic Outcomes for Human Immunodeficiency Virus-Infected Persons in Clinical Trials**. *CID* 2002; **34**(8): 1115–1121.
7. Nachega JB, Hislop M, Dowdy DW, Lo M, Omer SB, Regensberg L, *et al*. **Adherence to highly active antiretroviral therapy assessed by pharmacy claims predicts survival in HIV-infected South African adults**. *J Acquir Immune Defic Syndr* 2006; **43**:78–84.
8. Abaasa AM, Todd J, Ekoru K, Kalyango JN, Levin J, Odeke E *et al*. **Good adherence to HAART and improved survival in a community HIV/AIDS treatment and care programme: the experience of The AIDS Support Organization (TASO), Kampala, Uganda**. *BMC Health Serv Res* 2008; **8**:241.
9. Kebba A. **Antiretroviral therapy in sub-Saharan Africa: myth or reality?** *J Antimicrob Chemother* 2003; **52**:747–749.
10. Yazdanpanah Y. **Costs associated with combination antiretroviral therapy in HIV-infected patients.** *J Antimicrob Chemother* 2004; **53**:558–561.

11. Desclaux A, Ciss M, Taverne B, Sow PS, Egrot M, Faye MA, *et al*. **Access to antiretroviral drugs and AIDS management in Senegal**. *AIDS* 2003; **17(suppl 3)**:95–101.
12. Taverne B, Desclaux A, Sow PS, Delaporte E, Ndoye I. *Evaluation de l'impact bioclinique et social, individuel et collectif, du traitement ARV chez des patients VIH-1 pris en charge depuis 10 ans dans le cadre de l'ISAARV - Cohorte ANRS 1215. Rapport final, mai 2012*. Dakar: CNLS, CRCF, IRD, ANRS; 2012. 415 p.
13. Lanièce I, Ciss M, Desclaux A, Diop K, Mbodj F, Ndiaye B, *et al*. **Adherence to HAART and its principal determinants in a cohort of Senegalese adults**. *AIDS* 2003; **17(suppl 3)**:103–108.
14. Etard JF, Lanièce I, Fall MBK, Cilote V, Blazejewski L, Diop K, *et al*. **A 84-month follow up of adherence to HAART in a cohort of adult Senegalese patients**. *Trop Med Int Health* 2007; **12**:1191–1198.
15. Bastard M, Fall MBK, Lanièce I, Taverne B, Desclaux A, Ecochard R *et al*. **Revisiting long-term adherence to highly active antiretroviral therapy in Senegal using latent class analysis**. *J Acquir Immune Defic Syndr* 2011; **57**(1), 55–61.
16. Etard JF, Ndiaye I, Thierry-Mieg M, Guèye NFN, Guèye PM, Lanièce I *et al*. **Mortality and causes of death in adults receiving highly active antiretroviral therapy in Senegal: a 7-year cohort study**. *AIDS* 2006; **20**:1181–1189.
17. McLachlan G, Peel D. *Finite Mixture Models*. Wiley: New York; 2000.
18. Skrondal A, Rabe-Hesketh S. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Boca Raton, FL: CRC Press; 2004.
19. Nagin D. *Group-based Modeling of Development*. Cambridge, MA: Harvard University Press; 2005.
20. Cox DR. **Regression models and life tables (with discussion)**. *J R Statist Soc B* 1972; **63**: 269–76.
21. Crowley J, Hu M. **Covariance analysis of heart transplant survival data**. *J Am Statist Assoc* 1977; **72**: 27–36.
22. Altman DG, De Stavola BL. **Practical problems in fitting a proportional hazards model to data with updated measurements of the covariates**. *Stat Med* 1994; **13**: 301–41.
23. R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

24. Glass TR, Battegay M, Cavassini M, De Geest S, Furrer H, Vernazza PL, *et al*. **Longitudinal analysis of patterns and predictors of changes in self-reported adherence to antiretroviral therapy: Swiss HIV Cohort Study**. *J Acquir Immune Defic Syndr* 2010; **54**:197–203.
25. Focà E, Odolini S, Sulis G, Calza S, Pietra V, Rodari P, *et al*. **Clinical and immunological outcomes according to adherence to first-line HAART in a urban and rural cohort of HIV-infected patients in Burkina Faso, West Africa**. *BMC Infect Dis* 2014; **14**:153.
26. Farmer, K. C. **Methods for measuring and monitoring medication regimen adherence in clinical trials and clinical practice**. *Clinical therapeutics* 1999; **21**(6):1074–1090.
27. Turner B. **Adherence to antiretroviral therapy by human immunodeficiency virus-infected patients**. *J Infect Dis* 2002; **185(Suppl 2)**:143–151.
28. Costagliola D, Barberousse C. **Comment mesurer l'observance?** In: *L'observance aux traitements contre le VIH/sida: mesure, déterminants, évolution,* (edited by ANRS). Paris: ANRS, Collection Sciences Sociales et Sida; 2001. pp. 33–42.

Table 1 – Distribution of the patients according to the classes of average and standardized variance of adherence to HAART.

	Average			
Variance	cH	HsD	DrI	Total
Low	40	20	11	71 (22.40%)
Moderate	127	14	14	155 (48.90%)
High	54	20	17	91 (28.70%)
Total	221 (69.72%)	54 (17.03%)	42 (13.25%)	317 (100%)

cH: constantly high average adherence - HsD: high but slowly decreasing average adherence - DrI: decreasing then rapidly increasing average adherence.

Table 2 - Relationship between adherence (average and standardized variance) and the viral load according to the mixed logistic regression.

Adherence groups	OR average	OR variance	95% CI
DrI average adherence	1		-
High variance		1	-
Moderate variance		1.28	0.64 - 2.58
Low variance		1.73	0.75 - 4.00
HsD average adherence	2.88		1.52 - 5.45
High variance		1	-
Moderate variance		0.69	0.37 - 1.28
Low variance		1.71	0.85 - 3.45
cH average adherence	2.91		1.74 - 4.86
High variance		1	-
Moderate variance		1.76	1.35 - 2.30
Low variance		1.82	1.25 - 2.66

DrI: decreasing then rapidly increasing average adherence - HsD: high but slowly decreasing average adherence - cH: constantly high average adherence

Table 3 - Relationships between adherence (average and standardized variance) and the CD4 cell count according to the mixed linear model.

Adherence groups	Monthly change in CD4 cell count (cells/mm^3 blood)	95% CI
Baseline *	0.04	-0.93 - 1.01
Moderate variance	2.43	1.20 - 3.66
Low variance	2.79	1.24 - 4.34
HsD average adherence	4.28	3.22 - 5.35
Moderate variance	-0.41	-1.44 - 0.63
Low variance	2.20	1.20 - 3.21
cH average adherence	4.70	3.82 - 5.58
Moderate variance	-0.33	-0.70 - 0.05
Low variance	0.76	0.26 - 1.25

* Corresponds to the case of DrI average (decreasing then rapidly increasing average adherence) and "High" variance. All the other values correspond to the deviation from this baseline.

Figure 1 – Trajectories of adherence averages (left panel) and standardized variances (right panel).

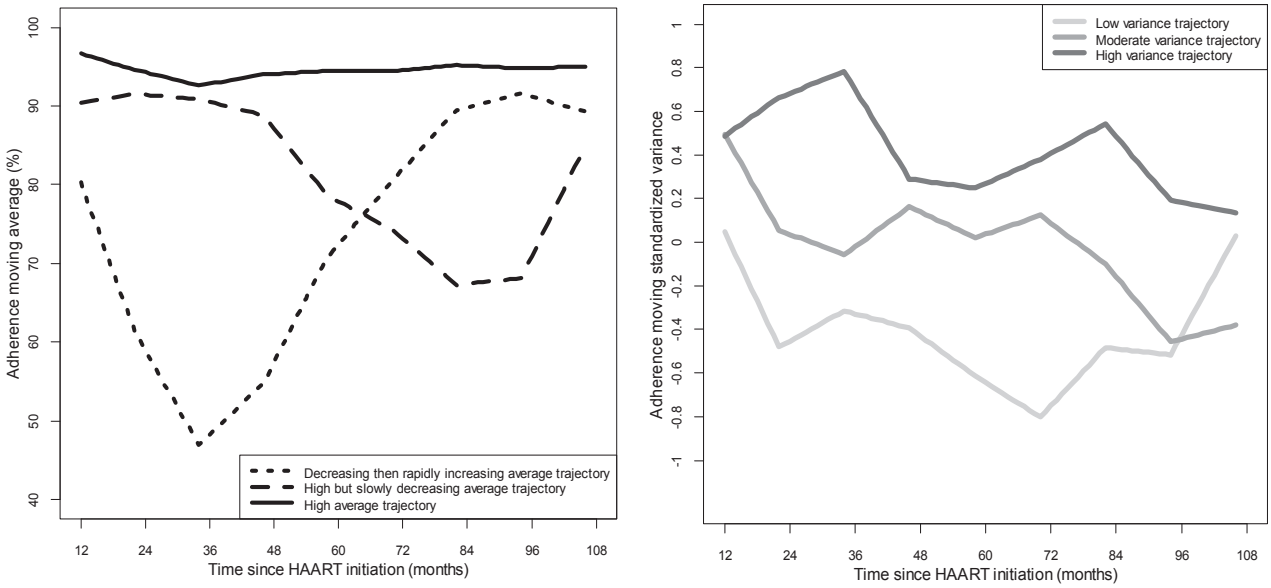
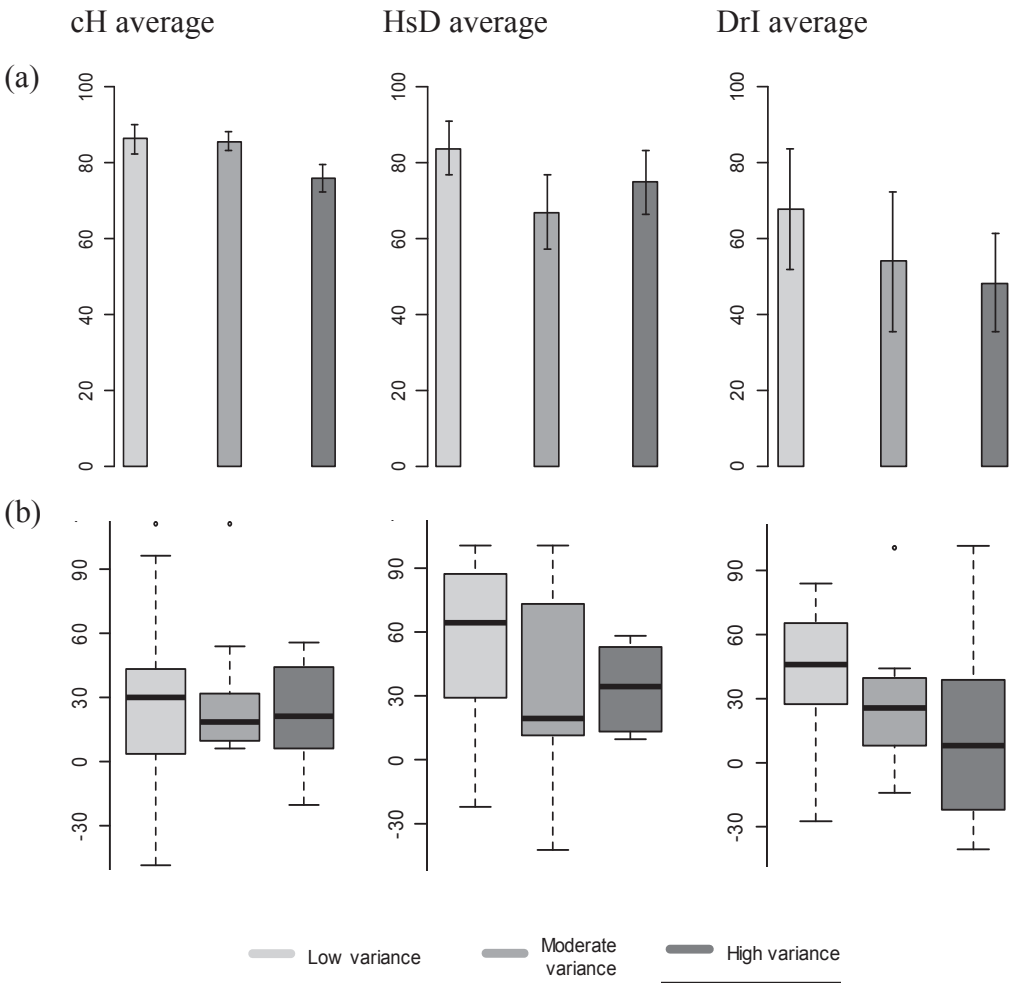


Figure 2 – Panel a: Mean percentages of undetectable viral loads (with their 95% CIs) starting from the 6th month after HAART initiation. Panel b: Variations, by six-month intervals, of the CD4 cell counts starting from the 6th month after HAART initiation. cH: constantly high average adherence - HsD: high but slowly decreasing average adherence - DrI: decreasing then rapidly increasing average adherence.



Appendix

ANRS 1215 study group (2009–2011)

Ibra Ndoye (National Council for the Fight Against AIDS, Dakar, Senegal); Eric Delaporte, Jean-François Etard, Martine Peeters, Alice Desclaux, Pierre de Beaudrap, Christian Laurent, Julie Coutherut, Tidiane Ndoye, Nicole Vidal, Claire Moquet, Sabrina Eymard-Duvernay, Cécile Cames, Kirsten Bork, Sabah Boufkhed, Mathilde Couderc, Amandine Cournil, Bernard Taverne (IRD, UMI 233, Montpellier 1 University, France); Assane Diouf, Mame Basty Koïta Fall, Alle Baba Dieng, Christian Eric Massidi, Adama Sarr, Khoudia Sow, Mariane Ndiaye Berthe, Saïdou Ba, Absa Ba, Catherine Lissoune Fall Sané, Sokhna Boye, Caroline Desclaux-Sall, Héléne Dior Mbodj, Coumba Gueye Cissé, Frédérique Muller, Kouro Bousso Niang, Marie-Louise Sarr, Estelle Simen, Alassane Sow, El Hadj Malick Sy Camara (Regional Center for Research and Training in Clinical Treatment, National University Hospital Center in Fann, Dakar, Senegal); Maryvonne Maynart, Isabelle Lanièce, Vanina Cilote (Department of Cooperation and Cultural Affairs, French Embassy, Dakar, Senegal); Papa Salif Sow, Ibrahima Ndiaye, Cheickh Tidiane Ndour, Viviane Pierre Marie Ciss, (National University Hospital Center in Fann, Infectious and Tropical Disease Unit, Dakar, Senegal); Ndeye Fatou Ngom Guèye, Djibril Baal, Batista Gilbert, Andréa Robalo Diassy (National University Hospital Center in Fann, Ambulatory Treatment Center, Dakar, Senegal); Jeanne Diaw (General Hospital in Grand Yoff, Dakar, Senegal); René Ecochard, (Claude Bernard Lyon I University, France); Mathieu Bastard (Epicentre, Paris, France); Kadidiatou Ba Fall, Pape Madoumbé Guèye, Pape Samba Ba, Madiouba Diawara (Principal Hospital in Dakar, Senegal); Souleymane Mboup, Pape Alassane Diaw, Halimatou Diop Ndiaye, Ndeye Coumba Touré Kane, Moussa Thiam (Le Dantec National University Hospital Center, Bacteriology and Virology Laboratory, Dakar, Senegal); Karim Diop (Ministry of Health, Division of AIDS Control, Dakar, Senegal); Bara Ndiaye (National University Hospital Center in Fann, Dakar, Senegal).

Conclusion et Perspectives

Ce travail a été construit autour de deux problématiques de santé relatives aux deux plus grandes pandémies actuelles en Afrique sub-saharienne. Un des problèmes posé est celui de l'hétérogénéité spatiale et temporelle de la répartition des vecteurs de paludisme. Le second problème posé est celui de la variabilité de l'observance aux traitements antirétroviraux par les personnes vivant avec le virus de l'immunodéficience humaine. Pour le statisticien, censé éclairer les praticiens de la santé et les décideurs en santé publique, ces deux problématiques conduisent à une même problématique méthodologique bien connue : la modélisation de données répétées, en particulier les données longitudinales. Dans notre travail, l'objectif dans la modélisation est double en ce sens qu'il faudra fournir non seulement des estimations assez proches des observations mais aussi pouvoir fournir dans chaque cas, une classification des unités statistiques assez sensée et ayant un intérêt pratique.

Les modèles à classes latentes (que nous identifions ici aux modèle de mélanges finis) répondent à cette double exigence grâce à leur flexibilité qui permet en outre de bénéficier de l'ensemble des résultats de la statistique mathématique : lois multivariées paramétriques, estimation, choix de modèles. En effet, dans ces modèles, l'ensemble du processus ayant généré les données, qu'elles soient observées ou manquantes, est totalement modélisé ; d'où leur qualificatif de modèles génératifs. Pour le praticien, les modèles de mélange offrent un cadre formel assez confortable d'utilisation car, sans un quelconque effort supplémentaire, il peut au travers des paramètres du modèle, obtenir un résumé exhaustif et très souvent interprétable de sa structure de partition.

Cependant, bien qu'assez efficace, il est important de ne pas perdre de vue que les

modèles de mélange ne sont qu'un simple outil de modélisation et de recherche d'un partitionnement existant au sein d'une population par exemple. Par conséquent, il serait dangereux de résumer la problématique de la classification à un simple sous produit d'un modèle de mélange. Le statisticien ou le praticien qui se sert de ces modèles comme outil statistique doit faire preuve de réalisme en prenant en compte, aussi bien dans la phase d'estimation que pour le choix de modèles, des informations supplémentaires en rapport avec l'objectif initial recherché. Par exemple, le nombre de classes peut être fixé a priori en raison des connaissances empiriques sur la problématique de santé traitée.

Dans les régions où le paludisme demeure un problème de santé publique, la connaissance du niveau de contact homme-vecteur et son évolution saisonnière est l'un des indicateurs incontournables qui interviennent en amont comme en aval de la lutte antivectorielle. Très souvent, les méthodes de LAV sont déployées sur un ensemble de localités sans une bonne connaissance de l'hétérogénéité spatiale et temporelle de la densité vectorielle ou du niveau de contact homme-vecteur qui y prévaut. Nous avons montré dans ce travail qu'il est possible de prendre en compte cette hétérogénéité à travers les modèles de mélange et de dégager des sous groupes de localités beaucoup plus spécifiques en termes de contact homme-vecteur de même que les facteurs influençant cet indicateur. Ces résultats pourraient aider, en amont, à rendre les méthodes de LAV beaucoup plus spécifiques des sites et en aval, à évaluer l'impact de ces méthodes. Les modélisations séparées du profil annuel de contact homme-vecteur et de son intensité moyenne ont conduit à deux partitions différentes des localités considérées et ont montré que ces deux caractéristiques du contact homme-vecteur ne sont pas forcément reliées aux mêmes facteurs environnementaux ou liés à la présence humaine.

Or dans la pratique ces deux caractéristiques seront prises en compte simultanément dans la spécification des sites. L'idée de modéliser conjointement le profil (annuel) du contact homme-vecteur et son intensité moyenne, avec toujours comme

outil les modèles de mélange, semble l'une des suites logiques au travail que nous avons présenté.

Par ailleurs l'hétérogénéité spatiale et temporelle du contact homme-vecteur mise en évidence dans ce travail ouvre naturellement sur une autre problématique : celle de l'optimisation du dispositif d'échantillonnage dans le cadre de la capture sur appât humain. En effet, il est important pour les entomologistes de bien sélectionner les points de captures dans un site donné afin d'optimiser la collecte des moustiques. Par exemple, on peut questionner la pertinence de garder le même point de capture durant toute la durée de l'essai (plutôt que de changer de point de capture à chaque enquête). De même est-il souhaitable, lors d'une mission de capture, de faire des captures sur des jours consécutifs ou non ? Est il important de garder les mêmes heures de captures ou de calquer les captures sur une variable climatique et/ou environnementale particulière ? Nous envisageons donc aborder ces aspects dans le cadre de travaux allant au delà de cette thèse.

Une autre partie de notre travail a été consacrée à la modélisation des trajectoires de variances d'observance aux traitements antirétroviraux par des personnes vivant avec le virus de l'immunodéficience humaine. Dans la prise en charge des patients VIH, en particulier dans le contexte des pays à ressources limitées, la question de l'observance au traitement se pose avec beaucoup d'intérêt car l'efficacité des traitements en dépend directement. Nous avons montré grâce aux modèles à trajectoires latentes que non seulement le niveau moyen d'observance, mais aussi la stabilité dans l'observance au cours du temps ont un impact sur le devenir des patients. Grâce aux modèles de mélange, nous avons, d'une part obtenu une catégorisation des patients par types de trajectoires d'observance moyenne et de trajectoires de variance de l'observance et d'autre part établi le lien entre ces deux caractéristiques de l'observance et la réponse immunovirologique de même que la survie. Ces résultats semblent des outils à prendre en compte par le praticien dans l'éducation thérapeutique des patients ; en outre ils pourront l'aider à comprendre l'échec de certaines lignes

thérapeutiques chez certains patients.

Sur le plan méthodologique, la notion de trajectoires de variance d'observance est transposable à des études menées dans le cadre d'autres maladies. Par ailleurs, la classification de courbes est un domaine de recherche actif en statistiques et dans lequel beaucoup reste à faire. Il serait sans doute utile de comparer les résultats obtenus dans le chapitre 6 avec ceux que l'on obtiendrait en utilisant d'autres méthodes de classification de courbes. Ceci permettrait d'approfondir la réflexion sur le choix de modélisation ou de prise en compte de la variabilité et constitue ainsi une ouverture possible pour la suite de ce travail de recherche.

Bibliographie

A.M. Abaasa, J. Todd, K. Ekoru, J.N. Kalyango, J. Levin, E. Odeke, and C.A.S. Karamagi. Good adherence to HAART and improved survival in a community HIV/AIDS treatment and care programme : the experience of the AIDS support organization (TASO), kampala, uganda. *BMC Health Services Research*, 8(1) : 241, 2008.

G. B. Airy. *On the algebraical and numerical theory of errors of observations and the combination of observations*. Macmillan & Company, 1861.

M. Aitkin. A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and computing*, 6(3) :251–262, 1996.

M. Aitkin. A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, 55(1) :117–128, 1999.

D.R. Bangsberg, S. Perry, E.D. Charlebois, R.A. Clark, M. Roberston, A.R. Zolopa, and A. Moss. Non-adherence to highly active antiretroviral therapy predicts progression to AIDS. *AIDS*, 15(9) :1181–1183, 2001.

C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(7) :719–725, 2000.

K.A. Bollen. *Structural equations with latent variables*. New York : Wiley, 1989.

H. Bozdogan. Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the inverse-fisher information matrix. In *Information and classification*, pages 40–54. Heiderberg : Springer-Verlag, 1993.

M. Caraël. *Twenty years of intervention and controversy. In : The HIV/AIDS epidemic in sub-Saharan Africa in a historical perspective.* Becker PDaC, 2006.

P. Carnevale, V. Robert, S. Manguin, V. Corbel, D. Fontenille, C. Garros, C. Rogier, and J. Roux. *Les anophèles : biologie, transmission du plasmodium et lutte antivectorielle.* Marseilles : IRD Editions, 2009.

A. Castro. Adherence to antiretroviral therapy : merging the clinical and social course of AIDS. *PLoS Medicine*, 2(12) :e338, 2005.

CDC. Pneumocystis pneumonia. *Centers for Disease Control : Morbidity and mortality weekly report*, 30 :250–252, 1981.

G. Celeux and G. Govaert. A classification EM algorithm for clustering and two stochastic versions. *Computational statistics & Data analysis*, 14(3) :315–332, 1992.

M. Coetzee, R.H. Hunt, R. Wilkerson, A. Della Torre, M.B. Coulibaly, and N.J. Besansky. *Anopheles coluzzii* and *Anopheles amharicus*, new members of the anopheles gambiae complex. *Zootaxa*, 3619(3) :246–274, 2013.

T. Coffinet, C. Rogier, and F. Pages. Evaluation de l'agressivité des anophèles et du risque de transmission du paludisme : méthodes utilisées dans les armées françaises. *Médecine tropicale*, 69(2) :109–122, 2009.

M. Coluzzi, A. Sabatini, V. Petrarca, and M.A. Di Deco. Chromosomal differentiation and adaptation to human environments in the anopheles gambiae complex. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 73(5) :483–497, 1979.

V. Corbel and R. N'Guessan. Distribution, mechanisms, impact and management of insecticide resistance in malaria vectors : a pragmatic review. In *Anopheles mosquitoes - New insights into malaria vectors.* InTech, 2013.

J. Cox-Singh and B. Singh. Knowlesi malaria : newly emergent and of public health importance ? *Trends in parasitology*, 24(9) :406–410, 2008.

G.B. Damien, A. Djènontin, C. Rogier, V. Corbel, S.B. Bangana, F. Chandre, M. Akogbéto, D. Kindé-Gazard, A. Massougbodji, and M.-C. Henry. Malaria infection and disease in an area with pyrethroid-resistant vectors in southern benin. *Malaria journal*, 9(1) :380, 2010.

A. Desclaux, M. Ciss, B. Taverne, P.S. Sow, M. Egrot, M.A. Faye, I. Lanièce, O. Sylla, E. Delaporte, and I. Ndoye. Access to antiretroviral drugs and AIDS management in senegal. *AIDS*, 17 :S95–S101, 2003.

A. Djènontin, S. Bio-Bangana, N. Moiroux, M.-C. Henry, O. Boussari, J. Chabi, R. Ossè, S. Koudénoukpo, V. Corbel, M. Akogbéto, et al. Culicidae diversity, malaria transmission and insecticide resistance alleles in malaria vectors in ouidah-kpomasse-tori district from benin (west africa) : A pre-intervention study. *Parasit Vectors*, 3 :83, 2010.

A.M. Dondorp, F. Nosten, P. Yi, D. Das, A.P. Phyo, J. Tarning, K.M. Lwin, F. Ariey, W. Hanpithakpong, S.J. Lee, et al. Artemisinin resistance in *Plasmodium falciparum* malaria. *New England Journal of Medicine*, 361(5) :455–467, 2009.

J.J. Droesbeke, G. Saporta, and C. Thomas-Agnan. *Modèles à variables latentes et modèles de mélange*. TECHNIP OPHRYS EDITIONS, 2013.

E. Fee and M. Parry. Jonathan mann, HIV/AIDS, and human rights. *Journal of public health policy*, 29(1) :54–71, 2008.

R. A. Fisher. Statistical methods for research workers. *Oliver and Boyd : London*, 1925.

M.T. Gillies. Studies on the dispersion and survival of anopheles gambiae giles in east africa, by means of marking and release experiments. *Bulletin of Entomological Research*, 52(01) :99–127, 1961.

M.T. Gillies and T.J. Wilkes. A study of the age-composition of populations of *An. gambiae Giles* and *An. funestus Giles* in north-eastern tanzania. *Bulletin of entomological research*, 56(02) :237–262, 1965.

W.C. Gorgas. Malaria prevention on the isthmus of panama. *The Prevention of Malaria*, pages 346–352, 1910.

B. Grassi. Ciclo evolutivo delle semilune nell' « anopheles claviger » ed altri studi sulla malaria dall'ottobre 1898 al maggio 1899. *Atti d. Soc. Per gli Studi d. Malaria*, 14, 1899.

S.M. Hammer, K.E. Squires, M.D. Hughes, J.M. Grimes, L.M. Demeter, J.S. Currier, J.J. Eron Jr, J.E. Feinberg, H.H. Balfour Jr, L.R. Deyton, et al. A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and cd4 cell counts of 200 per cubic millimeter or less. *New England Journal of Medicine*, 337(11) :725–733, 1997.

J.A. Hartigan. *Clustering algorithms*. New York : John Wiley & Sons, Inc., 1975.

R.J. Hathaway. Another interpretation of the EM algorithm for mixture distributions. *Statistics & Probability Letters*, 4(2) :53–56, 1986.

R.H. Haubrich, S.J. Little, J.S. Currier, D.N. Forthal, C.A. Kemper, G.N. Beall, D. Johnson, M.P. Dubé, J.Y. Hwanga, J.A. McCutchan, et al. The value of patient-reported adherence to antiretroviral therapy in predicting virologic and immunologic response. *AIDS*, 13(9) :1099–1107, 1999.

M. Imwong, M.E. Boel, W. Pagornrat, M. Pimanpanarak, R. McGready, N.P.J. Day, F. Nosten, and N.J. White. The first plasmodium vivax relapses of life are usually genetically homologous. *Journal of Infectious Diseases*, 205(4) :680–683, 2012.

N.L. Johnson and S. Kotz. *Distribution in statistics : Discrete distribution*. Wiley Online Library, 1969.

A. Kebba. Antiretroviral therapy in sub-saharan africa : myth or reality ? *Journal of Antimicrobial Chemotherapy*, 52(5) :747–749, 2003.

C. Keribin. Consistent estimation of the order of mixture models. *Sankhya Ser. A*, 62(1) :49–66, 2000.

B. Korber, M. Muldoon, J. Theiler, F. Gao, R. Gupta, A. Lapedes, B.H. Hahn, S. Wolinsky, and T. Bhattacharya. Timing the ancestor of the HIV-1 pandemic strains. *Science*, 288 :1789–1796, 2000.

N.M. Laird and J.H. Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982.

D. Lambert. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1) :1–14, 1992.

R. Landman, R. Schiemann, S. Thiam, M. Vray, A. Canestri, S. Mboup, C.T. Kane, E. Delaporte, P.S. Sow, M.A. Faye, et al. Once-a-day highly active antiretroviral therapy in treatment-naive HIV-1-infected adults in senegal. *AIDS*, 17(7) :1017–1022, 2003.

A. Laveran. Note sur un nouveau parasite trouvé dans le sang de plusieurs malades atteints de fièvre palustre. *Bulletin de l'Académie médicale*, 2 :1235–1236, 1880.

I.C. Lerman. *Classification automatique et analyse ordinale des données*. Paris : Dunod, 1981.

G.A. Livadas. Do anophelines acquire resistance to DDT ? *WHO Expert Panel on Malaria*, 1951.

G.A. Livadas and G. Georgopoulos. Development of resistance to DDT by anopheles sacharovi in greece. *Bulletin of the World Health Organization*, 8(4) :497, 1953.

S. Mannheimer, G. Friedland, J. Matts, C. Child, M. Chesney, et al. *Clinical infectious diseases*, 34(8) :1115–1121, 2002.

P. McCullagh and J.A. Nelder. *Generalized linear models*. London : Chapman and Hall/CRC, 1989.

G.J. McLachlan and D. Peel. *Finite Mixture Models*. New York : Wiley, 2000.

M.H. Merson, J. O'Malley, D. Serwadda, and C. Apisuk. The history and challenge of HIV prevention. *The Lancet*, 372(9637) :475–488, 2008.

E.J. Mills, J.B. Nachega, D.R. Bangsberg, S. Singh, B. Rachlis, P. Wu, K. Wilson, I. Buchan, C.J. Gill, and C. Cooper. Adherence to HAART : a systematic review of developed and developing nation patient-reported barriers and facilitators. *PLoS medicine*, 3(11) :e438, 2006.

A. Mocroft, B. Ledergerber, C. Katlama, O. Kirk, P. Reiss, A. Monforte, B. Knysz, M. Dietrich, A.N. Phillips, and J. D. Lundgren. Decline in the AIDS and death rates in the eurosida study : an observational study. *The Lancet*, 362(9377) :22–29, 2003.

J. Mouchet, P. Carnevale, J. Julvez, S. Manguin, D. Richard-Lenoble, and J. Sircoulon. *Biodiversité du paludisme dans le monde*. Paris : John Libbey Eurotext, 2004.

J.B. Nachega, M. Hislop, D.W. Dowdy, M. Lo, S.B. Omer, L. Regensberg, R.E. Chaisson, and G. Maartens. Adherence to highly active antiretroviral therapy assessed by pharmacy claims predicts survival in HIV-infected south african adults. *JAIDS*, 43(1) :78–84, 2006.

J.A. Najera and M. Zaim. *Lutte contre les vecteurs de paludisme. Critères et procédures pour prise de décisions pour une utilisation raisonnée des insecticides.* Geneva : World Health Organization, 2005.

J. Neyman and E.L. Scott. Statistical approach to problems of cosmology. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20(1) :1–43, 1958.

ONUSIDA. Rapport 2013 sur l'épidémie mondiale de sida. 2013.

F. Pages, E. Orlandi-Pradines, and V. Corbel. Vecteurs du paludisme : biologie, diversité, contrôle et protection individuelle. *Medecine et Maladies Infectieuses*, 37(3) :153–161, 2007.

F.J. Palella Jr, K.M. Delaney, A.C. Moorman, M.O. Loveless, J. Fuhrer, G.A. Satten, D.J. Aschman, and S.D. Holmberg. Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. *New England Journal of Medicine*, 338(13) :853–860, 1998.

F.A. Plummer, N.J. Nagelkerke, S. Moses, J.O. Ndinya-Achola, J. Bwayo, and E. Ngugi. The importance of core groups in the epidemiology and control of HIV-1 infection. *AIDS*, 5 :S169, 1991.

H. Ranson, R. N'Guessan, J. Lines, N. Moiroux, Z. Nkuni, and V. Corbel. Pyrethroid resistance in african anopheline mosquitoes : what are the implications for malaria control ? *Trends in parasitology*, 27(2) :91–98, 2011.

R. Ross. On some peculiar pigmented cells found in two mosquitoes fed on malarial blood. *British medical journal*, 18 :1786–1788, 1897.

G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2) : 461–464, 1978.

A.J. Scott and M.J. Symons. Clustering methods based on likelihood ratio criteria. *Biometrics*, 27(2) :387–397, 1971.

J.B. Silver. *Mosquito ecology : field sampling methods*. London : Springer, 2008.

A. Skrondal and S. Rabe-Hesketh. *Generalized latent variable modeling : Multilevel, longitudinal, and structural equation models*. Boca Raton, FL : CRC Pres, 2004.

R.W. Snow, C.A. Guerra, A.M. Noor, H.Y. Myint, and S.I. Hay. The global distribution of clinical episodes of plasmodium falciparum malaria. *Nature*, 434(7030) : 214–217, 2005.

M.J. Symons. Clustering criteria and multivariate normal mixtures. *Biometrics*, 37 :35–43, 1981.

B. Taverne, A. Desclaux, P.S. Sow, E. Delaporte, and I. Ndoye. Evaluation de l'impact bioclinique et social,individuel et collectif, du traitement arv chez des patients vih-1 pris en charge depuis 10 ans dans le cadre de l'ISAARV - cohorte ANRS 1215. rapport final, 2012.

WHO. *Adherence to Long-Term Therapy : Evidence for Action.* Geneva : World Health Organization, 2003.

WHO. *World malaria report 2013*. WHO Library Cataloguing-in-Publication Data, 2013.

WHO, March 2014. URL `http://www.who.int/mediacentre/factsheets/fs094/fr/`.

Y. Yazdanpanah. Costs associated with combination antiretroviral therapy in HIV-infected patients. *Journal of Antimicrobial Chemotherapy*, 53(4) :558–561, 2004.