Principes de méthodes " non classiques, non statistiques et massivement multivariées " et de réduction de la complexité. Applications en épidémiologie sociale et en médecine légale

Thomas Lefèvre

THESE DE DOCTORAT DE
L'UNIVERSITE PIERRE ET MARIE CURIE

Spécialité

Epidémiologie sociale

ECOLE DOCTORALE PIERRE LOUIS DE SANTE PUBLIQUE A PARIS :
EPIDEMIOLOGIE ET SCIENCES DE L'INFORMATION BIOMEDICALE

Présentée par

M. Thomas LEFEVRE

Pour obtenir le grade de

DOCTEUR de l'UNIVERSITÉ PIERRE ET MARIE CURIE

Sujet de la thèse :

Principes de méthodes « non classiques, non statistiques et massivement multivariées » et de réduction de la complexité. Applications en épidémiologie sociale et en médecine légale

soutenue le 22 juillet 2015

devant le jury composé de :

M. Pierre CHAUVIN, Directeur de thèse
M. Cyrile DELPIERRE, rapporteur
M. Paul DOURGNON, examinateur
Mme Marie-Christine JAULENT, examinatrice
Mme Florence JUSOT, rapporteuse

Résumé

Si le vivant est complexe, l'introduction de la dimension sociale semble ne pouvoir qu'ajouter en complexité. Si la médecine est une discipline entre savoir et savoir-faire, la médecine légale est un domaine qui démultiplie ces dimensions, puisqu'elle doit articuler des disciplines peu conçues les unes pour les autres, tout en traitant également des rapports entre les vivants. La complexité qui traverse ces deux disciplines que sont l'épidémiologie sociale et la médecine légale du vivant est celle que l'on cherche à saisir par la variété des observations et par l'intrication de phénomènes, de points de vue d'échelles différentes – l'individu, le groupe, la société. Les méthodes scientifiques du biomédical sont encore peu adaptées au traitement de la complexité, à sa représentation qui ne soit pas normative d'emblée, statistique le plus souvent. Il existe pourtant un ensemble d'approches non statistiques, « non classiques », qui puissent traiter simultanément un grand nombre de dimensions et enfin qui permettent de réduire la complexité apparente ou intrinsèque du réel en en dégageant des objets d'étude spécifique. Nous présentons ici les principes et l'utilisation des techniques de reconnaissance de forme – de *clustering* – dans le cadre de l'épidémiologie sociale, en les appliquant à la recherche d'une typologie de recours aux soins, sur la base des données de la cohorte francilienne SIRS. Nous expliquons en quoi ces approches ont leur place, épistémologiquement et techniquement parlant, aux côtés des méthodes expérimentales classiques type essais randomisés contrôlés. Nous exposons également un autre moyen, complémentaire, de réduire la complexité apparente des données, tout en en préservant les qualités topologiques. Nous introduisons ainsi la notion de dimension intrinsèque, plus petite dimension nécessaire et suffisante à la description des données, et de techniques non linéaires de réduction de la dimension dans le champ de la médecine légale du vivant. Nous en appliquons les principes dans le cadre de l'intégration de sources d'information multiples pour l'estimation de l'âge chez les adolescents migrants. Enfin, nous discutons les avantages et limites de ces différentes approches, ainsi que les perspectives qu'elles ouvrent à ces deux disciplines, a priori faites pour se compléter, en pratique peu souvent en contact.

Abstract

The living is a complex topic to study, and the irruption of the social dimension may increase its complexity. Medicine is often considered as a hybrid field, partly scientific and partly made of specific know-hows. Clinical legal medicine, which stands for the "medicine of violence" is all the more a hybrid object in the sense that it articulates several fields, while also accounting for interpersonal relationships in the society. The complexity that characterizes both the social epidemiology and the clinical legal medicine is investigated through a collection of points of view and dimensions, from different observation scales: the individual scale, the group scale and the society scale. The techniques that are used in the biomedical field are not designed to properly deal with such a complexity, to render adequate observations that are not intrinsically normative. A wide range of alternative non statistical, "non-classical" methods exist that can process simultaneously a lot of various dimensions so that we can reduce the apparent or intrinsic complexity of reality while discovering genuine scientific objects in data. In this work, we present the principles and the use of techniques known as pattern recognition or clustering techniques, applied in the context of social epidemiology. More specifically, we applied different clustering techniques on data derived from the French SIRS cohort to build a typology of healthcare utilization in the Paris metropolitan area. From an epistemological and technical point of view, we explain why these methods should take place beside other recognized but limited techniques such as the randomized controlled trials. We also introduce another but complementary kind of complexity reduction technique. The concept of intrinsic dimension is explicated – the littlest dimension needed to describe properly data – and techniques termed nonlinear dimensionality reduction techniques are applied to a typical topic in clinical legal medicine. With these tools, we explore whether the integration of multiple information sources is relevant in the context of age estimation in living migrants. Finally, we discuss the strengths and shortcomings of these newly introduced methods, as well as the opportunities they may create for both fields of social epidemiology and clinical legal medicine.

Avant-propos

Ces travaux s'inscrivent dans un contexte où il est fait grand cas de la pluri, multi ou transdisciplinarité, mais pour lesquelles il semble que cette ambition reste majoritairement lettre morte. La tâche n'est pas nécessairement la plus simple, peut-être plus dans sa mise en pratique qu'en théorie. La multiplicité des équipes de recherche, de leurs thèmes, leurs tailles variables, leur dispersion géographique – même modeste, à l'échelle d'une ville ou moins – ou encore les politiques d'évaluation et de valorisation de la recherche, nationales ou internationales, tous ces facteurs font d'un véritable travail de recherche transdisciplinaire une gageure.

Je sais donc gré tant à l'équipe de Pierre Chauvin qu'au service de Patrick Chariot d'avoir permis les rencontres, favorisé les travaux transdisciplinaires dont il sera question dans ce manuscrit.

Evidemment, le biais choisi, celui des méthodes, autorise une approche de la transdisciplinarité plus aisée, s'il s'avère que ces méthodes sont en effet d'intérêt transdisciplinaire. Néanmoins, l'introduction de nouvelles méthodes dans un champ ou une discipline donnés nécessite de les situer dans la nomenclature des méthodes déjà disponibles, d'un point de vue épistémologique.

Les deux disciplines mises en avant ici sont l'épidémiologie sociale et la médecine légale du vivant. Elles partagent des intérêts évidents, mais leur rencontre concrète n'a pour ainsi dire quasiment aucune réalité. Nous souhaitons donc contribuer par ces travaux à illustrer, justifier et promouvoir cette collaboration.

Cette thèse de science est le prolongement d'une thèse d'exercice en médecine, soutenue en 2013, et dont les travaux effectués au sein de l'équipe de Pierre Chauvin en sont la base. Elle est également le prolongement et la mise en application de certains principes et techniques développés dans le cadre d'une thèse de sciences en mathématiques appliquées.

Table des matières

Tables des figures

Table des tableaux

Abréviations et acronymes

ACM : analyse en correspondances multiples

ACP : analyse en composantes principales

ACS : aide à la complémentaire santé

AME : aide médicale d'état

ANOVA : *analysis of variance*

ANR : agence nationale de la recherche

ASMR : amélioration du service médical rendu

BEH : bulletin épidémiologique hebdomadaire

CMP : centre médico-psychologique

CMU : couverture médicale universelle

CMUc : couverture médicale universelle complémentaire

CNAMTS : caisse nationale d'assurance maladie des travailleurs salariés

CNRS : centre national de la recherche scientifique

CNS : comptes nationaux de la santé

CSBM : consommation de soins et de biens médicaux

DCS : dépense courante de santé

DESC : diplôme d'études spécialisées complémentaires (médecine)

DIANA : *divisive analysis*

DREES : direction de la recherche, des études, de l'évaluation et des statistiques

DS3 : (équipe) déterminants sociaux de la santé et du recours aux soins

EHESS : école des hautes études en sciences sociales

EHPA : établissements d'hébergement pour personnes âgées

EHPAD : établissements d'hébergement pour personnes âgées et dépendantes

ENS : école normale supérieure

ERIS : équipe de recherche sur les inégalités sociales

ERES : équipe de recherche en épidémiologie sociale

ESAT : établissement et service d'aide par le travail

GHM : groupe homogène de malades

GHS : groupe homogène de séjours

HAD : hospitalisation à domicile

HAS : haute autorité de santé

HCLUST : *hierarchical clustering*

HCSP : Haut conseil de la santé publique

HLLE : *hessian locally linear embedding*

HPST : (loi) hôpital, patient, santé et territoire

IC95 : intervalle de confiance à 95%

IDH : Indice de développement humain

IGAS : Inspection générale des affaires sociales

IMC : indice de masse corporelle

INCa : institut national du cancer

INSEE : institut national de la statistique et des études économiques

INSERM : institut national de la santé et de la recherche médicale

IRDES : institut de recherche et de documentation en économie de la santé

IReSP : institut de recherche en santé publique

IRIS : îlots regroupés pour l'information statistique

ISOMAP : *isometric feature mapping*

LLE : *locally linear embedding*

LTSA : *local tangent space alignment*

MCO : médecine, chirurgie, obstétrique

MDS : *multidimensional scaling*

NHS : *national health service*

NICE : *national institute for clinical excellence*

NLDR : *nonlinear dimensionality reduction*

OCDE : organisation de coopération et de développement économique

ONG : organisation non gouvernementale

oNML : observatoire nationale de la médecine légale

OR : odd-ratio

PASS : permanence d'accès aux soins de santé

PAM : *partitioning around medoïds*

PIB : produit intérieur brut

PMSI : programme de médicalisation des systèmes d'information

PSPH : participant au service public hospitalier

REFINEMENT : *research on financing systems' effect on the quality of mental health care*

RIM-P : recueil d'informations médicalisées pour la psychiatrie

SAU : service d'accueil des urgences

SEM : *structural equations modeling*

SIRS : santé, inégalités et ruptures sociales

SMR : service médical rendu

SOM : *self-organizing map*

SOTA : *self-organizing tree algorithm*

SSR : soins de suite et de réadaptation

T2A : tarification à l'activité

SVM : *support vector machine* (ou : séparateur à vastes marges)

ZUS : zone urbaine sensible

Epidémiologie sociale, médecine légale et méthodes

Le sujet de ces travaux tient en l'application de méthodes que nous avons qualifiées de « non classiques, non statistiques et massivement multivariées », aux champs de l'épidémiologie sociale et de la médecine légale. Un tel choix appelle plusieurs remarques : d'une part, pourquoi la juxtaposition de deux domaines – l'épidémiologie sociale et la médecine légale – qui, historiquement et selon le sens commun, ne semble pas aller de soi ? D'autre part, quelle justification, sinon quel sens, aux qualificatifs employés pour désigner ces méthodes que nous entendons appliquer à ces deux domaines apparemment dissemblables ?

Par l'exposé de ces travaux, nous entendons illustrer la nécessité d'approches pluridisciplinaires de problèmes complexes – nous discuterons la notion de complexité – nécessité justifiée tant par la nature de ces problèmes que par les moyens réduits que nous imposent les cadres théoriques dominants en biomédecine lorsqu'il s'agit de production de connaissance et de niveaux de preuve. Plus généralement, nous utiliserons une partie des principes et techniques qui ont pu être développés dans le cadre d'une précédente thèse de sciences en mathématiques appliquées, où nous avons pu discuter l'intérêt d'une approche morphodynamique basée sur l'information et la géométrique, dans le champ de la biomédecine.[1–3]

Approches pluridisciplinaires de problèmes complexes

L'épidémiologie sociale peut se définir comme le domaine s'intéressant aux déterminants sociaux de la santé, c'est-à-dire, aux mécanismes d'origines sociales conditionnant l'état de santé des personnes et de groupes sociaux.[4–6] Il s'agit ici d'écarter la critique facile objectant d'emblée que les facteurs sociaux ne sauraient rendre compte de la santé des personnes, ou qu'il n'existe pas de déterministe social – entendu du point de vue mécaniste le plus pur et si possible mono causal. Dès lors, parler d'épidémiologie sociale revient à parler de contributions plus ou moins fortes, plus ou moins causales, de certains aspects sociaux à l'état de santé des personnes et des populations. A contrario, il tient à l'épidémiologie sociale de savoir éliminer tout facteur social qui apparaîtrait corrélé à l'état de santé, mais ne dit rien sur les mécanismes menant à des états de santé distincts. Par ailleurs, s'il existe des facteurs associés ou contribuant à des différences d'états de santé entre certains groupes sociaux, on pourra parler d'inégalités sociales de santé. Le débat peut alors être reporté sur le caractère modifiable d'une société, pour tout ou partie, et les voies permettant la réduction de ces

inégalités sociales de santé. L'épidémiologie se place donc comme une science au croisement de la sociologie, dont les objets d'études sont virtuellement infinis, et de l'épidémiologie. L'épidémiologie « classique » souffre de problèmes propres que sont, de manière non exhaustive, la difficulté de définir correctement les populations d'étude, de définir et mesurer les expositions, d'identifier des relations causales entre exposition et pathologie ou encore de mettre en place des interventions. Ces difficultés se retrouvent mécaniquement dans le champ de l'épidémiologie sociale, où s'y additionnent des difficultés spécifiques, par exemple héritées du champ de la sociologie. Parmi ces difficultés, nous voudrions mentionner les conflits certainement plus apparents que dans d'autres domaines entre objectivisme et relativisme, et l'apparente infinité des objets définissables. Enfin, s'ajoute la dimension de l'échelle : l'épidémiologie et la santé publique s'occupent certes de populations, mais de telle sorte qu'elles visent finalement l'individu, tant comme cause partielle que comme objet subissant des conséquences. Dire que l'épidémiologie sociale s'attaque à des problèmes complexes relève par conséquent de l'euphémisme.

La médecine légale s'entend fréquemment comme la médecine s'intéressant aux causes et manières de décès, aux scènes de crime et éventuellement à la médecine pénitentiaire.[7,8] Cette vision n'est que très partiellement correcte. Correcte dans le sens où elle comprend effectivement la thanatologie, donc la science des morts, de la mort ; très partiellement dans le sens où cette activité n'est en réalité que très minoritaire, ne serait-ce qu'en France, par rapport aux autres champs qu'elle recouvre – peut-être 5 ou 10 % des actes réalisés.[9] Quels sont ces autres champs ? Hors activité d'expertises médicales, la médecine légale est aussi la médecine légale des vivants, que l'on peut voir définie comme la médecine des violences, des situations de violence.[7] Par violence, on entend plus volontiers violences interpersonnelles, et la dichotomie violences volontaires et violences involontaires (on pensera par exemple aux accidents de la voie publique). Classiquement et relativement artificiellement, les violences sont catégorisées comme violences physiques, psychologiques ou sexuelles, sans que ces catégorises soient mutuellement exclusives. S'y ajoutent la maltraitance et les négligences lourdes. Deux activités importantes en médecine légale du vivant sont enfin l'examen de personnes placées en garde à vue – la contrepartie d'une privation de liberté est de pourvoir à la sécurité des personnes concernées, notamment sanitaire, et l'un des droits d'une personne placée en garde à vue est de pouvoir demander à voir un médecin – et l'estimation d'âge pour des personnes ne pouvant justifier de leur identité ou de leur date de naissance. Toutes ces situations ont en commun le fait qu'elles relèvent, au moins potentiellement, du système

judiciaire, et plus particulièrement, d'un traitement pénal. La population rencontrée dans la pratique de la médecine légale est une population complexe, et n'est pas donnée a priori, elle est sélectionnée et le produit vraisemblable de plusieurs mécanismes.[10,11] Pour ce qui concerne la population des personnes examinées dans le contexte d'une garde à vue, ou examinées dans le but d'estimer leur âge, la part sociale paraît évidente : ces populations « n'existeraient pas » si le code pénal ne caractérisaient pas un ensemble de situations comme autant d'infractions (détention, cession ou consommation de stupéfiants dont le cannabis, être sur le territoire français sans pouvoir justifier de sa nationalité ou de son identité), ni si le travail policier correspondant ne s'appliquait pas. En conjonction avec ces origines ou indépendamment de celles-ci, il existe évidemment une participation sociale à l'état de santé des personnes vues et aux situations de violences dans lesquelles elles sont impliquées, souvent complexes. On pensera par exemple aux violences conjugales, à la maltraitance ou encore à l'infraction « d'outrage et rébellion », facilement associée à celle de « violences volontaires sur personnes dépositaires de l'autorité publique ». Enfin, santé et violences sont fortement associées, tant ponctuellement (traumatismes physiques ou psychologiques), que s'inscrivant dans le temps (développement d'un état de stress post-traumatique ou troubles psychologiques,[12–14] trajectoire de personnes maltraitées plus exposées au risque suicidaire, aux addictions ou aux histoires de « victimes répétées »[15–18]). Les situations rencontrées en médecine légale sont, on le comprend bien, également hautement complexes et riches en intrications. A ce titre, elles peuvent en partie faire l'objet d'une approche type épidémiologie sociale. Enfin, elles sont toutes sauf marginales : rien qu'en France, on recensait environ 700 000 mesures de garde à vue, quand on examinait, rien que dans le service de médecine légale de Bondy (93), plus de 11 000 victimes en 2011.

Par leur objet d'étude et historiquement, ces deux domaines – l'épidémiologie sociale et la médecine légale – traitent de problèmes complexes, et convoquent volontiers des profils professionnels variés.

Production de la connaissance et niveaux de preuve

En restant simple, la science connaît deux grands modes de production de connaissance, incompatibles et ne rendant pas compte correctement de la réalité, qui soit reconnue comme « scientifique » : l'inductionnisme et le falsificationnisme.[19]

L'inductionnisme est une position reposant sur l'observation et la répétition (la reproductibilité) de ces observations, dont on tirera une vérité générale par induction. Le falsificationnisme prend un parti radicalement différent : une théorie – des énoncés généraux portant sur un domaine particulier – est tenue comme valide tant qu'elle résiste aux tests auxquels on la soumet. Si la théorie échoue à un test visant à vérifier un cas couvert par ses énoncés généraux, alors elle est décrétée fausse, et doit être remplacée par une autre. Il existe des critiques plus ou moins radicales de ces deux positions, voire de la valeur de la science si ce n'est sur sa définition, ainsi que des variantes ou améliorations de ces deux positions principales. Les deux positions sont tenues pour insuffisantes, mais rendent compte partiellement d'une part des pratiques, et reflètent certainement les raisonnements frustes des scientifiques selon leur domaine, qu'ils en soient conscients ou non d'autre part.

En biomédecine, actuellement, il semble qu'il n'y ait que peu de place pour ces deux positions, et l'inductionnisme paraît prévaloir, tout en se parant à la fois d'un paradoxe méthodologique et d'ornements méthodologiques censés augmenter la valeur de la production de connaissance. Le paradoxe tient dans le fait qu'il est fait un recours non totalement exclusif mais très largement dominant du test d'hypothèse.[20–22] Un test d'hypothèse consiste à 1) formuler une hypothèse dite « nulle » et une hypothèse dite « alternative » - par exemple, l'hypothèse nulle peut être qu'il n'existe aucune différence de poids, en moyenne, entre deux groupes d'individus, l'un suivant un régime A l'autre un régime B ; l'hypothèse alternative est alors qu'il existe une telle différence - et à 2) rejeter ou non cette hypothèse nulle, en indiquant avec quel degré de certitude ce rejet n'est pas dû uniquement à la « chance », mais bien à l'action d'un facteur – ici, l'action plus importante du régime A par rapport au régime B.[23] Ainsi, une approche par test d'hypothèse semblerait davantage l'apanage d'une méthode concernant le falsificationnisme. Or, il n'en est rien, car on utilise le test d'hypothèse non pas pour réfuter ce que l'on entend démontrer, mais pour affirmer qu'il existe une différence que l'on désire mettre en évidence : une preuve positive, et non une réfutation. En ce sens, la recherche biomédicale, s'appuyant théoriquement sur la reproductibilité des expériences (finalement : la réplication de ce test d'hypothèse, dans conditions plus ou moins similaires), ainsi que sur l'accumulation d'observations convergeant vers le même résultat, par des voies différentes (expériences cliniques, arguments biologiques, arguments physico-chimiques), est essentiellement une science se constituant par un mécanisme inductionniste : on tire des conclusions générales sur l'effet d'une intervention ou d'un facteur, sur la base d'observations répétées, avec la conviction d'une plus grande confiance ou valeur de « vérité » si le nombre

d'expériences aux résultats similaires ou compatibles est d'autant plus important. Par ornements méthodologiques, on entend le recours croissant à la statistique, et à un niveau plus profond, aux probabilités, pour assoir davantage la conviction – et la précision ? – de résultats fiables.

Par ailleurs, la recherche biomédicale est également une illustration exemplaire des limites les plus importantes reconnues à l'inductionnisme. Brièvement, un problème majeur de l'inductionnisme se formule ainsi : quel que soit le nombre d'observations effectuées, on ne saura jamais assuré qu'on ne fera jamais une observation contraire à l'énoncé général que l'on a induit des premières observations. Dans le cas de la biomédecine, cette critique prend une autre dimension : il est connu que les études dites « négatives », c'est-à-dire se basant sur un test d'hypothèse et dont la conclusion en a été l'absence de rejet de cette hypothèse, ne sont pas ou très marginalement publiées.[24] Ainsi, dans un contexte où seules les études positives sont connues et mobilisables, on en déduira des énoncés généraux qui pourraient être contredit s'il n'existait pas cette asymétrie d'information.

Nous en venons aux niveaux de preuve, admis et hiérarchisés en recherche biomédicale. Les plus hauts standards en termes de valeur de la preuve produite sont fournis par les essais randomisés contrôlés et les méta-analyses.[25] Les essais randomisés contrôlés ont évidemment des avantages certains, mais souffrent en contre partie des limites du point de vue inductionniste, sans parler du caractère artificiel des résultats ou de l'intensité de l'effet mis en évidence. Les méta-analyses, se basant sur les essais randomisés contrôlés et sur les résultats essentiellement publiés, souffrent donc des deux niveaux de limites de l'inductionnisme que nous venons de discuter. Enfin, c'est bien toute la chaîne de production de preuves qui est depuis quelques années critiquée en biomédecine, à différents niveaux : outils méthodologiques inadaptés à la question de recherche ou au schéma d'étude, manipulations de données, protocoles non respectés ou mal décrits, biais de publication et survalorisation des questions originales, fraude pure et simple ou encore mauvaise interprétation des résultats ou résultats ne soutenant pas les conclusions.[26–30] Nous retiendrons essentiellement ici l'inadéquation ou le mésusage des méthodes utilisées pour traiter des problèmes complexes.

Nonobstant ces importantes objections, il apparaît que l'épidémiologie sociale ainsi que la médecine légale ont peu accès, en pratique, à ce type d'études, essentiellement en raison de leurs objets d'étude. S'il reste possible a priori de mener des revues systématiques de la littérature, les études interventionnelles sont plus délicates, pour des raisons soit pratiques soit

éthiques. Un exemple extrême concerne les recherches sur les causes de décès, par exemple entre suicides et homicides. Il paraît difficile de monter une étude interventionnelle, en double aveugle et randomisée. Si des interventions de type prévention des situations de violence ou mode de prise en charge des personnes victimes de violences sont envisageables, il n'en reste pas moins qu'elles ne sont pas simples à mettre en place, et n'ont pas pour perspective l'assurance de mettre en évidence des effets facilement visibles ou « convaincants ». Quelques études interventionnelles ont été menées en épidémiologie sociale, par exemple pour mesurer l'impact du type de voisinage sur le bien-être à long terme,[31] ou pour mesurer la possibilité de réduire les inégalités socio-économiques de santé par l'aide financière.[32] Enfin, dans une optique où l'on favorise la production de résultats positifs, notamment par l'utilisation de tests d'hypothèse, une étude qui échouerait à mettre en évidence des résultats suffisamment évidents voire des résultats quantifiés, peut entraîner des effets pervers : les résultats seront interprétés dans le sens d'une conclusion négative – l'intervention n'a pas fonctionné. C'est d'autant plus dommageable que les rares études tentées sont justement publiées, en raison de leur rareté, alors que dans d'autres domaines, la « faiblesse » de leurs résultats ou la négativité de leurs résultats auraient entraîné une abstention de publication. Par ces mécanismes, l'épidémiologie sociale et la médecine légale se voient non seulement privées des moyens classiquement reconnus comme idéaux pour la production de savoir scientifique valable, mais se retrouvent également dans une situation où leurs tentatives peuvent les desservir, faute de techniques adaptées au problème.

Enfin, il est intéressant de remarquer une certaine mise en abîme concernant la médecine légale : la médecine légale est renvoyée par ses deux origines – la médecine et la loi – à devoir justifier de la valeur des preuves en termes de production scientifique, et de la valeur des preuves produites en matière légale, devant un tribunal. Comme nous venons de l'expliquer, même dans des conditions considérées comme idéales, expérimentales, comme en recherche clinique, la valeur de la preuve scientifique n'est pas évidente ou étayée de manière particulièrement convaincante – or la médecine légale, en tant que discipline médicale, place sa production scientifique sous les mêmes appréciations que le reste de la recherche biomédicale, alors qu'elle peut difficilement accéder aux plus hauts niveaux de preuve. Elle se retrouve dans la position paradoxale d'une science en panne d'un appareil efficace pour sa construction. Dans le même temps, la loi demande à la médecine légale de fournir le plus sûrement et le plus précisément – en termes par ailleurs fréquemment incompatibles avec le savoir et la pratique médicaux – des éléments attestant de la réalité d'un fait : la réalité d'un

viol, l'âge précis d'une personne, la crédibilité d'une victime ou encore la dangerosité d'un auteur présumé de violences.

La nécessité de méthodes variées et complémentaires

Le but de ces travaux n'est pas de délibérer sur la position philosophique à favoriser, si elle existe, concernant la valeur des savoirs scientifiques. Il convient cependant d'avoir en tête le contexte dans lequel le savoir est produit, selon quels choix, et pour quelle utilité.

Si l'*evidence-based medicine* – médecine basée sur les preuves – est de plus en plus critiquée, force est de constater que les solutions proposées pour pallier les failles dans la chaîne de production du savoir biomédical visent essentiellement à améliorer les pratiques reposant sur les techniques décrites ci-dessus (les tests d'hypothèse).[33–35] S'il est certain qu'un grand progrès peut être accompli si la méthodologie était mieux assise, comprise et exécutée, si les résultats étaient systématiquement publiés et accessibles, il n'en demeure pas moins qu'il n'est pas envisagé de changer ou diversifier qualitativement les approches. Il s'agit certainement d'une erreur fondamentale.

Il existe, concrètement, un besoin pour des méthodes variées et complémentaires, comme nous avons pu le montrer dans le cas de l'épidémiologie sociale et de la médecine légale.

Nous n'aborderons pas ici toutes les alternatives, embryonnaires ou bien établies, qui peuvent exister.[1] Nous présenterons ici deux types d'approches, que nous appliquerons aux champs de l'épidémiologie sociale et de la médecine légale.

Indépendamment de la vision de la science que l'on peut favoriser, entre inductionnisme et falsificationnisme, l'on peut tomber d'accord que le premier critère de scientificité tient dans le comportement méthodique, à opposer au comportement de « l'extravagant ».[19,36] Si une approche est retenue, un autre critère possible de scientificité peut être l'identification d'objets invariants, du moins invariants au regard d'un certain nombre de caractéristiques et d'observateurs.[37] De ce point de vue, nous proposerons le recours à des techniques qui permettent de mettre en évidence l'existence et de caractériser des groupes homogènes à partir des données – nous les appliquerons à l'épidémiologie sociale, en cherchant à dégager une typologie de recours aux soins et ses facteurs associés.

Un autre point technique important nous semble être l'intelligibilité et la préservation de la complexité des objets d'études. A ce titre, nous présenterons des techniques permettant de réduire un problème, un ensemble de données, à sa plus petite expression tout en en

préservant la complexité – nous appliquerons ces méthodes à un problème de médecine légale : l'estimation de l'âge chez les personnes ne pouvant justifier de leur âge, les adolescents migrants. Nous dégagerons, à partir de données, l'espace minimal de représentation de ces données, ce qui nous permettra de travailler efficacement et de les visualiser.

Types, typologies et classifications

Un des points de départ de ce travail repose sur le postulat suivant : il existe des objets, des invariants, qu'il est possible d'identifier. Il est possible que ces objets soient des objets absolus, invariants pour tout observateur, qu'ils soient des objets construits, relatifs à un processus d'observation-production, ou encore qu'ils soient des objets mixtes, hybrides. Ces objets peuvent également être fixes, immuables, ou bien déformables, soumis à l'action de facteurs temporels ou autres.

Ici, nous ne nous intéresserons pas à ces distinctions, seulement à nous donner les moyens théoriques et pratiques, étant donné un ensemble d'observations, de dégager de tels objets. Leur caractère artificiel ou « naturel » sera cependant discuté.

De l'identification d'objets au sein d'un espace de description particulier, il peut devenir possible de dégager une typologie : une description plus ou moins ordonnée, associée à une sémantique plus ou moins précise, de ces objets et, potentiellement, des relations qu'ils entretiennent implicitement ou explicitement.

Dans cette première partie, nous présenterons le cadre général qui explicite les notions de types, de typologies et de classifications. Nous partirons de définitions formelles et d'exemples historiques, pour aborder ensuite un point central de la question de l'objet en science : la dualité entre le discret et le continu. Nous esquisserons également un panorama des techniques et formalismes mathématiques correspondant à un discours sur la qualité des objets. Enfin, nous ferons un premier pas vers les outils développés disponibles sur ces bases formelles.

Définitions générales et exemples de typologies et classifications

Définitions d'une typologie, d'une classification

Définir ce qu'est une typologie est certainement académiquement plus simple à faire qu'à mettre en pratique, eu égard aux nombreux flous et pièges que recèle le concept. Notre propos n'est cependant pas de tout aborder ici. Nous proposons de nous baser sur 4 sources très différentes, qui nous permettent d'accéder à des points de vue croisés et variés. Prenons une définition accessible aisément au plus grand nombre, c'est-à-dire disponible sur Wikipedia :[38]

« *Une **typologie** est une démarche, souvent scientifique et toujours fondée sur une étude, consistant à définir un certain nombre de types afin de faciliter l'analyse, la classification et*

l'étude de réalités complexes. Par extension, le terme typologie désigne parfois la liste des types propres à un domaine d'étude. Le terme doit alors s'employer au singulier : la typologie (singulier) détaille un ensemble de types (pluriel). »

L'argument de scientificité ne présume pas de la méthode à employer, laquelle peut être de divers types. La nôtre, reposant aussi sur des données d'études, est davantage à classer du côté des méthodes types « quantitatives », bien, que comme nous l'expliquerons plus loin, elle ne doit pas être entendue en tant que telle, comme l'opposé du qualitatif. Elle se range dans le quantitatif dans le sens où elle n'est pas une méthode reposant sur l'analyse approfondie des cas restreints, en entretiens, par exemple.

Si l'on se rapporte au Trésor de la Langue Française informatisé,[39] une typologie est définie de la manière suivante :

« **-** Science de l'analyse et de la description des formes typiques d'une réalité complexe, permettant la classification; *p. méton.*, systèmes de types. Synon. *classification, systématique, taxinomie*. *Typologie des dictionnaires*. *La typologie, ou classification des types, s'exerce aussi bien sur les individus humains (envisagés du point de vue physique ou psychologique) que sur les espèces animales ou même les plantes et les simples phénomènes de la nature*. - *LING.* Méthode de classification des langues qui s'appuie sur leurs caractéristiques internes, telles qu'elles se dégagent d'une analyse rigoureuse.

- *PSYCHOL.* Étude des types humains du point de vue de leur constitution physique, mais, d'ordinaire, en corrélation avec leurs dispositions morales ou psychiques. *Typologie anthropologique*. *L'une des tendances les plus représentatives de la typologie contemporaine fut l'œuvre de Kretschmer sur les correspondances entre certains types morphologiques et certaines constitutions mentale*.

- *SOCIOL.* Étude des traits caractéristiques d'un ensemble de données empiriques complexes d'un phénomène social, en vue de les classer en types, en systèmes. Synon. moins usuel *typique*. *Typologie économique; typologie des structures sociales, des voies urbaines*. *Lorsqu'on utilise plusieurs critères de classification pour répartir des individus et que la combinaison des critères permet de définir des classes non hiérarchisées, on parle plutôt de* typologie *que de classification*.

- *SOCIOL. POL. Typologie politique*. Classification des régimes politiques fondée sur un ou plusieurs critères. *Les nomenclatures d'Aristote, de Bodin ou de Montesquieu sont des typologies politiques*.

Typologique, adj. [En parlant d'une classification] Qui appartient à la typologie; fondé sur une typologie. *Schéma typologique*. *C'est un des mérites de Kretschmer d'avoir toujours souligné, dans ses descriptions typologiques, l'ambivalence des traits, et les possibilités supérieures qu'offre chaque syndrome caractérologique à côté des types de qualité médiocre*. »

A ce titre, il est intéressant de remarquer les citations répétées pour illustrer ce qu'est une typologie, des domaines de la psychologie / psychiatrie, et des sciences sociales (sociologie, sciences politiques). Bizarrement, Linné est absent de ces définitions, là où il aurait pu être légitime, au moins historiquement, pour son œuvre de classification des règnes animaux et végétaux (Systema Naturae – 1735,[40] Species Plantarum – 1753).

En 1977, dans un numéro spécial de « Informatique et Sciences Humaines », Jean Paul Grémy et Marie Joëlle Le Moan publient leur compte-rendu d'étude de recherche commandée par la Délégation Générale à la Recherche Scientifique et Technique.[41] Le document s'intitule «Analyse de la démarche de construction de typologies dans les sciences sociales.»

D'après les auteurs, « élaborer une typologie consiste à distinguer, au sein d'un ensemble d'unités (individus, groupes d'individus, faits sociaux, etc.), des groupes que l'on puisse considérer comme homogènes d'un certain point de vue. Le contenu de cette notion d'homogénéité varie selon les auteurs et les domaines d'application ; elle se fonde généralement sur une certaine ressemblance définie à partir d'un sous-ensemble de caractéristiques servant à décrire les unités étudiées » (p15). En outre, ils ajoutent 2 conditions indépendantes mais nécessaires, que sont :

- L'exhaustivité : toute personne doit pouvoir être attribuée à un des types, un des groupes ;
- L'exclusivité : une personne ne peut appartenir à 2 groupes à la fois.

De façon intéressante, les auteurs proposent également 5 raisons principales qui mènent à se pencher sur la nécessité d'une typologie dans un domaine donné (pp16-17) :

- les exigences de l'application, car la typologie tend à fournir des moyens pratiques d'agir et d'évaluer l'action, si les groupes sont stables,
- l'importance du volume des données à traiter,
- l'inefficacité du modèle explicatif général,
- l'impossibilité d'aboutir à un modèle unique,
- la dynamique interne du système étudié, qui impose de penser en termes de typologie.

Non contents de recenser les motivations conduisant à la recherche d'une typologie, ils exposent également 3 méthodes de construction de typologies, dans le champ des sciences sociales :

- Les idéaux-types,
- La réduction de l'espace d'attributs,
- L'agrégation des unités.

La méthode dite des idéaux-types repose sur la construction de cas considérés comme typiques, voire archétypiques, lesquels accentuent, caricaturent dans une certaine mesure des situations réelles, analysées selon une grille de lecture prédéterminée. Les données réelles sont positionnées par rapport à ces idéaux. L'approche, de tradition Wébérienne, relève essentiellement d'une démarche abstraite, déductive. Une illustration de l'utilisation des idéaux-types dans le domaine de la santé est par exemple proposée par Parizot.[42]

La méthode par réduction de l'espace d'attributs est une méthode en 2 étapes. La première dénombre tous les attributs (ou dimension) a priori nécessaire pour caractériser les objets de la recherche, les personnes, les faits étudiés, pouvant mener à un espace de description de dimension très élevée. Toutes les combinaisons d'attributs sont candidates à être des types. La seconde étape vise précisément à filtrer tous les « faux » types, en confrontant l'espace d'attributs aux données réelles. Considérant que l'on recueille par l'expérience tous les vrais types observables, on réduit alors l'espace d'attributs initial au strict nécessaire en termes de dimension, quitte à le réarranger. En termes de techniques statistiques, cette méthode est

proche dans son idée des techniques d'analyse en composantes principales ou en correspondances multiples.

La méthode par agrégation des unités opère sensiblement différemment, dans le sens où l'on considère, d'une certaine manière, que l'on dispose déjà d'unités suffisamment typiques, et que l'on puisse comparer à d'autres selon un référentiel commun. Sous-entendu, parmi les données recueillies, la forme de recueil est telle que les objets sont déjà identifiés, et l'on peut, par différentiation, regrouper préférentiellement certaines unités avec une unité choisie comme « noyau ». L'approche est praticable « à la main » essentiellement lorsque l'on dispose de données concernant des unités en relativement petit nombre.
La méthode que nous mobiliserons dans les travaux présentés ici ne répond exactement d'aucune de ces 3 possibilités – éventuellement, elle se rapproche peut-être le plus de la 2e proposition. La 3e méthode, dans son principe, sert également de base à l'algorithme principal que nous utiliserons. De fait, les 3 méthodes exposées ne sont pas totalement exclusives les unes des autres, et sont davantage une possible décomposition des méthodes de base mobilisables. Comme nous le verrons, la méthode retenue fait intervenir, à divers degrés, des éléments de chacun de ces 3 types, selon l'étape : construction, validation, interprétation.
Enfin, il peut être profitable de consulter une 4e source, complémentaire, entre sciences sociales et typologies, classification, proposée par J. Coenen-Huther.[43] Il y aborde notamment l'utilité des typologies en termes de réduction de complexité d'un champ de connaissance, et y discute les liens entre discipline scientifique, consensus autour des objets de ces disciplines, et le degré d'acceptation des typologies au sein de ces disciplines. En particulier, il note à propos de la sociologie : « Depuis plusieurs décennies, on a le sentiment que certaines sciences humaines – la science économique, la démographie, la linguistique, l'ethnologie même – ont accepté plus facilement que d'autres l'idée de sélection de quelques dimensions privilégiées aboutissant à un objet relativement facile à isoler. La sociologie, discipline fourre-tout par excellence, ne peut se prévaloir d'un consensus solide sur la définition de son objet. Elle ne suscite l'accord ni sur la classification des variables dépendantes ni sur l'énumération des variables ayant une vocation particulière à être des variables indépendantes. »

L'exemple de Durkheim

En 1894, paraissent dans la Revue philosophique, « Les règles de la méthode sociologique », d'Emile Durkheim.[44] Il y expose les conditions nécessaires selon lui pour que la sociologie puisse accéder pleinement au statut de science. La sociologie sera la science du fait social, pris comme objet d'étude, et selon une méthode invariable et propre à la discipline.

Trois ans plus tard, en 1897, il publie comme exemple d'application de sa méthode « Le suicide »,[45] où celui-ci est considéré comme un fait social à part entière. Il étudie, sur des données statistiques, le suicide à la lumière de ce qu'il considère comme ses causes, au nombre de deux : l'intégration et la régulation (de la personne, par la société).

Il en dérive 4 types de suicide, selon les écarts qualitatifs par rapport à ces 2 causes : soit un défaut de régulation ou d'intégration sociale, soit à l'inverse un excès de régulation ou d'intégration sociale. A noter que dans l'ouvrage à proprement parler, Durkheim ne met en exergue que 3 des 4 types recensés, en leur consacrant chacun une section entière. Le 4ème existe (suicide fataliste) et est néanmoins bien explicitement énoncé en fin de chapitre V.

Il existe en ce sens 4 types de suicide (voir tableau 1) : le suicide égoïste (défaut d'intégration : la personne n'est pas reliée suffisamment à ses semblables), le suicide anomique (défaut de régulation : les normes sociales, les conduites ont des limites plus floues, tout peut sembler alors possible, mais le principe de réalité rattrape la personne, et lui prouve qu'il n'est pas omnipotent), le suicide altruiste (excès d'intégration : la personne perd de son autonomie au profit du groupe, le suicide peut devenir un devoir), et le suicide fataliste (excès de régulation : la société contraint trop la personne, qui n'a plus de marges de manœuvre suffisantes).

Tableau 1 : Différents types de suicides selon Durkheim

	Régulation	Intégration
Excès	Suicide fataliste	Suicide altruiste
Défaut	Suicide anomique	Suicide égoïste

A titre plus ou moins anecdotique, il est remarquable de constater que les facteurs de protection ou les facteurs de risque recensés encore à l'heure actuelle concernant le suicide, ou le comportement suicidaire, recoupent une bonne partie de ceux avancés par Durkheim, même s'il apparaît difficile de considérer la conduite suicidaire comme un fait social pur. En

ce sens, la conduite suicidaire – dont il a été envisagé assez tôt dans le processus de conception du manuel qu'elle devienne une entité nosologique à part entière dans la nomenclature du DSM-V à venir (« *Suicidal behavior disorder* ») (3),[46] et non plus une « simple issue » à une pathologie psychiatrique primaire – est un autre exemple de la tension qui préoccupe les partisans des facteurs environnementaux, ceux des facteurs génétiques, ou encore ceux qui se placent en faveur d'une interaction des 2 précédents. Les arguments biomoléculaires récents ne suffisant de toute façon pas à lever l'ambigüité, ou à disqualifier l'approche sociologique du suicide. Le comportement suicidaire, l'impulsivité, les relations entre dépression majeure et suicide sont actuellement approchés via les variations et mutations, le polymorphisme génétique impactant la sérotonine, son utilisation, sa distribution, et les structures neuronales associées, de l'échelle sub-neuronale aux réseaux synaptiques.[47–50] Pour autant, des études récentes conduites en Orient remettent en cause les approches et facteurs de risques classiques occidentaux, et privilégient sans le mentionner directement, l'importance du milieu et de la société.[51]

Les travaux fondateurs de Durkheim sur la sociologie du suicide se poursuivent en droite ligne par ceux de Maurice Halbwachs, « les causes du suicide ».[52] Notamment, il réexamine, analyse de nouveau, prolonge l'étude de Durkheim, critique tant les sources de données que les méthodes employées, et en propose d'autres. La place de Halbwachs est ici importante, dans le sens où si l'on replace « les causes du suicide » dans son œuvre, cet ouvrage continue, et renforce l'explication sociologique du suicide. En effet, il convient de suivre le raisonnement de l'auteur quand il expose que les souvenirs individuels sont évoqués par des événements, des personnes extérieures à soi (voir la notion de « mémoire collective »[53]) – sans parler de la toute relative robustesse absolue d'un souvenir, d'une évocation, et la plasticité mémorielle. De là, c'est tout le fonctionnement intellectuel, les idéations, qui – pour autant que leur support matériel soit bien composé de neurones, de connexions et de transmissions électrochimiques – relèvent de l'expérience sociale. Les éléments positivistes, matérialistes du suicide, ne sont aucunement exclusifs des éléments sociologiques qui l'instancient pour une personne, une histoire données. Par ailleurs, Halbwachs s'intéressera tout naturellement aux formes, au sens large, donc en un sens à la typologie, de la société : « la morphologie sociale ».[54]

Enfin, pour élargir le sujet, un pendant des typologies liées aux causes et pathologies dites relevant de la psychiatrie, comme le suicide, l'impulsivité, les troubles de la personnalité (dont les personnalités borderline, antisociale ou psychopathique) tient dans la volonté,

possiblement dans la dérive, de vouloir rendre les catégories déterminantes des personnes qui sont censées les constituer, de les y enfermer. Ainsi, en criminologie, et suivant les faits divers récurrents, tout un chacun a pu entendre parler des échelles « actuarielles », en liaison avec la notion de dangerosité.

L'idée sous-jacente est alors de valider des échelles prédictives de « dangerosité » pour l'autre (plus souvent que pour lui-même ; les échelles de risque suicidaire étant classées à part), et de fournir des éléments objectifs quant à la conduite à tenir et à la réponse adaptée de la société face à telle personne considérée plus ou moins pathologique. Un certain nombre d'échelles ont été proposées (par exemple : l'ERRRS (Hanson, 1997), le SACJ-Min de Thornton (Grubin, 1998) ou encore la Static-99 (Hanson, Thornton) 1999 puis Static 2002, entre autres). Certaines sont d'usage courant dans plusieurs pays dans le cadre des expertises médico-légales, tels que les Etats-Unis, le Canada et la Belgique.

En 2008, un rapport canadien proposant une méta-analyse sur l'acuité des outils de prédiction des récidives chez les délinquants sexuels met en évidence une meilleure précision des échelles actuarielles par rapport au jugement du clinicien.[55] Il est cependant notable, et c'est un problème central qui dépasse le cadre de ces échelles, que les auteurs aient pris soin de présenter l'opposition cliniciens / partisans des échelles comme un antagonisme entre désir de compréhension et d'explication d'un cas particulier, et désir de prédire au mieux, au sacrifice même total de toute intelligibilité du modèle.

Si l'idée initiale est séduisante (déterminer qui récidivera), elle pose immédiatement des problèmes éthiques, sociaux et sociétaux. Dans l'hypothèse d'un pouvoir prédictif très important, quel choix, quelle prise en charge correspondant à, disons, 70% de risque de récidive ? Un haut risque supposé constitutionnel de la personne plus qu'environnemental ou conjoncturel devrait-il déterminer l'enfermement, la condamnation, l'obligation de traitement de cette personne ? Dans le système de soins actuels, même pour des interventions d'importance a priori mineure, les faux positif ou les faux négatifs sont ressentis facilement comme intolérables, dans une culture du déterminisme, du mécanisme et du risque zéro. Devrait-on alors souffrir que les « faux positifs » inévitables prédits par ces outils soient traités différentiellement de la population « normale » ?

L'Académie nationale de médecine s'est prononcée sur la question de l'évaluation de la dangerosité, et la confrontation évaluation personnelle ou personnalisée versus approche statistique des échelles actuarielles. Elle souligne l'intérêt des deux outils, et profite finalement de l'opportunité pour faire l'état des lieux des experts disponibles face à la demande croissante d'expertise médico-légale et réclamer une formation de chacun des

acteurs à ces outils statistiques.[56] Néanmoins, il est bien noté la différence entre la mission expertale première qui est à vocation diagnostique, tandis que la mission expertale visant l'évaluation de la dangerosité relève désormais du domaine du pronostic, donc de la prédiction. Cette activité de prédiction, selon les auteurs, ne sauraient être de la seule responsabilité de la médecine.[57]

Dans le prolongement de ce raisonnement, l'idée d'une médecine prédictive s'accompagne voire se confond généralement avec l'idée d'une médecine de prédiction. Cette médecine de prédiction repose tacitement et historiquement, sur des considérations génétiques, sur la disponibilité des informations génomiques. Elle sous-entend donc une forte proéminence du code génétique sur les autres types de facteurs explicatifs, dont les facteurs sociaux.[58] La connaissance du code devrait pouvoir révéler les variations individuelles, livrant elles-mêmes les clés des trajectoires sanitaires, parfois même sociales, de chacun. Parmi les voix qui se font le plus entendre, la médecine prédictive est une médecine du gène ; certains ont pu critiquer assez sévèrement cette approche,[59–61] au final très peu élaborée et ne rendant pas compte des savoirs et des connaissances en complexité issues de plusieurs disciplines – notamment en sciences physiques.

Dualité qualitatif-quantitatif, dualité discret-continu

La notion de typologie, de classification, donc de classes, entretient des relations ambiguës avec les concepts supposés opposés que sont le continu et le discret, ou le qualitatif et le quantitatif. En effet, à première vue, si l'on désire établir une typologie ou une classification, on sous-entend qu'il existe des groupes séparés les uns des autres, si possible en nombre fini.[41] L'existence d'une classification s'oppose a priori à la possibilité d'un continuum. Il s'agit néanmoins d'être prudent. La détermination de classes à partir de données présente plusieurs difficultés, et peut mener à la conclusion de classes artéfactuelles. Pensons au cas inverse, où les groupes sont déterminés a priori, et sont connus, par exemple dans le cas d'un essai randomisé contrôlé. Un groupe recevra un placebo tandis que l'autre recevra l'intervention, supposée suffisamment efficace pour que l'on en mesure simplement l'effet. Supposons pour les besoins de la démonstration, que cette intervention vise à diminuer le poids des personnes. On testera l'efficacité de cette intervention en comparant les moyennes des poids dans les deux groupes, poids dont on fera l'hypothèse qu'ils suivent une loi gaussienne. Le problème de rejeter l'hypothèse nulle, à ce stade, c'est-à-dire, que les

moyennes sont identiques dans les deux groupes, revient à celui de « démêler » les deux distributions gaussiennes. Or, les deux distributions ont des queues infinies et sont symétriques : elles sont nécessairement entremêlées. Que dire des points, des personnes qui se trouvent dans la zone intermédiaire, entre les deux moyennes ? Sont-ils plutôt dans un groupe ou dans l'autre, si l'on n'est pas au fait de leur appartenance a priori ?
Dans une situation plus réaliste, où l'on va recueillir des observations diverses, sans constitution a priori de groupes distincts, il n'est pas impossible de « couper » artificiellement un ensemble de personnes simplement parce que la densité des observations est plus faible dans une zone donnée, par analogie avec le cas décrit précédemment pour les deux gaussiennes. Se pose donc la question de l'échantillonnage, où il s'agit de ne pas soit surreprésenter certaines observations, soit sous-représenter certaines autres. Dans un tel cas, il n'est pas impossible qu'il n'existe pas en soi de classe « naturelle », mais plutôt un continuum.

Prenons un autre cas, sensiblement différent, mais pouvant se représenter de la même manière : deux zones plus densément peuplées en observations, agissant comme deux pôles relativement éloignés, et une zone intermédiaire moins dense en observations, mais présentant bel et bien des observations. On identifiera vraisemblablement deux classes correspondant aux deux pôles. Néanmoins, que faire des observations intermédiaires ? Faut-il les rejeter, décrétant qu'elles n'appartiennent ni à une classe ni à l'autre, ou bien les attribuer à l'une des deux classes, selon leur proximité à un pôle plutôt que l'autre ?
Supposons en effet qu'il existe deux classes « naturelles », c'est-à-dire, pouvant s'expliquer par une structure spécifique, un mécanisme entretenant leur existence par deux voies séparées. On peut par exemple prendre au pied de la lettre l'idée de deux pôles magnétiques différents : il existe deux aimants, bien identifiables. Les données sur lesquelles on se base sont des observations à un instant donné, et il est possible que les observations de la zone intermédiaire soient en fait des positions de transition entre les deux classes. C'est potentiellement le cas si les valeurs des observations sont continues, et non strictement quantifiées. Comme on le voit, il peut exister des classes, mais dont l'assise est plongée dans un espace continu : les observations ne sont pas tenues de « sauter » d'une classe à l'autre, mais peuvent suivre une transition plus ou moins lisse de l'une à l'autre. La remarque peut paraître purement théorique. Cependant, que penser des deux cas suivants : 1) il existe bien deux pôles, mais la réalité est telle que finalement, les personnes ne font qu'alterner entre ces deux pôles, 2) il existe autant de personnes appartenant à chacun des pôles que de personnes « en transition » ?

La notion de classe est intéressante si elle représente soit un invariant, un groupe particulièrement stable et représentatif, soit un pôle entraînant un comportement dynamique (par exemple, la circulation des personnes de l'un à l'autre pôle). Les deux cas ne seront pas à interpréter de la même manière. Enfin, Il faut encore prendre garde à ne pas confondre classe et élément constituant de cette classe, de la même manière qu'il existe des faits sociaux, au sens de Durkheim, survivant à l'individu et s'imposant à lui, et des personnes dotées de psychologies individuelles. Une classe peut très bien se définir par un ensemble de caractéristiques qui ne varieront pas, validant par là son statut d'objet scientifique, bien que les personnes les constituant ne soient évidemment pas toujours les mêmes. Le passage d'une personne d'une classe à une autre n'infirme l'existence de ces classes : ce sont deux niveaux d'interprétation différents.

De la sorte, il faut d'une part prêter attention au possible caractère artificiel des classes identifiées, et d'autre part, s'interroger sur le caractère purement discret, quantifié de l'espace de description de ces classes.

Opposition des sciences humaines et des sciences dures, rôle du social dans la détermination de l'objet scientifique

On oppose généralement sciences humaines et sciences dures. Ces oppositions tiennent par exemple dans le caractère expérimental plus accessible et reproductible dans le cas des sciences dures, en la possibilité de mesurer des quantités relativement précisément ou encore de se reposer sur des énoncés généraux qui tiennent correctement l'épreuve du temps et des tests. Les objets en sciences dures en sont presque présentés comme donnés a priori, ce qui est pourtant loin d'être le cas. D'une part, il n'y a d'observations descriptibles, donc d'observations, sans cadre théorique, implicite ou non. Les objets sont donc au moins en partie construits, et n'ont pas l'heur de bénéficier d'un statut parfaitement consensuel ou fixé. Il se peut en outre que certains objets, vus comme radicalement différents selon le point de vue théorique, puissent finalement être caractérisés effectivement de manières apparemment paradoxales, comme pour le cas de la lumière en physique.
L'histoire des sciences regorge ainsi d'exemples tendant à montrer la composante éminemment sociale de la construction des savoirs et des objets auxquels ils se rapportent.[62–

[65] La science demeure une activité humaine, qui plus est de plus en plus segmentée, spécialisée, ce qui la ramène à la production de groupes plus ou moins restreints.

Aussi, nous voulons évacuer ici la critique qui se voudrait définitive d'une approche de l'épidémiologie sociale qui soit en partie basée sur l'identification de typologies et de classes, sous le prétexte qu'elle serait par trop biaisée ou reflétant davantage la position des chercheurs que l'existence réelle de telles classes. Pour autant que l'on soit conscient des limites et des conditions d'applications et d'interprétations des méthodes que l'on mobilise, de même que de la nécessaire composante sociale dans la découverte ou co-découverte d'un objet, d'une classe, et que la caractérisation de cet objet, de cette classe, s'effectue de toute façon à la lumière d'un vocabulaire, d'une théorie, la valeur de notre démarche ne peut déméríter.

A ce jour, le distinguo entre sciences dures et sciences humaines tient certainement davantage à la « jeunesse » des secondes, mais également à une forme d'immaturité des premières, dans le sens où elles s'avèrent incapables de rendre compte de la part de social et de subjectivité, en pratique, dans la construction de ses objets.
L'état actuel nous semble se résumer à deux considérations importantes selon lesquelles d'une part, nous ne travaillons pas sur des « données, mais sur des obtenus »,[64,66] et, d'autre part, le fait scientifique est produit socialement.[64] Cependant, peut-on dire du fait scientifique qu'il est un fait social ? Cette question ne nous paraît pas de résolution si évidente, si l'on s'en tient aux 4 critères de Durkheim définissant le fait social :[44]

- Généralité : un objet qui doit être représenté un minimum dans la société, sur un horizon temporel et spatial donné, et invariant sur cet horizon. Au-delà de cet horizon, le fait peut voir sa forme changer. Ce critère rappelle la distinction qui existe entre relativité restreinte et relativité générale ;
- Extériorité : l'objet est extérieur aux individus, c'est-à-dire que s'il n'existe que parce que des individus existent, c'est uniquement parce que ceux-ci font société. L'objet caractérise et relève de la société, pas de l'individu lui-même ;
- Pouvoir coercitif : le fait social s'impose aux individus, et en contraint les activités et les comportements ;
- Historicité : il est nécessaire, pour se distinguer d'une mode, qu'un certain recul temporel soit possible afin d'identifier et caractériser la stabilité de l'objet social.

Aussi, à première vue, un fait scientifique répond du critère de généralité, d'une forme d'extériorité, est coercitif dans le sens où des éléments comme la physiologie contraignent les modes d'être, et le critère d'historicité peut se discuter par le processus même de découverte et de fabrication du fait scientifique.

Dualités discret-continu : physique fondamentale, théorie de la communication et échantillonnage

Il semble être naturel que l'on soit enclin soit à préférer une description discrète de la nature, soit une description continue de celle-ci. L'histoire des sciences recèle de plusieurs exemples où selon les savants, selon le point de vue et le moment de l'histoire, une description discrète est avancée plutôt qu'une description continue, et inversement – ce, parfois même de manière concurrente et simultanée. Les descriptions atomistes remontent au moins à l'antiquité (*De Natura Rerum*, Lucrèce),[67] et les interprétations continues également (le fleuve d'Héraclite).[68] Le débat s'est fait particulièrement plus concret lorsque l'on en est arrivé au problème du comportement de la lumière, et par extension avec la mécanique quantique, au comportement de ce que l'on appelle actuellement "les particules élémentaires". Certains soutenaient que la lumière étaient constituée de petits grains de lumière, équivalents de billes ou autres figurations, quand d'autres soutenaient que la lumière devait se décrire comme une onde capable de se propager dans l'espace et au cours du temps, donc un élément continu. Selon les configurations (calcul de la pression due aux impacts de photons, ces grains de lumières, ou l'effet photo-électrique décrit par Einstein pour le cas discret ; calcul des figures de diffraction et d'interférences, pour le cas continu), les deux théories ont largement prouvé leur pertinence. La coexistence d'un état à la fois discret et d'un état continu est ainsi acceptée, sans qu'elle soit pour autant vécue comme satisfaisante. Cette coexistence – cette dualité – a néanmoins été mieux vécue, semble-t-il pour le cas de la lumière, que pour l'extension de ce principe de dualité à l'ensemble de la matière. En effet, via l'introduction d'une longueur d'onde associée à n'importe quelle particule élémentaire par De Broglie, notion concrétisée et parfaitement intégrée par la mécanique quantique (qui est une mécanique ondulatoire), ce n'est pas seulement la lumière qui se trouve avoir un comportement discret et continu, mais bien toute la gamme de la matière, sous certaines conditions (petites échelles, en général). C'est ce que l'on nomme la dualité "onde-corpuscule". Des expériences célèbres, comme celle dite des fentes d'Young, ont mis en évidence cette dualité troublante pour des particules matérielles comme les électrons : ceux-ci, s'ils sont examinés dans des conditions similaires et adaptées à

celles qui permettent de montrer que la lumière peut créer des figures d'interférences (phénomènes inexplicables si la lumière n'était que corpusculaire), forment également des figures d'interférence, attestant d'un comportement ondulatoire des particules. Enfin, tout récemment, une expérience a permis de mettre en évidence simultanément, et non alternativement ou selon des schémas expérimentaux différents, le caractère ondulatoire et corpusculaire de la lumière.[69]

Ainsi, une science dure comme la physique accepte la dualité du discret et du continu, de deux descriptions qui correspondent à des manipulations conceptuelles reconnues comme radicalement différentes, sinon opposées. Le domaine de la théorie de l'information a reconnu, d'une autre manière, la possibilité de décrire toute l'information contenue dans un signal continu par un ensemble de valeurs discrètes, sous certaines conditions. Ces conditions sont énoncées par le théorème fondamental de l'échantillonnage, ou théorème de Shannon-Weaver[1].[70,71] Par ailleurs, on sait également que la qualité d'un échantillonnage, l'information contenue dans un ensemble d'observations discrètes, dépend de considérations de résolution, donc d'échelle d'observation.

En statistiques "générales", la notion d'échantillonnage existe également, et son but poursuivi est similaire à celui d'un bon échantillonnage en termes de théorie de l'information : il s'agit d'obtenir un ensemble d'observations qui soient fidèles à la "réalité", c'est-à-dire, dont on puisse inférer des vérités générales. Ainsi, dans l'idéal, l'échantillon est censé pouvoir correspondre à un ensemble suffisant d'observations qui permettent de "reconstruire" le signal réel, de la même manière que l'on échantillonne de la musique, une image ou un film dans une procédure que l'on appelle désormais "la numérisation". Les règles qui guident un bon échantillonnage en statistiques sont néanmoins orientées un peu différemment : elles sont liées, de manière générale, à un plan d'expérience plus ou moins complexe. Selon la discipline (plan de sondage, essai clinique, étude observationnelle), les techniques mobilisées et les objectifs plus fins seront sensiblement différents – les plans d'expérience étant différents.[72] L'un dans l'autre, à l'heure actuelle, il manque très certainement une connaissance plus robuste qui serait capable de définir ou d'étendre de manière satisfaisante la définition et l'application des règles d'un "bon échantillonnage" en biomédecine – la notion, par exemple, de puissance et de risque de première espèce étant des critères bien faibles et finalement relativement peu

[1] Le théorème de l'échantillonnage énonce que tout signal de spectre (d'énergie) fini, soit a priori les signaux « réels », peut être échantillonné (numérisé) sans aucune perte d'information et reconstitué ad integrum à partir des échantillons, si la fréquence d'échantillonnage est au moins double de la fréquence maximale du signal.

utiles pour témoigner de la réalité, donc pour servir de règles fiables pour une démarche d'inférence.

On le voit, certaines disciplines ont pleinement accepté que qualitatif et quantitatif puissent être unis et duaux l'un de l'autre, certaines étant même parvenue à formuler explicitement un lien de passage de l'un à l'autre. Nous en sommes encore loin, semble-t-il, en biomédecine ou en sciences humaines en général – dont l'économie ne fait pas exception. La psychiatrie, ou plus largement le domaine de la santé mentale, avec notamment ses systèmes de classifications en entités nosologiques séparées, semblent osciller entre, ou à tout le moins, persister à vouloir opposer, descriptions catégorielles et descriptions continues.[73–76]

Il existe enfin une autre façon de concilier discret et continu, quantité et qualité, que nous exposons plus en détails ci-dessous, dont nous avons pu par ailleurs montrer l'applicabilité en santé mentale.[1,77]

Approches mathématiques de la qualité des objets

Les mathématiques ont longtemps séparé le discret et le continu. Quelques branches et théorèmes permettent de naviguer entre ces deux univers, mais ceux-ci restent encore très autonomes et présentent des développements également riches. En particulier, les mathématiques du discret ont suscité un regain d'intérêt au cours des dernière décades. Parmi les domaines des mathématiques jonglant avec ces deux aspects, la place de la géométrie est notable, dans le sens où elle est un point de rencontre privilégié pour la plupart des autres grands domaines : géométrie algébrique, géométrie différentielle, géométrie analytique...

Les développements de la topologie et la géométrie différentielle ont inspiré les travaux de R. Thom, qui a fondé la théorie dite des catastrophes : une théorie permettant de représenter ou d'expliquer le discret au sein d'un cadre continu.[78] Nous en abordons rapidement les principes ci-dessous, au titre de la topologie, des variétés, et du lien discret-continu.

Topologie et variétés

La topologie est l'étude du "lieu", et en mathématiques, la topologie s'intéresse essentiellement aux transformations continues que l'on peut faire subir à un espace donné, souvent caractérisé par sa dimension et une manière de mesurer des distances dans cet

espace.[79] La topologie permet notamment d'établir des classifications d'espaces selon leurs qualités topologiques (dimensions, distance, présence de "trous", de "nœuds"), ou des relations d'équivalences topologiques entre espaces qui selon d'autres considérations, pourraient sembler différents.

Les variétés sont, en topologie et géométrie différentielles, des objets géométriques obtenus par des opérations de recollement d'espaces particuliers, des ouverts d'espaces vectoriels.[80] Le vocabulaire emprunte au lexique de la cartographie : on y parle certes de recollement, mais aussi de cartes locales, d'atlas. L'idée sous-jacente est que l'utilisation de référentiels, de manières de décrire localement l'information, permet, si ces référentiels se recoupent au moins partiellement, sur leurs bords, et que l'on peut définir une transformation établissant une correspondance entre leurs deux manières de situer un point sur leur carte, de décrire tout l'espace que recouvre l'ensemble de ces cartes, en un seul grand atlas : on peut donc, par des collections d'informations parcellaires et souvent plus simples, reconstituer des ensembles complexes. Nous ne sommes pas loin de la notion d'échantillonnage décrit plus haut. Ces objets sont les objets de base utilisés en géométrie différentielle, et permettent de légitimer tout un ensemble de manipulations et de résultats les concernant.

Théories des catastrophes

Thom a fondé sa théorie des catastrophes sur l'intuition très topologique que beaucoup de phénomènes usuels, observables, en "régime permanent", sont des phénomènes simples et triviaux, ne nécessitant pas une complexité de représentation importante.[78,81] En revanche, il existe des lieux et des instants, où la qualité de la représentation utilisée peut changer du tout au tout, peut qualitativement changer, par exemple bifurquer d'un comportement vers un autre brutalement. L'idée force des catastrophes est de décrire, selon les dimensions de l'espace et les paramètres qui le structurent, les conditions et la forme de ces changements brusques – nommés catastrophes. En basses dimensions, ces catastrophes peuvent également être totalement décrites et avoir force de classification. Pour Thom, le changement de qualité est équivalent à la notion de catastrophe.

Depuis, la théorie des catastrophes a été plusieurs fois critiquée,[82] finalement relativement peu appliquée et testée dans des environnements ou cas contrôlables. Néanmoins, que ce soit pour ou contre, les arguments pêchent par un certain manque de force, sinon de pertinence. Enfin, initialement autonome, la théorie s'est trouvée de nouveau légitimée par son intégration dans

une théorie plus vaste : la théorie des systèmes dynamiques. L'application pratique de ses concepts reste à cette heure marginale voire inexistante en biomédecine, en épidémiologie ou en santé publique, où les idées ont pu être évoquées et finalement assez rarement mises en œuvre – elle a été plus souvent rencontrée en biologie ou en écologie.

Variables de contrôle

Une catastrophe – ou un système dynamique – est caractérisée de manière duale, là aussi : les variables décrivant le système d'intérêt, reliées entre elles par des relations différentielles (l'accroissement d'une variable dans le temps exprimée selon une fonction de cette variable ou d'autres, cette fonction n'étant pas nécessairement linéaire en les variables), et ces relations sont paramétrées par un autre ensemble de variables, qui peuvent être vues comme des variables de contrôles – ou pour reprendre un terminologie plus connue en statistiques, des variables explicatives.[1,83] Ces variables de contrôle sont donc les paramètres du système, et elles règlent la structure de ce système : ce sont elles et les valeurs qu'elles peuvent prendre qui détermineront la présence et le type de catastrophes rencontrées par un système.

Une analyse des systèmes selon l'approche des systèmes dynamiques ou des catastrophes permet donc d'éclairer notre compréhension de ces systèmes selon plusieurs aspects : le comportement du système en termes de ses variables descriptives, le comportement structurel du système en termes d'un ensemble de variables qui le structure directement (variables "explicatives"), et d'en déterminer les configurations critiques, c'est-à-dire, les configurations qui peuvent mener à des présentations qualitativement différentes de ce système, au point où l'on pourrait penser qu'il s'agit de systèmes différents, s'ils étaient étudiés spécifiquement et exclusivement des autres configurations possibles. Cette analyse doit donc permettre de caractériser le système, mais également d'en dégager des objets spécifiques, support d'une typologie des configurations de ce système.

Malheureusement, à l'heure actuelle, nous disposons encore de peu de techniques qui autorisent à aborder les études, en particulier observationnelles,[84–88] à la lumière de cette théorie. Il existe une autre raison, partiellement liée au manque de techniques, qui est que les systèmes dynamiques prennent pour parti d'étudier plus spécifiquement les systèmes évoluant dans le temps – les séries temporelles. Or, en biomédecine, le temps est souvent aboli (par exemple dans les essais cliniques randomisés, qui ne sont pas explicitement des études de survie), ou bien peu échantillonné (par exemple, au mieux dans certaines grandes cohortes, type Whitehall II[89] ou Gazel,[90] qui commencent à prendre de l'âge, et à proposer un

échantillonnage temporel plus satisfaisant, mais pas nécessairement adapté à l'étude de n'importe quel aspect, puisqu'il est au mieux de l'ordre de l'année).

Vers les outils sous-tendant la recherche de typologies

Différences opérationnelles entre typologie et classification

Des différences peuvent être faites entre typologie et classification. Jusqu'ici, nous avons traité l'une comme l'autre et inversement, sans distinction particulière. Parmi les différences qui peuvent séparer sémantiquement les notions de classification et de typologie, nous retiendrons essentiellement une différence opérationnelle, voire temporelle. Une typologie est avant tout une démarche d'identification, de réduction de la complexité, avec sa part d'arbitraire de la part de celui qui cherche à réduire cette complexité.[66,91] Cette opération d'identification et de réduction se veut une opération liée au savoir, à la compréhension, selon certains aspects plus ou moins féconds ou fonctionnels – finalement : utiles – du monde environnant. Elle n'a pas pour but de réduire la complexité du monde à cette unique typologie, c'est-à-dire de réduire la complexité de situations relevant de dimensions autres ou supplémentaires que celles considérées pour la construction de la typologie à la complexité de la typologie dégagée. Similairement, si une typologie a été dégagée dans une démarche de prédiction ou d'explication de phénomènes associés à cette typologie, elle n'a pas pour vocation de pouvoir prédire ou expliquer tout autre phénomène qui ne serait pas explicitement relié à cette typologie. Enfin, la construction d'une typologie implique en sus de points de vue forcément non exhaustifs, la notion d'échelle : les objets obtenus et structurant une typologie ne permet en aucun cas de réduire exclusivement les éléments aux types qu'ils constituent. Enfin, les relations inter et intra-types, ainsi que les relations inter-échelles peuvent rester mal connues.

Une classification relève davantage, pour nous, d'une opération reposant sur l'utilisation d'une typologie, et des types qui la structurent. Une classification consiste en l'attribution d'un type ou d'une classe à un individu qui se présente à nous, avec la possibilité de se « tromper » de classe. Cette opération est au moins aussi délicate sinon plus dangereuse que celle qui permet de construire une typologie : elle est délicate au sens qu'il est toujours délicat d'utiliser un outil, l'outil ne renseignant pas directement sur les intentions de celui qui s'en sert. Par ailleurs, cette opération, étant données les limites de validité exposée concernant la construction et l'interprétation d'une typologie, possède un domaine de validité propre limité.

En effet, si l'on considère que l'on ne peut réduire un élément d'un type à ce type exclusivement, l'opération de classification portera pourtant bel et bien sur des éléments et non directement sur les types, puisqu'il s'agit d'attribuer un type à chaque élément considérer.
Ces mises en garde faites, nous nous référerons dans la suite toujours indifféremment à la notion de classification ou de typologie, même si le terme de typologie serait potentiellement à préférer. La raison principale à cette indistinction tient dans le fait suivant : les techniques que nous emploierons dans la suite sont rarement sinon jamais décrites dans la littérature comme des outils de construction de typologie, mais bel et bien des outils de classifications, éventuellement de classement, ou finalement comme techniques de partitionnement ou d'identification de formes.

Identifier des types, reconnaître la dimension et la forme de l'espace

Les deux grands types d'outils que nous mobiliserons dans la suite de ces travaux seront les techniques dites de *clustering* ou de *pattern recognition* (indifféremment appelées, en général, dans la littérature, techniques de classification, de partitionnement ou de classement),[92–96] et les techniques de réduction de la dimension des espaces,[97–100] opération qui entraîne de facto l'apprentissage de la forme des espaces considérés.
A partir d'un ensemble de données, obtenues par des mesures, des observations, on dispose d'un espace d'une certaine dimension. Dans cet espace, nous chercherons à déterminer s'il existe des objets, des types qui structurent l'espace : nous utiliserons des techniques de *clustering*. En procédant ainsi, nous ne cherchons pas réellement à identifier ou à réduire la complexité de l'espace lui-même, mais la complexité "en nombres", et "en échelle" de l'espace : on identifie un nombre limité de types, qui permettent de grouper de bien plus nombreux éléments disparates et variés décrivant tout l'espace.
Un autre moyen de réduire la complexité de ce que nous manipulons tient dans la réduction de l'espace lui-même, étant entendu que l'espace initial est constitué d'observations qui ont toutes les chances d'être partiellement redondantes – mais possédant un sens plus ou moins évident pour l'intelligence ou eu égard aux cadres théoriques historiquement acquis et construits. En cherchant la plus petite dimension nécessaire et suffisante à décrire l'espace initial, puis en réduisant à cette dimension l'espace initial à un ensemble de données équivalent en termes de complexité, nous obtenons à la fois des données plus facilement manipulables, de complexité apparente réduite mais de complexité "réelle" conservée.

Inégalités et systèmes de soins, système judiciaire

Inégalités de santé et inégalités sociales de santé : généralités

Il existe des disparités de santé : à un instant donné ou sur toute la durée d'une vie, toutes les personnes ne présentent pas le même état de santé, qu'ils soient a priori exposés à des facteurs différents observables ou non. Tout un pan de la recherche biomédicale, directement ou indirectement, vise à identifier les raisons, à défaut les facteurs associés, à ces différences, à ces inégalités face à la santé. Parmi les grands ensembles de facteurs associés à des différences de santé, nous distinguons les facteurs environnementaux, les facteurs génétiques – minoritaires – les facteurs liés au système de soins, les comportements de santé. Au-delà de l'échelle individuelle, nous savons désormais qu'une part de ces inégalités de santé est constituée d'inégalités d'origine sociale – par exemple, en lien avec la catégorie professionnelle.[4] De même qu'il existe des interactions gène-environnement qui soient associées à des différences éventuellement plus accentuées que celles qui seraient attribuables d'une part à un profil génétique particulier et d'autre part à une exposition environnementale spécifique, il existe des interactions entre facteurs non sociaux et facteurs sociaux. Plusieurs modèles théoriques ont pu être proposés, concernant les déterminants de la santé et leurs interactions : modèle de Dahlgren et Whitehead (1991),[101] modèle de Diderichsen et Hallqvist (1998), ou encore le modèle de la CSDH/CDSS de l'OMS[102] (Commission on social determinants of health / Commission des déterminants sociaux de la santé) et d'autres variantes.[103] Le modèle de Dahlgren et Whitehead, modèle « en arc-en-ciel », propose une architecture des déterminants en 4 niveaux, non indépendants les uns des autres. Le premier niveau, « facteurs liés au style de vie personnel », concerne les comportements et les styles de vie personnels. Le deuxième niveau, « réseaux sociaux et communautaires », prend en compte les influences sociales et collectives, le soutien social. Le troisième niveau, « facteurs liés aux conduites de vie et de travail », fait référence à l'accès au travail, aux services de santé, aux équipements essentiels tels que l'habitat, la nourriture. Le quatrième et dernier niveau, « conditions socio-économiques, culturelles et environnementales », intègre les facteurs qui influencent la société dans son ensemble. Cela peut correspondre au niveau de vie global d'une société, les relations et différences liées au genre ou certaines représentations et croyances culturelles.

Le modèle de la CSDH se présente sensiblement différemment, et organise les déterminants reconnus en déterminants dits « structurels » des inégalités sociales de santé et des déterminants « intermédiaires » de l'état de santé. Les déterminants structurels sont ceux qui font référence aux contextes socio-économique et politique d'un pays. Les déterminants intermédiaires concernent les conditions matérielles, psychologiques, les facteurs biologiques et génétiques, les comportements ou encore l'accès au système de santé.
Bien entendu, comme pour beaucoup de facteurs, ces inégalités sociales sont associées à des effets moyens, et il n'y a pas de déterminisme strict, directement transposable à l'échelle de l'individu ; cet aspect n'est pas propre au caractère social.
La reconnaissance d'inégalités sociales de santé est une opportunité pour lier, renforcer les liens entre santé, biomédecine, santé publique et politiques de santé. Pourtant, les interventions visant explicitement les inégalités sociales, considérées comme levier particulier, ne semblent pas majoritairement représentées.

Le lien entre la santé et la situation socio-économique d'une personne a été mis en évidence depuis plusieurs décennies.[104] Ce lien est d'autant plus intéressant que, si l'on constate et enregistre un progrès biomédical, technologique continu ces 50 dernières années, les inégalités sociales concernant la mortalité perdurent, voire s'aggravent entre groupes socioéconomiques, et ce particulièrement depuis les années 90.[105]

On ne saurait s'arrêter à ce constat, et il est devenu indispensable de chercher des éléments explicatifs à ce tableau, de mieux saisir les mécanismes à l'œuvre derrière l'observation des inégalités sociales de santé.[106] Un obstacle important à ce projet s'est cependant fait jour, rapidement : les indicateurs sociaux déjà disponibles et étudiés dans la recherche internationale[107] ou en France,[108] tels que la catégorie socio professionnelle (CSP), la classe de revenu ou encore le niveau d'éducation, sont insuffisants à décrire la position socioéconomique des personnes. Cette insuffisance est encore plus prononcée quand on se risque à étudier la vulnérabilité associée à leurs conditions de vie[109] et leurs conditions dites « néo-matérielles »[110] (facteurs associés aux modes de vie des sociétés post-industrielles, urbanisées,[111] comme le régime alimentaire spécifique, les loisirs et vacances, l'accès à l'information).

Il existe donc à la fois un champ d'investigation pertinent, important, celui de l'épidémiologie sociale, et un vide qualitatif et quantitatif à expliquer les inégalités sociales, les données classiques s'avérant d'efficacité restreinte. De nouveaux déterminants sont ainsi à envisager,[6]

à l'instar de la qualité de l'intégration des personnes, des ruptures sociales[4] – support, capital social,[108,112] ruptures dans l'enfance ou étant adulte[113] – leurs caractéristiques psychosociales,[114] ou même les représentations, expériences tant personnelles que familiales liées à la santé, la maladie, aux soins.[115]

Néanmoins, et en France en particulier, les études en population générale qui s'intéresseraient à ce type de facteurs sont longtemps demeurées rares ; on a le plus souvent eu recours à des enquêtes transversales de personnes en situation dite précaire,[116] parfois à des cohortes d'actifs.[117]

Traditionnellement, la plupart des études en épidémiologie sociale se sont intéressées aux déterminants de la santé mesurés au niveau individuel dans la suite d'une longue tradition de l'épidémiologie moderne qui s'est largement construite en opposition avec l'utilisation, scientifiquement dangereuse et régulièrement critiquée, de données agrégées et d'analyses écologiques.[118] Plus récemment, le souci de replacer l'individu dans son contexte s'est fait jour dans la discipline - à la fois parce que les résultats se sont accumulés sur les effets de l'insertion sociale des individus sur leur santé et, plus généralement, parce que cette perspective réductionniste a été, à son tour, largement critiquée en épidémiologie sociale. De fait, l'existence d'effets du contexte sur la santé des individus et leur accès aux soins fait l'objet d'une reconnaissance croissante en santé publique.[119] On citera par exemple, le développement des recherches en santé urbaine dans les pays anglo-saxons (et, particulièrement, des travaux sur les effets de l'environnement construit[120]). La littérature en « épidémiologie contextuelle », en plein essor, s'attache ainsi à mesurer conjointement les effets de caractéristiques individuelles et contextuelles sur la santé par l'emploi de méthodes statistiques appropriées. En France, les cohortes SIRS et RECORD ont successivement été mises en place dans cet objectif.

Dans le même temps, on observe que les inégalités sociales de santé ne se réduisent pas spontanément par l'augmentation quantitative et qualitative de l'offre de soins curatifs, comme en témoigne notamment leur persistance malgré la part croissante des dépenses publiques affectées à la santé.[121] De surcroît, l'objectif d'équité poursuivi par les systèmes de santé européens n'est pas entièrement atteint,[122] y compris dans les pays où une assurance santé systématique (comme, en France, la Couverture maladie universelle) permet théoriquement l'accès de tous aux soins, ou plus exactement une réduction de ses obstacles financiers. La question de l'accès aux soins ne se résume donc pas aux capacités financières et

matérielles des personnes, ni à la disponibilité de l'offre. Les comportements de recours aux soins dépendent aussi des facteurs cités plus haut. Ainsi, les inégalités de santé et de recours aux soins ne seraient pas tant liées par une relation directe de cause à effet que par des déterminants en partie communs, dessinant un gradient social semblable qu'il s'agit d'étudier.[103,123]

Dans une perspective de maîtrise des dépenses et un contexte de crise économique depuis 2008, tous les organismes publics, tous les systèmes administrés voient, dans des proportions variables, leurs ressources allouées diminuer. C'est également le cas du système de soins. Si l'on désire mener ce genre de politique d'une manière adaptée aux réalités, il faut encore pouvoir documenter cette réalité et ajuster au mieux les attentes, les besoins des personnes en termes de soins avec l'offre disponible. Si un certain nombre de déterminants de recours ou de non-recours aux soins sont connus, il est nécessaire d'approcher le problème globalement, intégrativement. Ainsi, quelques études se sont intéressées au mode de recours au système de soins d'un point de vue systémique ou du point de vue décisionnel, autant qu'aux modes de recours par des sous-types de population, comme les patients atteints de telle ou telle maladie grave ou chronique (cancer, diabète, VIH/Sida), les migrants, les pauvres et les démunis. Parallèlement, les déterminants de recours à certaines composantes du système ont été étudiés : en santé mentale, en médecine d'urgence, en soins primaires, en soins dentaires ou en soins spécialisés.

Il a été proposé que les systèmes de soins en eux-mêmes ne peuvent pas être correctement analysés selon une approche classique et réductionniste, et devraient être vus comme des systèmes complexes – lesquels appellent des techniques d'analyse dédiées. Dans cette optique, il peut être pertinent de rechercher l'existence de profils de recours aux soins, de groupes homogènes d'utilisateurs vis-à-vis de l'ensemble de l'offre disponible.

Le modèle d'Andersen et ses variantes comme cadre d'analyse

A la fin des années 60, Andersen a proposé un premier modèle, qui sera suivi de plusieurs autres, visant à fournir un cadre d'analyse pour l'utilisation des systèmes de soins, basé sur les comportements.[124,125] L'auteur a identifié 3 grands groupes de facteurs structurant son modèle comportemental : les *predisposing factors*, les *enabling factors* et les *needs factors*.

La caractérisation d'un système de soins selon ces trois groupes de facteurs, et selon les différentes versions du modèle comportemental ainsi que la manière dont ces facteurs sont

représentés, doivent éclairer quant à l'équité du système de soins en termes d'accessibilité. Selon les facteurs concernés, l'iniquité peut être plus ou moins réduite, de la même manière qu'il existe des facteurs de risque dits modifiables et non modifiables. Parmi les facteurs modifiables, certains demandent des interventions ou des efforts plus ou moins réalistes et bénéfiques au vu de l'effet escompté sur la réduction des inégalités de recours aux soins.
Les « predisposing factors » sont représentés par des caractéristiques telles que l'âge, le genre, le niveau d'éducation, l'activité professionnelle, les relations sociales, les attitudes envers le système de soins et les professionnels de santé et leur connaissance de ceux-ci.
Les « enabling factors » sont ceux qui sont à même de permettre le recours aux soins, de « passer à l'acte », si les conditions sont réunies, telles qu'un contexte favorable au recours (*predisposing factors*), et l'identification de besoins perçus ou réels (*needs factors*). Parmi ces facteurs, on mentionnera les rôles prépondérants des revenus et de la couverture maladie. Ainsi, des revenus suffisants et l'existence d'une couverture maladie assurant la prise en charge totale ou quasi-totale des frais de santé seront autant de facilité de recours effectif.
Les « needs factors » représentent l'opportunité pour un recours aux soins : peut estimer devoir recourir aux soins celui ou celle qui présentent un état de santé nécessitant intervention. Ainsi, on comptera parmi les besoins l'état de santé perçu ou encore les incapacités fonctionnelles.

Selon les facteurs qui seront les déterminants de recours aux soins les plus importants pour un système donné, ce système sera considéré plus ou moins équitable, en termes d'accès. Ainsi, si les déterminants principaux de recours sont représentés par des *needs factors*, par des besoins perçus ou réels, l'accès au système paraît équitable on accède au système en raison de besoins, et l'utilisation correspond à une adéquation entre besoins exprimés et besoins satisfaits. A l'opposé, si les principaux déterminants sont constitués de facteurs sociaux, des croyances ou d'*enabling factors* (revenus, couverture maladie), l'accès au système a peu de chances d'être équitable.

Inégalités ou différences de recours aux soins ?

Les recours aux soins, les circuits d'accès et l'orientation des malades ainsi que leur prise en charge ne sont pas identiques pour tous.[42,126] Ils peuvent différer - voire révéler d'authentiques discriminations - selon, par exemple, l'origine migratoire des personnes, leur situation sociale,

leur couverture maladie, ou certaines de leurs pratiques ou habitudes de vie. La « distance sociale » à l'égard des institutions de soins, ou l'image qu'en ont les personnes, induit de plus des phénomènes de dissuasion symbolique qui éloignent certaines personnes de certaines structures : cette dissuasion s'opérant parmi les plus pauvres mais aussi parmi les personnes en haut de l'échelle sociale (qui fréquentent peu, par exemple, les centres de PMI). En tant que rapport à une institution centrale de notre société (la médecine, et ses différentes structures), les recours aux soins mobilisent un vaste ensemble de dimensions qui renvoient à des normes, des représentations, des processus identitaires et relationnels, et au rapport à la société dans son ensemble.[127] Ceci est d'autant plus marqué que les frontières se troublent, dans notre société, entre le médical et le social, entre la santé et l'exigence normative de performance.[128]

Récemment, enfin, une étude examinant le « paradoxe » de la persistance des inégalités sociales de santé a recensé les différentes explications plausibles à cet état de fait.[129] Des 9 explications possibles, l'auteur en retient 3 qu'il considère plus vraisemblables à l'heure actuelle : 1) en dépit des mesures prises, les inégalités d'accès aux ressources matérielles et immatérielles par les plus pauvres demeurent, 2) l'association entre mobilité sociale et santé, dans un sens comme dans l'autre (à savoir que la mauvaise santé mène à des conditions socioéconomiques défavorables, et inversement, que des conditions socioéconomiques défavorables mènent à un mauvais état de santé) s'est accrue et 3) les comportements de santé et de consommation sont devenus les premiers déterminants de santé en importance, amenant les plus aisés à plus en tirer profit et creusant l'écart avec les plus défavorisés.

Quelques typologies déjà connues en santé

Nous présentons à la suite plusieurs typologies qui ont pu être dégagées jusqu'à présent dans le domaine de la santé. Ces typologies peuvent s'ordonner selon le fait que l'on mette l'accent sur le type de population ou des caractéristiques de ces populations, sur les pathologies, ou sur des aspects du recours aux soins. Assez étonnamment, les approches par typologies semblent se rencontrer plus fréquemment dans la littérature française qu'anglo-saxonne.

Typologies selon le type de population ou leurs caractéristiques.

- **L'analyse et les typologies différenciées par genre**

Les différences de genre concernant les comportements de santé, consommations de soins ou simplement en termes de pathologies sont connues, et le genre figure de façon quasi systématique dans toute étude, au moins initialement, quitte à le retirer des analyses si son impact est négligeable.

Dans l'étude de la DREES,[130] les différences de genre sont ici examinées au travers des résultats de l'enquête Handicap-santé 2008. Il y est analysé les différences d'état de santé perçue, selon que l'on est homme ou femme, ou bien de recours au moins une fois au spécialiste, au généraliste, dans les 12 derniers mois. On en retient par exemple la propension plus importante des femmes à consulter le généraliste aussi bien que le spécialiste (gynécologue inclus ou non) par rapport aux hommes, et ce de façon significative sur au moins toute la première partie de la vie.

- **L'analyse concernant les populations dites vulnérables ou à risque : exemple des migrants**

Dans un ensemble d'études, les auteurs s'intéressent aux populations migrantes ; notamment, ils lient ou rapprochent inégalité de santé, et inégalités de recours aux soins.[131]De leur côté, Berchet et Jusot notaient :[132] «Les inégalités de santé liées à la migration sont aussi confirmées par des inégalités en matière de recours aux soins. Les résultats des diverses études françaises sont relativement convergents et soutiennent l'idée d'un moindre recours au généraliste ou au spécialiste de la population immigrée. À besoin de soins équivalents, les immigrés de première génération recourent moins souvent au généraliste et au spécialiste, alors que les immigrés de seconde génération ne se distinguent pas des Français nés de parents français. Après ajustement des conditions socio-économiques, seules les disparités d'accès au spécialiste persistent. Les disparités de recours aux soins sont, par ailleurs, cohérentes avec le plus important taux de renoncement aux soins mis en évidence dans certains travaux. Ainsi, d'après une enquête réalisée auprès des bénéficiaires de l'Aide médicale de l'État (AME) en 2007, 25% des immigrés déclarent renoncer aux soins. Les difficultés financières figurent parmi les premières raisons de ce renoncement. L'exploitation de l'enquête «Trajectoires et origines» indique que les individus originaires d'Afrique

sahélienne renoncent plus fréquemment aux soins pour des raisons financières que l'ensemble des autres groupes d'immigrés.»

Nous pouvons également évoquer le recours à la prévention médicalisée, qui diffère selon que l'on est immigré, nés de parents immigrés ou nationaux. Une étude menée en France a pu montrer l'existence d'un gradient prononcé en fonction de l'origine migratoire des femmes, quant au retard ou à l'absence de dépistage des cancers du sein et du col de l'utérus. Ce gradient persistait par ailleurs lorsque le statut socio-économique des femmes était pris en compte.[133]

Le recours aux soins s'analysent donc ici selon une double grille, celle évidemment des catégories de recours aux soins (généraliste, spécialiste, prévention médicalisée), et celle des origines ; les immigrés sont opposés aux non immigrés, et les différences concernant d'autres paramètres, comme le revenu, sont examinées et comparées.

- **L'analyse selon la couverture sociale : les bénéficiaires de la CMU**

Ainsi que nous l'avons déjà précisé, la condition d'universalité d'accès (financier) aux soins est une réponse fréquemment proposée aux inégalités de recours. Puisque l'universalité, en France, s'est concrétisée par l'instauration de la CMU et de la CMUc, reflétant de fait en même temps qu'elle crée une hétérogénéité supplémentaire dans les modalités d'accès, il est légitime de s'interroger sur l'impact de ces types de couverture.

En 2004, Boisguérin note ainsi :[134]

« Une autre manière d'appréhender l'impact de la CMU sur le recours aux soins est de repérer les personnes qui, ayant renoncé à des soins avant de se voir ouvrir le bénéfice de la CMU, ont été amenés à engager des soins depuis. La proportion des ménages dans cette situation demeure à la fois stable et forte entre 2000 et 2003 : comme en 2000, 71 % des ménages de nouveaux affiliés ont ainsi entamé des soins depuis qu'ils disposent de la CMU. Cette proportion a augmenté entre 2000 et 2003 pour les prothèses dentaires (de 35 % à 49 %) et pour les soins optiques, (passant de 47 % à 61 %). En 2003, ce sont les ménages dont la personne de référence est inactive qui déclarent le plus fréquemment avoir commencé des soins depuis que le bénéfice de la CMU leur a été accordé. Au niveau individuel, ce sont, de la même façon, plus de 70 % des nouveaux bénéficiaires qui avaient renoncé à au moins un soin avant la CMU qui en ont entamé depuis son obtention. Pour les prothèses dentaires et les soins optiques la proportion de personnes ayant commencé des soins a aussi progressé,

respectivement de 38 à 47 % et de 48 à 62 %. Comme en 2000, ce « rattrapage » engagé en matière de soins a surtout profité aux enfants et plus généralement aux jeunes de moins de 20 ans dont 81 % sont concernés. »
Ainsi, en suivant les auteurs, on constate des effets différentiels de l'attribution de la CMU, notamment selon l'âge, ou le type de soins nouvellement engagé. Evidemment, la seule attribution de la CMU n'explique pas intégralement le comportement, mais davantage la modification a priori transitoire, de comportement.

Typologies selon le type de pathologie.

- **L'exemple des patients dépressifs**

L'étude DREES concernant « la classification des dépressifs selon leur type de recours aux soins » se base sur les données de l'enquête Santé mentale en population générale, menée entre 1999 et 2003.[135] Cette enquête recensait 5 grandes catégories de recours aux soins concernant les personnes dépressives : la consultation d'un professionnel de santé, les séjours dans les structures de soins, la consommation de médicaments, le recours aux médecines douces, et les traitements dits traditionnels.
Les auteurs distinguent 3 premiers niveaux d'interprétation des résultats, avant de parler de la typologie elle-même, constituée de 8 groupes de patients :
« Les comportements de recours aux soins des personnes dépressives sont contrastés. Le premier type de distinction oppose les personnes dépressives qui se soignent et celles qui ne se soignent pas. Le second oppose celles qui recourent à des soins dans le domaine de la santé mentale et celles qui recourent à d'autres types de soins. Le troisième oppose celles qui recourent à des soins conventionnels et celles qui recourent aux médecines douces ou à des traitements traditionnels. »

Par suite, la typologie comporte 8 types différents de patients dépressifs, vis-à-vis de leur recours aux soins :

- Les personnes dépressives ne requérant aucun soin (28%)
- Les personnes qui consultent un professionnel de santé (le généraliste, essentiellement), et consommant des médicaments « recommandés » (30%)

- Les personnes qui consultent un professionnel de santé (le généraliste, essentiellement), et consommant des médicaments de toute sorte, et peu de psychotropes (5%)
- Les personnes utilisant les plantes (6%)
- Les personnes utilisant l'homéopathie et autres médecines douces en dehors des plantes (12%)
- Les personnes requérant à des traitements dits traditionnels (guérisseurs, marabouts… 5%)
- Les personnes ayant été hospitalisées en structures spécialisées (9%)
- Les personnes ayant été hospitalisées en hôpital général (6%)

Typologies selon le type de recours.

- **Le recours aux soins urgents ou non programmés en médecine de ville**

Une question récurrente dans l'organisation des soins, et le type de recours aux soins selon certaines caractéristiques socio-économiques et démographiques, est de savoir si la structure actuellement proposée pour gérer les soins urgents (objectivement urgents ou ressentis comme tels, mais également non programmés) est efficace et adaptée aux besoins.

En particulier, existe-t-il un phénomène de substitution entre recours aux urgences hospitalières et recours au médecin généraliste de ville, pour des soins non programmés ? Une étude de la DREES proposait d'examiner les recours au médecin généraliste de ville pour les soins urgents ou non programmés, et d'en dégager une typologie, avant toute action ou remaniement envisageable de l'activité.[136,137]

L'étude propose une typologie en 7 types :

- Les recours des enfants et adolescents pour épisode infectieux aigus (43%)
- Episode aigu non infectieux touchant des patients plus âgés (24%)
- Recours relatifs à des maladies chroniques stables (6%)
- Manifestations allergiques et lésions dermatologiques nécessitant des soins (4%)
- Problèmes traumatiques (10%)
- Troubles psychiques des adultes (8%)
- Urgences somatiques critiques (5%)

Cette typologie présente 3 intérêts distincts : d'une part, elle confirme par l'analyse ce que l'on présentait, dans les grandes lignes, comme classification de recours urgents en médecine de ville ; d'autre part, elle identifie des types qualitatifs de recours relativement décorrélés les uns des autres ; enfin, elle quantifie la fréquence de chaque type de recours.

- **Le recours spécialisé en santé mentale**

Une étude basée sur l'Enquête décennale santé 2002-2003 de l'INSEE[138] a mis l'accent sur le type de recours aux soins spécialisés en santé mentale, notamment via le type de professionnel requis : le psychiatre, le psychologue ou le psychanalyste, et le recours à l'hospitalisation. Cette typologie a priori s'est révélée pertinente, puisque dans le cadre de l'enquête, 9 personnes sur 10 ont eu recours à un seul type de spécialiste, suggérant ainsi un recours très différencié. Il faut cependant noter la faible durée (8 semaines) prise en compte dans les questions de recours, qui peut laisser dans l'ombre la possible co-consultation ou réorientation vers un autre type de professionnel au-delà de ces 8 semaines.
On en retient notamment les grands résultats suivants :

- Globalement : les consultants sont plus souvent des femmes, ne vivant pas en couple, ayant conscience d'un trouble psychique et de niveau scolaire élevé ;
- Les personnes consultant des psychiatres sont le plus souvent des adultes en forte détresse psychique, au parcours professionnel très perturbé ;
- Les personnes consultant des psychologues sont souvent soit des jeunes de moins de 20 ans, soit des adultes à la vie perturbée aussi bien dans le domaine privé que professionnel, avec un bon niveau d'études ;
- Les personnes consultant les psychanalystes semblaient s'être trompé de professionnels, car ils espéraient un ou deux rendez-vous maximum, ce qui est incompatible avec les cures psychanalytiques ou les psychothérapies d'inspiration psychanalytique ;
- Les personnes hospitalisées en psychiatrie cumulaient de lourdes difficultés et un recours important aux soins non psychiatriques.

Notons que si le choix a priori de catégoriser les personnes ayant recours à des soins spécialisés en santé mentale par le type de professionnel consulté s'est révélé pertinent, rien ne peut permettre de conclure que ce soit la « meilleure » typologie de recours aux soins

spécialisés. Tout au plus peut-on conclure qu'il existe effectivement des différences significatives dans les caractéristiques des personnes consultant plutôt un type qu'un autre.

- **Le recours au spécialiste en médecine de ville en 2007**

Citons pour conclure l'existence d'une enquête de la DREES, concernant le recours au spécialiste en médecine de ville.[137] Là où on s'était intéressé aux recours au généraliste de ville pour soins urgents ou non programmés, il est ici question de mieux connaître, côté spécialiste et côté patient, les motifs, conditions et déterminants de consultation du spécialiste de ville.

Les spécialistes concernés sont le cardiologue, le gastro-entérologue, le pédiatre, le gynécologue, le psychiatre, l'ophtalmologue, le dermatologue, l'ORL et le rhumatologue. En particulier, il est noté qu'il existait 5 types de recours aux médecins spécialistes d'accès indirect : les patients n'ayant pas déclaré de médecin référent, ceux consultant pour une affection chronique, les consultations d'urgence, les consultations de dépistage et les recours n'entrant dans aucune de ces catégories. Il était également noté que les jeunes et les actifs en emploi étaient plus fréquemment en position de « hors parcours », donc ne passant pas par un médecin référent.

Aspects du système de soins français

Structuration du système

Dans cette partie et les suivantes, nous nous référerons préférentiellement à l'ouvrage de Durand-Zaleski, Chevreul et al, dans la collection *Health Systems In Transition* (HIT), où le système de soins est présenté de façon synthétique et dans une démarche de comparabilité avec les systèmes de soins d'autres pays.[139] Nous utilisons ici cette seule référence dans un dessein de simplicité et de cohérence ; d'autres choix sensiblement différents dans leur présentation auraient été possibles, sans toutefois bouleverser les fondements du travail.

Le système de soins français fait intervenir à la fois des acteurs publics et des acteurs privés. Les soins primaires sont essentiellement délivrés en secteur ambulatoire, notamment par des professionnels d'exercice libéral. Les soins secondaires se partagent eux entre secteur ambulatoire et secteur hospitalier. Depuis 1990, la place des médecins généralistes a évolué, puisqu'ils ont été placés à l'entrée du parcours de soins, et au centre (en théorie) de la

coordination des soins entre professionnels. Le dispositif est un peu semblable en ce sens à celui existant en Angleterre (« *gate keeping* »). Les établissements hospitaliers se répartissent entre hôpitaux publics, hôpitaux privés à but non lucratif (participant ou non au système public), et hôpitaux privés à but lucratif. Les soins de longue durée pour personnes âgées et pour personnes handicapées sont assurés tant dans le secteur institutionnel que par les soins à domicile.

D'un point de vue macroscopique, structurel le système de soins se retrouve façonné et contraint par 3 grands types d'entrées :

- La santé publique (lois, problèmes, priorités, état de santé des populations…)
- La planification et la régulation
- Le financement

Concernant la santé publique, au sens strictement réglementaire et législatif, le code de santé publique a vu sa taille et les domaines concernés augmenter exponentiellement en quelques décennies, et est appelé à s'enrichir encore dans un futur proche. Nous ne reviendrons pas sur la chronologie des différentes lois de santé publique ou afférentes, mais soulignerons simplement qu'elles poursuivent dans leur esprit, la clarification des responsabilités et droits respectifs de tous les acteurs du système de soins (du patient, devenu usager du système de soins, mais aussi des professionnels de santé) et la réaffirmation du rôle central de l'usager du système au sein du dispositif entier (individuellement ou par les associations d'usagers).
Les buts poursuivis par la santé publique y sont décrits comme la caractérisation du lien sanitaire entre individu et population, l'identification des problèmes de santé, leur priorisation, et doit s'articuler aux politiques de santé. Il est cependant également possible de considérer ce développement réglementaire de la santé publique sous le jour de la généralisation d'un encadrement légal du risque, et partant de sa monétisation.

Offre générale de soins

Nous reprenons ici à dessein, telle quelle, la présentation de l'offre de soins figurant dans Chevreul et al.[139] Nous dressons très brièvement un aperçu du contenu de chacune des catégories.

A - Le parcours de soins

Depuis 1990, il est difficile de parler du système de soins français et de l'offre de soins sans préciser l'introduction de la notion de parcours de soins. Il s'agit d'une version faible du système anglais de « *gate keeping* », où le médecin généraliste référent est le passage obligé – du moins, fortement recommandé – à l'entrée dans le système de soins.

Le médecin référent d'un usager est le médecin de son choix (pas nécessairement un généraliste, par ailleurs), déclaré auprès de l'Assurance maladie, et doit être consulté en priorité pour les soins primaires, avant qu'il n'oriente éventuellement le patient vers des soins spécialisés s'il considère que c'est nécessaire. Ce système repose sur des pénalités financières, et n'est pas totalement verrouillé : si un patient consulte directement un spécialiste, il ne pourra prétendre au remboursement maximal prévu.

Cette restriction ne tient pas pour certaines spécialités, accessibles directement. Les spécialistes d'accès direct sont : les gynécologues, les pédiatres, les ophtalmologues, les stomatologues, les psychiatres entre 16 et 25 ans (en dehors, en réalité, de toute consultation d'un psychiatre en CMP – centre médico-psychologique, où un psychiatre peut recevoir en théorie directement n'importe quel patient).

Dans la suite, nous tiendrons compte de cette distinction, entre spécialistes accessibles directement ou indirectement, via le médecin référent.

B - Les soins ambulatoires

Les soins ambulatoires peuvent se définir par opposition à l'hospitalisation, c'est-à-dire requérant le coucher. Cette définition souffre un contre-exemple, dans le cas de la santé mentale : l'hospitalisation à temps partiel peut ne concerner que la journée, sans qu'il y ait hébergement.

Les soins ambulatoires concernent aussi bien les consultations, actes diagnostiques, thérapeutiques délivrés par les généralistes, les spécialistes, les infirmiers etc. L'essentiel est assuré en cabinet de ville, en activité libérale, même si une petite proportion est assurée par les centres de santé, les dispensaires, les soins ambulatoires hospitaliers (consultations et soins externes) où le personnel peut être salarié ou libéral.

C - Les soins d'hospitalisation

L'hôpital revêt différentes formes : on distingue notamment les hospitalisations pour soins de courte durée, en médecine, chirurgie et obstétrique (MCO), les hospitalisations dites de jour, ou encore l'hospitalisation à domicile (HAD). Les établissements sont soit publics, soit privés, avec la possibilité d'établissements privés participant au service public hospitalier (PSPH), ou d'établissements privés à but lucratif (de plus en plus concentrés entre les mains de quelques opérateurs privés).
Le secteur public représente environ les trois quarts de la capacité hospitalière et du volume des actes effectués avec, cependant, de grandes variations géographiques.[140]

D - Les soins en santé mentale

La structuration des soins en santé mentale, même si elle présente d'évidentes similarités avec le reste des soins dispensés en MCO, reste un cas à part. La critique d'hospitalo-centrisme demeure, la crainte d'une persistance d'enfermement « à vie » en lit d'hôpital est toujours présente, mais des efforts ont été faits afin de réduire la part de l'hospitalisation « fermée » et complète, ainsi que la durée de celle-ci (cf plan psychiatrie et santé mentale 2005-2008 et son évaluation par le HCSP[141]).
Néanmoins, la place, la définition plus ou normative de la pathologie mentale, ses liens avec un autre régime d'enfermement (la prison) notamment via les modes de privations de liberté (par exemple : hospitalisation d'office, à la demande d'un tiers, intervention du juge des libertés et de la détention dans le processus de contrôle des hospitalisations sous contrainte), font de l'organisation de la santé mentale un problème complexe assez différent de celui du somatique « pur ».
La santé mentale se structure autour de nombreuses instances : l'hôpital, évidemment, en unités ouvertes ou fermées, de jour, en temps plein, en temps partiel, en séquentiel ; le secteur et ses centres médico-psychologiques, comme première liaison et alternative à l'hôpital ; d'autres structures, notamment diverses solutions d'hébergements selon l'encadrement offert (appartements thérapeutiques, familles d'accueil thérapeutique, foyer logements, maisons d'accueil spécialisé…) ou encore des structures dédiées pour l'emploi (ESAT – établissement et service d'aide par le travail).

Globalement, les parcours de patients dans les différentes structures de santé mentale restent assez mal connus, et présenteraient des trajectoires relativement indépendantes et différentes des trajectoires MCO, ne serait-ce que par la sous prise en charge des pathologies somatiques chez les patients psychiatriques, ou le manque de coordination généraliste-psychiatre de ville, qui tendent chacun de leur côté à assumer l'intégralité de la prise en charge (selon des travaux non publiés, en cours, sur le PMSI et le RIM-P). Hormis les consultations psychiatriques ou autres professionnels de santé mentale (psychologues…), le recours aux structures de soins en santé mentale ne sera pas pris en compte dans notre étude.

E - Les urgences, ou recours aux soins non programmés

Les urgences recouvrent classiquement 2 domaines dans le système français : les urgences pré-hospitalières et les urgences hospitalières.
Les urgences pré-hospitalières sont assurées par la permanence des soins et la régulation des centres 15. Elles concernent, de facto, toute consultation pour motif urgent, toute consultation non programmée, ou en dehors des heures ouvrables.
Les urgences hospitalières sont assurées par les services d'urgences. Elles sont abordables directement si le patient s'y présente spontanément, ou éventuellement comme orientation post urgences pré-hospitalières.

F - Les soins dentaires

Les soins dentaires sont essentiellement assurés par des professionnels exerçant en libéral (91%), les autres étant des salariés des hôpitaux ou de centres spécialisés en soins dentaires. Les dentistes sont les plus représentatifs de cette activité, auxquels il faut ajouter les stomatologues (médecins spécialistes), qui en règle général prennent en charge les interventions à orientation davantage chirurgicale ou au bloc opératoire (par exemple : extraction de dents de sagesse).
Etant donné leur niveau de remboursement par l'Assurance maladie, les soins dentaires sont l'un des soins les plus inégalitaires en termes d'accessibilité. Les soins d'orthodontie, toute l'activité des prothésistes, sont en effet peu ou mal remboursés par l'Assurance maladie, et

relativement mal complétés par les assurances complémentaires « d'entrée ou de milieu » de gamme.

Une étude de 2006 estimait que 9% de la population avait renoncé à des soins dentaires dans les 12 derniers mois, pour raisons financières (à comparer aux 14% de renoncement tous soins confondus).[142]

G - Médecines complémentaires et alternatives

Les médecines complémentaires (acupuncture, ostéopathes, phytomédecine) et alternatives (homéopathies, rebouteux), ne sont pas soumises à une réelle régulation en France. Les catégories les plus représentatives et les plus intégrées au système de soins officiel sont, par exemple, l'homéopathie, l'acupuncture, l'ostéopathie. Un des problèmes soulevés de façon récurrente par l'absence de régulation ou de surveillance, d'évaluation de ces domaines d'exercices, est la crainte de l'exercice illégal et frauduleux, voire dangereux au-delà de la simple escroquerie.

Il est cependant intéressant de noter que la France représente le marché le plus important pour l'homéopathie, tout en étant le premier producteur mondial dans ce domaine.

La régulation existante concernant l'homéopathie est définie par le code de santé publique, faisant pendant aux médicaments stricto sensu. Il existe par ailleurs deux régimes pour les produits homéopathiques, selon leur degré de dilution (simple déclaration), et leur indication thérapeutique ciblée (nécessité d'AMM pour les nouveaux produits).

L'acupuncture a été incluse dans la nomenclature générale des actes médicaux depuis une vingtaine d'années, et est remboursée sur la base d'une consultation de médecine générale.

L'ostéopathie, quant à elle, n'est pas reconnue encore comme activité médicale, et n'est remboursée que par un certain nombre de complémentaires santé.

Pour des raisons culturelles, de croyances, par effet éventuellement de contexte et de voisinage, pour des raisons d'insatisfaction, le recours à ces soins n'est pas exceptionnel. Il nous a paru important de les prendre en compte.

H - L'offre médicamenteuse

Le médicament, la pharmacopée tient une place importante, dans l'arsenal thérapeutique, mais aussi en valeur et en volume dans les comptes de la santé et la pratique courante, ce qui rend compte d'un usage, mais ne reflète pas nécessairement ou proportionnellement le service médical rendu attendu. Dans notre étude, nous ne disposons pas à proprement parler de la consommation médicale et médicamenteuse des personnes. En outre, en dehors des cas de l'automédication – certes non négligeable – et de la procuration de médicaments reconnus ou étiquetés comme tel par des voies extra légales, l'accès aux médicaments est contrôlé par le mécanisme de l'ordonnance, ce qui contraint assez fortement ce type de recours.

I - Les soins de suite et de réadaptation (SSR)

Ces structures sont des entités dédiées aux suites d'intervention, essentiellement. Leur accès est donc fortement conditionné à un recours initial pour un problème de santé, au moins, à des soins plus ou moins choisis et consentis. Les SSR sont par conséquent des modes de soins d'accès très particuliers, et ils ne seront pas pris en compte dans notre étude.

J - Les soins de longue durée

Les soins longue durée se partage entre les soins délivrés en institution, avec une option d'hébergement plus ou moins médicalisé, et les soins à domicile. Parmi les différentes solutions institutionnelles, on trouve aussi bien les foyers hébergements, que les EHPA et EHPAD, que les unités de soins longue durée. Globalement, la majorité des solutions concernent les personnes âgées et / ou dépendantes, en raison d'un handicap.
De la même manière que les SSR, les soins longues durées sont suffisamment particuliers, et d'accès fléché pour ne pas figurer explicitement dans les possibilités retenus dans l'étude.

K - Les soins palliatifs

Les soins palliatifs ont été reconnus, intégrés au système de soins officiel à partir de 1986. Ces soins procurés pour l'accompagnement des fins de vie ne sont pas explicitement inclus dans notre étude, même s'ils ne sont pas formellement exclus, notamment via la HAD.

L - Les soins adressés à des populations spécifiques

Par « soins adressés à des populations spécifiques », on regroupe essentiellement 3 grandes catégories, à savoir :

- Population carcérale.
- La santé des armées.
- Les populations en situation illégale, réfugiés et demandeurs d'asile.

L'organisation des soins pour les personnes incarcérées reste compliquée pour diverses raisons. En théorie, chaque établissement pénitentiaire est lié contractuellement à un établissement de santé public de référence, pour le secteur MCO, et pour la psychiatrie. Il y a eu transfert de responsabilités entre l'administration pénitentiaire et la santé depuis 1994, ce qui a permis d'améliorer la situation. L'accès aux soins, notamment relevant de la psychiatrie, reste perfectible

La santé des armées est appelée à muter, et mute déjà depuis quelques années, avec une ouverture totale sur le service public et non plus réservée aux seuls personnels militaires et à leurs familles.

Les personnes en situation illégale, les réfugiés et les demandeurs d'asiles sont en pratique reçus par les hôpitaux publics, en particulier par les services d'urgences, les PASS (Permanences d'accès aux soins de santé) ou les ONG.[143]

Mentionnons l'existence de l'Aide médicale d'Etat (AME), ainsi que le droit au séjour pour soins. Les « sans papier » qui résident depuis au moins 3 mois sur le territoire français et perçoivent moins 661 euro par mois peuvent bénéficier de l'AME, droits ouverts pour une année. Les bénéficiaires de l'AME n'ont pas d'avance de frais et sont pris en charge à 100%.

Enfin, les personnes en situation illégales ne peuvent être expulsées du territoire français si leur expulsion entraîne d'une part de graves conséquences pour leur santé si leur prise en charge devait s'arrêter, d'autre part le risque de ne pas pouvoir recevoir de soins adaptés dans leur pays d'origine.

Inégalités face au système judiciaire : cas des adolescents migrants

S'il existe des frontières géographiques et des états, alors il existe des flux migratoires, dans un sens comme dans l'autre. S'il existe un droit européen, fédéral ou transnational comme le droit européen, les droits nationaux s'appliquent toujours sur les territoires respectifs. Ainsi, la nationalité ouvre des droits spécifiques à ceux qui la détiennent, et en contraste, ceux qui ne la détiennent pas répondent d'autres droits, d'autres obligations. Le critère d'âge en particulier, et plus précisément de majorité légale, qui peut différer d'une région à l'autre du monde, est décisif quant aux droits ouverts à la personne. Il existe par ailleurs une convention, la Convention internationale des droits de l'enfant (CIDE), adoptée voilà 25 ans (1990 pour la France) par l'Assemblée générale des Nations unies, qui définit les éléments d'une protection de l'enfance.[144] Entre autres droits, la Convention précise que toute décision concernant un enfant doit tenir pleinement compte de l'intérêt supérieur de celui-ci. De même, l'Etat doit assurer à l'enfant la protection et les soins nécessaires au cas où ses parents ou les autres personnes responsables de lui en sont incapables (article 3). L'un des 4 principes fondamentaux de la Convention est le droit à la « non-discrimination ».

En France, en 2008, on recensait 3,2 millions d'immigrés étrangers, tous âges confondus.[145] Les adolescents migrants sont généralement désignés sous le terme de mineurs isolés étrangers (MIE). Leur nombre annuel est estimé entre 4000 et 8000,[146] et le nombre de mineurs isolés étrangers pris en charge par les autorités, sans présumer de l'issue de cette prise en charge, était de 3734 en 2013.[147] Leurs origines géographiques principales étaient la Guinée, le Mali et le Congo (15% chacun, soit près de 45% pour ces 3 origines), suivies du Bangladesh, de l'Albanie et du Pakistan (7%-7%-5%), puis le Maroc et l'Algérie (4,5%). Les « autres pays » représentaient presque un quart des cas (24%).

En France, un mineur migrant qui ne détient pas la nationalité française restera en général, dans les faits, sur le sol français et bénéficiera de l'Aide sociale à l'enfance. Un majeur, quant à lui, sera reconduit à la frontière, renvoyé à son pays d'origine. Pour les personnes détenant la nationalité française, bien entendu, les droits diffèrent également du moment que l'on est mineur ou majeur, que l'on est légalement dépendant d'une autorité parentale ou non. Si pour ces derniers, les français, la question de la justification de leur minorité ou de leur majorité ne pose pas problème, il n'en va pas ainsi pour les migrants arrivant sur le territoire français – qu'ils aient en leur possession ou non une pièce qui justifierait de leur âge ou d'une date de naissance, à partir du moment où l'on considère leur entrée sur le territoire comme illégale ou irrégulière. Une question préliminaire à tout droit qui leur serait ouvert est alors celle de leur

âge, de leur minorité ou de leur majorité. Ainsi, face à un même système judiciaire, au sein d'un même état de droit, pour des adolescents qui présenteraient comme différence leur nationalité, le traitement diffère. Pour résoudre ce qui pourrait être une question d'éthique et de société, la science est convoquée, et doit se prononcer sur l'âge de ces adolescents. Sous les garanties scientifiques, on peut être en droit d'espérer une approche fiable, équitable et partagée au moins au sein de pays partageant une partie de leurs droits. Actuellement, cet appel à la science semble davantage relever d'un artifice ou d'une démission de responsabilités de l'Etat ou de la société : les arguments scientifiques évoqués dans les tribunaux ne sont en aucun cas une source infaillible de « vérité », voire d'objectivité alors qu'ils sont à même de modifier les décisions des magistrats.[148–151] Enfin, pour ce qui est des adolescents migrants, les pratiques présentées comme scientifiques pour se prononcer sur leur minorité ou leur majorité, sont régulièrement critiquées, mais sans succès jusqu'à présent. Un article du Monde de janvier 2015 résumait assez bien la situation :[152]

« Ainsi, dès juin 2005, le Comité consultatif national d'éthique (CCNE) soulignait *« l'inadaptation de ces méthodes »*, comme l'avait fait auparavant la Défenseure des enfants. Tour à tour, l'Académie nationale de médecine, le Comité des droits de l'enfant des Nations unies, l'ancien commissaire aux droits de l'homme du Conseil de l'Europe, le Haut Conseil de la santé publique, le Défenseur des droits, ont émis sur ce point les plus expresses réserves. Récemment, la Commission nationale consultative des droits de l'homme (CNCDH), dans un avis du 24 juin 2014 préconisait de *« mettre fin aux pratiques actuelles d'évaluation de l'âge »*. » En mai 2015, le même quotidien titrait néanmoins :[153] « Immigration : les députés maintiennent les tests osseux. »

Des populations sélectionnées, les adolescents migrants

Les populations concernées par les demandes d'estimation d'âge sont essentiellement les populations de migrants, que l'on qualifiera d'adolescents, puisque la majorité des cas jugés douteux quant à l'âge déclaré ou non justifié, impliquent des personnes à la charnière entre puberté et âge adulte. Le terme adolescent est suffisamment vague pour ne pas réduire les personnes désignées à des personnes mineures légalement.

La question des populations migrantes est vaste et dépasse le seul cadre de nos travaux. Nous en donnons en préambule une définition et un modèle, tous deux repris des travaux d'Anne Jolivet, portant sur la santé des migrants.[154]

Ainsi, la migration se définit comme « le déplacement d'une personne ou d'un groupe de personnes, soit entre pays (migrations internationales), soit dans un pays entre deux lieux situés sur son territoire (migrations internes). La notion de migration englobe tous les types de mouvements de population impliquant un changement du lieu de résidence habituelle, quelles que soient leur cause, leur composition, leur durée. »[155]

Par ailleurs, Jolivet souligne l'absence de définition consensuelle au niveau international. Il est néanmoins possible de se reporter à la définition que propose l'Organisation internationale des migrations (OIM) : le terme « migrant » s'applique « aux personnes se déplaçant vers un autre pays ou une autre région aux fins d'améliorer leurs conditions matérielles et sociales, leurs perspectives d'avenir ou celles de leur famille.[155] Le terme de migrant englobe de nombreuses situations, que l'on peut classer selon leur principale raison (migrants économiques, étudiants, réfugiés politiques, migrants environnementaux…), leur durée présumée de résidence (temporaire, permanente, intermittente), les circonstances de la migration (régulière ou « clandestine ») et selon la situation en regard du séjour à un moment donné, dans un territoire donné (migrant en situation régulière ou « légale » ou à l'inverse migrant en situation irrégulière ou « illégale » ou « sans papier »). »

La notion de migrant, quelle qu'en soit la définition stabilisée, sous-entend le mouvement, le déplacement : la migration. Si les représentations communes de la migration sont celles d'un déplacement d'un point de départ A à un point d'arrivée B, plus ou moins rapide ou direct par rapport au trajet « à vol d'oiseau », la réalité est plus complexe, spatialement et temporellement.[156–158] Un modèle de migration en 5 étapes a ainsi pu être proposé. Nous ne nous y attarderons pas plus longtemps que ce qu'il est nécessaire pour présenter l'une de ces étapes : l'étape d'interception.

Selon Jolivet toujours, cette étape ne concernerait qu'une minorité des migrants, tels que les demandeurs d'asile, les réfugiés, les personnes dites déplacées, les personnes victimes de traite humaine ou encore les immigrés en situation irrégulière (présents sur un territoire sans en avoir les droits nécessaires). Ainsi : « Depuis le début des années 1990, tous les États membres de l'Union européenne (UE) ont développé des dispositifs législatifs, administratifs et politiques destinés à accueillir, trier, contrôler et renvoyer les étrangers, qui se matérialisent notamment par l'installation de centres ou de camps dédiés.[159] Le nombre de ces lieux de regroupement d'étrangers ne cesse d'augmenter, ainsi que les durées de détention des migrants. En 2012, la capacité totale connue – soit pour les deux tiers des camps au sein de l'UE – est d'environ 37000 places.[159] Ces pratiques contemporaines répressives d'internement

et de logement contraints, ont souvent des effets délétères sur la santé mentale et physique des migrants.[160,161] Il existe une claire association entre la durée de détention et la sévérité des troubles mentaux des personnes concernées. Ces camps, souvent difficilement accessibles, sont également, parfois, le lieu de dérives, d'abus et de violation des droits humains.[159,162] »

C'est dans ce contexte que la justice, devant un doute ou les réclamations de la personne concernée, peut demander à ce que soit estimé l'âge de cette personne – en réalité, de renseigner sur la minorité ou la majorité légale de la personne, au regard de sa définition dans les textes de loi. Reste à savoir comment procéder pour répondre à cette question judiciaire.

La science convoquée comme arbitre judiciaire

En France, la médecine légale a vu son organisation clarifiée, remaniée et stabilisée par la réforme de 2011. Elle a eu comme effet notable de créer un maillage et une typologie des centres médico-légaux sur le territoire, en partant du préexistant, hautement hétérogène en pratiques et en répartition géographique. Une première évaluation de cette réforme a été menée en 2013 par l'IGAS, et rendue publique que très tardivement, fin 2014. Ses conclusions plaident en faveur du maintien de la médecine légale dans ses prérogatives, moyennant quelques ajustements nécessaires et justifiés.
En particulier, un observatoire, l'observatoire national de médecine légale (oNML), a pour fonction d'enregistrer l'activité médico-légale sur l'ensemble du territoire. Ainsi, en 2012, environ 1300 demandes d'estimation d'âge ont été formulées aux légistes (ce qui n'exclut pas formellement que des demandes aient été faites à des non légistes : pédiatres, radiologues par exemple).[9] Un peu plus d'un quart étaient réalisés dans le service de médecine légale de Bondy (93), ce qui s'explique probablement par la proximité de l'aéroport de Roissy (Charles de Gaule).[9] La tendance actuelle est à la diminution de ces demandes, qui persistent néanmoins. En 2014, le même service a réalisé 120 estimations, soit 3 fois moins.

La justice, estimant qu'elle ne dispose pas des capacités propres à estimer de manière fiable l'âge des personnes concernées, se retourne vers des auxiliaires, des avis techniques, comme elle peut le faire pour n'importe quel autre domaine. Le légiste est tout désigné pour s'acquitter de cette tâche. Plus généralement, il s'agit de faire appel à la science pour qu'elle fournisse les garanties d'une conclusion fiable, sinon étayée.

Le droit et la médecine s'accorde difficilement, et sont parfois en porte à faux dès la position du problème, l'un posant des questions en un langage qui lui est propre et selon des attentes, un fonctionnement spécifiques, l'autre y répondant en son langage, et selon l'état de ses connaissances. Ainsi, pour les morts violentes ou suspectes, quand la police judiciaire hésite entre suicide et homicide, l'attente de celle-ci vis-à-vis du légiste est double : il s'agit de trancher entre les deux modes de décès, et par ailleurs d'en dater l'occurrence. En effet, les procédures à déclencher, prévues par le code de procédure pénale, sont sensiblement différentes qu'il s'agisse d'une enquête en flagrance ou non. Or, le déclenchement de l'enquête en flagrance doit se baser sur la donnée des deux éléments de l'infraction (l'homicide), et du délai de commission de cette infraction. Dans la vaste majorité des cas, aucun légiste ne pourra se prononcer fermement sur ces deux éléments. Dans le cas de l'estimation de l'âge chez les adolescents migrants, la question est celle de la minorité légale – notion purement juridique, et appelée à varier dans le temps et les lieux. Cette notion étant juridique, il est a priori difficile de lui imaginer un pendant clair et équivalent en physiologie humaine. Néanmoins, la question peut avoir son intérêt, et il n'est pas illégitime de l'explorer scientifiquement. La science est donc appelée à déterminer, à des fins judiciaires, si une personne est mineure ou majeure. Comment peut-elle y répondre, et comment y répond-elle ?[163,164] La convocation de la science pour statuer sur ce problème n'est pas sans conséquence ou implication éthique.[164–166]

Pratiques européennes et problématique générale

Les pays européens semblent particulièrement préoccupés par cette thématique de l'estimation de l'âge chez les migrants.[167,168] Nous nous appuyons ici sur les résultats préliminaires d'une étude présentée par un auteur Autrichien, Ernst Rudolf, lors du 23e congrès international de médecine légale (IALM), tenu à Dubaï du 19 au 21 janvier 2015.[169] A titre d'exemple, une session entière était dédiée au thème de l'estimation de l'âge chez les vivants. Cette étude portait sur les pratiques et cadres de pratiques au sein des pays de l'union européenne.

Selon Rudolf, il n'existe que peu de pays ayant des recommandations de pratique au niveau national. En particulier, la France, l'Italie et l'Allemagne présentent des pratiques très hétérogènes sur l'ensemble de leur territoire respectif. L'Allemagne dispose cependant d'un groupe d'étude qui produit des recommandations et certifie des praticiens suite à une

formation qu'il dispense : le *Study Group on Forensic Age Diagnostics* (AGFAD).[170–172] Les recommandations produites sont de deux ordres : d'une part, les recommandations concernant la conduite d'études sur les méthodes d'estimation d'âge, d'autre part les recommandations concernant la conduite d'un examen dont la conclusion doit être une estimation de l'âge.

Pour les recommandations portant sur ce qu'est une « bonne » étude sur l'estimation d'âge, il est par exemple suggéré de séparer les résultats selon le sexe masculin ou féminin, et de représenter les résultats sous forme de moyennes et d'écarts-types.

Parmi les recommandations concernant les examens, il est préconisé de recourir aux investigations les moins ionisantes possibles. Pour les modalités prescrites, il est proposé de conduire un examen physique, un examen dentaire et un examen radiologique du poignet de la main non dominante. En cas de besoin (devant une fusion complète des épiphyses distales du radius et de l'ulna), il sera également pratiqué un scanner des clavicules.

L'examen physique s'attachera à mesurer et peser la personne, à en déterminer le stade pubertaire et éventuellement des signes de pathologies pouvant interférer avec une croissance « normale » et harmonieuse. L'examen dentaire a pour but de déterminer la présence des deuxièmes et troisièmes molaires (dents de sagesse).[173] Il est complété par un panoramique dentaire. Différentes étapes de minéralisation ont été décrites (stades de Demirjian), corrélées à l'âge.[174–176] La radiographie du poignet s'intéresse à la fusion des cartilages, et se rapporte aux stades décrits par exemple dans l'atlas radiologique de Greulich et Pyle.[177–179] Enfin, il a également été décrit plusieurs stades de minéralisation des clavicules, également corrélés à l'âge.[180–182]

A l'issue de cet examen, une conclusion doit être donnée quant à l'âge estimé de la personne. Les recommandations demeurent floues quant à ce dernier aspect. Lors du même congrès et de la même séance précédemment évoquée, modérée par le Pr Andreas Schmeling, secrétaire de l'AGFAD, celui-ci a pu répondre à certaines interrogations de l'assistance. En particulier, il a été demandé comment répondre à la question de l'estimation de l'âge, à partir de plusieurs modalités, ainsi qu'il est recommandé (4 modalités). La réponse en fut : « donner l'âge le plus probable et l'âge minimum par modalité », étant entendu que ces données doivent être cherchées autant que faire se peut parmi la littérature traitant d'une population similaire à celle dont se réclame la personne examinée. Plus précisément, l'âge le plus probable est décrit comme l'âge moyen correspondant à celui estimé par la modalité utilisée. Néanmoins, quel âge rendre parmi les 4 âges moyens estimés à partir des 4 modalités ? Est-ce une moyenne des 4 modalités, une moyenne pondérée ?

Concernant les âges minimaux, une question similaire se pose : si l'on dispose de 4 âges minimaux (autant d'âges minimaux que de modalités), comment doit-on conclure ? La réponse fournie fut la suivante : il est recommandé de rendre « l'âge maximum des âges minimaux estimés ».

En conclusion, deux âges sont censés être rendus à l'autorité requérante : un âge moyen le plus probable, dont il n'est pas clair comment il doit être calculé, s'il peut l'être (problème de la population de référence), et un âge minimal qui s'avère être le plus grand des âges minimaux estimés. Rien n'est dit par ailleurs sur la possibilité de rendre deux âges qui soient en contradiction quant à la minorité de la personne (un âge probable de 19 ans et un âge minimal de 17 ans, par exemple). Un article a tenté de proposer une technique d'intégration des informations, provenant de plusieurs modalités, dont le scanner des clavicules et l'examen des dents. Les performances n'en étaient pas meilleures.[183]

Actuellement, le mot d'ordre en matière d'estimation de l'âge chez les adolescents migrants, repris dans les communications orales ou écrites, est de fournir des résultats qui soient fiables, "au-delà de tout doute raisonnable" ("*beyond reasonable doubt*"). Cette formule, ne reposant actuellement sur aucune idée substantielle, à part peut-être faire appel ou écho à la notion de résultat "probable", ou de technique fiable "en moyenne", est à notre avis une formule dénuée de sens. Un des arguments pour étayer la critique de cette formule tient simplement dans la compréhension et l'interprétation que chacun fait et peut faire d'une probabilité, loin d'être univoque. A partir du moment où une probabilité n'a pas un sens consensuel,[22,184,185] il devient difficile de rallier la majorité à la définition d'un "doute raisonnable" quant aux performances d'un outil statistique.

Eléments de critique des méthodes utilisées pour l'estimation de l'âge

Un certain nombre de critiques peuvent être faites des méthodes utilisées jusqu'ici à des fins d'estimation de l'âge chez les adolescents migrants. Une des plus importantes étant d'ordre éthique, plusieurs fois discutée, souvent sinon systématiquement balayée par les partisans de ces méthodes.

Des critiques tout à fait factuelles peuvent être également formulées. Il peut être supposé que l'ethnie, l'origine migratoire des adolescents invalidaient l'utilisation de techniques généralement testées sur des populations différentes de celles des migrants.[186–190] Il existe des différences inter sexes, en termes de maturation osseuse par exemple, puisque cette

maturation est en partie sous contrôle hormonal. Les différences génotypiques et phénotypiques des populations d'origines migratoires différentes ont également des chances de pouvoir se traduire par des différences de maturation osseuse.
Une étude allemande a suggéré que l'origine migratoire n'impliquait pas de différence en termes de maturation osseuse mais que le niveau socio-économique, estimé via l'indice de développement humain (IDH) du pays d'origine de l'adolescent, expliquerait en revanche une partie des différences observées.[191] Reste qu'il semble peu probable que les adolescents migrants qui sont soumis aux examens d'estimation d'âge soient parfaitement représentatifs de la population "moyenne" du pays d'origine, donc reflété par l'IDH.
En outre, les mêmes études soulignent qu'un facteur important pouvant entraîner des différences de maturation osseuse ou de développement physiologique général, tient à l'état de santé des personnes. Or, si l'on sait que l'état de santé des migrants est en général meilleur que celui des non migrants, quand la raison de la migration n'est pas par ailleurs une raison sanitaire,[154] certains motifs poussant les adolescents à migrer ne semblent pas compatibles avec cette observation : les exilés, qui fuient une région en guerre, ont pu subir des persécutions et de mauvais traitements, les « exploités », victime de la traite humaine, ou encore les « errants », qui passent de pays en pays, traversent plusieurs frontières et vivent de divers expédients.[192] La situation des adolescents migrants en termes de santé paraît pour le moins contrastée.

Les études avec âge chronologique connu et contrôlé ne portent en général pas sur les populations concernées des pays d'origines, dont certaines n'ont pas même, dans leur pays d'origine, une date de naissance claire et connue. Les recommandations de l'AGFAD disant qu'il faut se référer aux données concernant les populations correspondant à l'origine migratoire de la personne correspondent donc avant tout à un vœu pieux. Dans une veine similaire, il n'existe pas de données ouvertes ou mises en commun et disponibles pour tester et reproduire les résultats obtenus par les diverses équipes : la reproductibilité des résultats, alors même qu'ils peuvent prêter à la critique dans leur exposé initial, ne peut être assurée.

Plus profondément encore, à notre sens, il existe dans les études d'estimation de l'âge une confusion des objectifs, et donc des méthodes mobilisées pour atteindre ces objectifs. Désire-t-on prédire ou bien plutôt expliquer ?[193] Formulé autrement, les études d'estimation de l'âge visent-elles à expliquer dans quelle mesure une observation, fût-elle radiologique, dentaire, physique, décrit ou reflète l'âge chronologique à différents moments, avec plus ou moins

d'incertitude, ou bien visent-elles davantage à « prédire », ou dirons-nous plutôt, à classer une personne donnée soit parmi les personnes mineures, soit parmi les personnes majeures, avec la meilleure efficacité (c'est—à-dire, avec des taux de faux positifs et faux négatifs les plus bas possibles, sinon nuls) ? En effet, selon les cas, les techniques disponibles et mobilisables les plus adaptées ne sont pas les mêmes.

De fait, en pratique, une grande partie des études d'estimation de l'âge tente a priori de répondre indifféremment aux deux questions sans prendre la peine de les formuler explicitement – ces études sont souvent des études de corrélation ou d'association :[194] corrélation entre des stades identifiés par une modalité donnée (par exemple, l'imagerie IRM), et l'âge chronologique, connu par ailleurs. Or, une bonne corrélation (linéaire) ne saurait systématiquement induire une capacité satisfaisante à classer des individus par rapport à la minorité. De fait, une étude de corrélation, qui n'est pas synonyme de causalité, renseigne davantage sur de possibles facteurs descriptifs de la croissance osseuse, par exemple, que sur des facteurs prédictifs satisfaisants. On peut en outre souligner qu'en d'autres contextes, on aurait plutôt tendance à connaître l'âge des individus, et d'étudier l'aspect des tissus, l'expression de substances selon une modalité d'observation en fonction de l'âge, et non l'inverse. Par capacité prédictive satisfaisante, nous entendons une technique, si elle existe, qui soit capable de déterminer avec la plus haute certitude un individu particulier, et non renseigner sur le fait que, possiblement, 85% des adolescents étudiés sont correctement classés, et donc 15%, en moyenne, ne le sont pas (que ce soit classés à tort majeur ou mineur).

L'autre type d'étude majoritaire consiste à déterminer des stades de développement (dentaires, osseux), et de déterminer l'âge moyen des personnes par stade. Les stades ainsi obtenus ne permettent pas de déterminer de plages d'âges qui ne se recouvrent pas : les classes d'âges associées aux différents stades ne sont pas exclusives les unes des autres.

Le développement actuel de la recherche sur le thème de l'estimation d'âge chez les adolescents migrants repose essentiellement sur la multiplication des modalités, notamment d'imagerie (TDM,[182] puis IRM,[195] mais également échographie osseuse[196]), des sites anatomiques (épiphyses distales du radius et de l'ulna, clavicule), et partant des stades ou sous-stades de développement, et enfin éventuellement sur l'intégration de plusieurs modalités. La logique reposant sur l'intégration de plusieurs modalités est celle d'un accroissement de la précision de l'estimation : puisque les stades décrits ne permettent pas de discriminer des plages exclusives, mais qu'éventuellement ces plages ne sont pas exactement

les mêmes selon les modalités, alors l'intégration de différentes modalités pourraient permettre des recoupements intéressants.[183]

Enfin, au-delà des techniques mobilisées, se pose le problème de la formulation de la réponse aux autorités requérantes : quel que soit l'examen, quelle que soit la méthode, la question demeure généralement celle de la majorité ou de la minorité de la personne présentée. Etant donnée l'absence de techniques permettant actuellement de discriminer parfaitement les âges chronologiques, il ne saurait être répondu péremptoirement un âge spécifique et annoncé comme certain. Quelles sont donc les formes des réponses médicales à la question judiciaire ? Le PHRC multicentrique (Estimation de l'âge de l'adolescent à des fins juridiques : déterminants de la réponse médicale. AOR 11 118 (2012-2015)) devrait pouvoir, pour le cas de la France, apporter des éléments de réponse à cette question. Il est déjà possible d'annoncer une grande variété des formes de réponses (compatibilité avec l'âge allégué, probabilités d'âge, probabilité de majorité, fourchettes d'âges, probabilités de fourchettes d'âges, réponses chiffrées ou réponses littérales) sans préjuger de l'utilité de ces différentes formulations pour l'autorité judiciaire.

Méthodes non classiques, non statistiques et massivement multivariées pour l'épidémiologie et la médecine légale

Rappels sur la méthode expérimentale en sciences biomédicales

L'épidémiologie sociale et la médecine légale sont toutes deux liées à la biomédecine, par construction et nature. Toutes deux ont un rapport particulier aux méthodes canoniques de la biomédecine. En raison de leurs objets en particulier (l'échelle des populations, les situations de violences et leur traduction en termes pénaux), la méthode expérimentale classique leur est peu accessible. Ce pourrait être vu au moins comme une limite, sinon comme un obstacle rédhibitoire à leur meilleure considération scientifique ou simplement fonctionnelle. Ce peut être également vu comme une opportunité et une chance de se dégager des principaux écueils d'une recherche standardisée qui, elle aussi, souffre par ailleurs d'importantes limites : le caractère artificiel des résultats, leur portée explicative ou prédictive limitée en contexte de « vie réelle » et le caractère très normatif qu'elle entraîne, en partie par construction.

Modèle de Descartes et de Bernard

Le modèle expérimental classique, traduit dans un langage technique propre à la biomédecine au travers des essais randomisés contrôlés, est celui de la méthode expérimentale et réductionniste. On peut rapporter, dans une première approche, cette méthode à deux noms ou moments de l'histoire des sciences, que sont René Descartes[197] et Claude Bernard[198] – il faut néanmoins se garder de généralisations hâtives ou réductrices : on sait que les sciences n'ont pas une progression linéaire.[19]

Le réductionnisme tient pour méthode qu'un problème, compliqué ou complexe, doit pouvoir se décomposer en un ensemble de problèmes plus simples et isolables ; la résolution des problèmes élémentaires doit permettre, par additivité des causes et des effets, de résoudre le problème global. L'additivité des causes et des effets s'entend comme la propriété de linéarité en mathématiques ou en physique : si une cause A entraîne un effet B, et qu'une cause C entraîne un effet D, alors une cause A + C entraîne un effet B + D. Un peu rapidement, le réductionnisme se condense souvent dans l'expression « le tout est la somme des parties ».

La méthode expérimentale repose sur la comparaison de deux objets, dont toute dissemblance est neutralisée, à l'exception d'une seule propriété. Cette propriété est celle que l'on fera

varier entre les deux objets d'étude ; la logique sous-jacente étant que si l'on observe, après avoir fait varier une propriété A entre les deux objets, une différence de caractéristique entre les deux objets (qui ne soit pas A, bien entendu), alors cette différence a de forte chance d'être due à l'action de A (puisque toutes les autres caractéristiques pouvant différencier les deux objets ont été neutralisées et fixées). Cette approche se condense souvent sous la formule « toutes choses égales par ailleurs ».

La traduction biomédicale de ces deux approches est l'essai randomisé contrôlé, tenu pour plus haut niveau de preuve quant à un possible lien de causalité entre une intervention et un effet observé. La neutralisation des sources de différences entre groupes est assurée par la randomisation : les variabilités inter individus (taille, poids…) sont censées être réparties également entre les deux groupes. La comparaison est assurée par le contrôle de l'expérience : il existe un groupe qui recevra l'intervention d'intérêt (classiquement, un médicament), et un groupe de contrôle, qui ne recevra par l'intervention, mais par exemple un placebo. La différence pertinente à étudier et quantifier est alors celle des effets observés entre les deux groupes.

Puisque l'on étudie le vivant, il est acquis qu'il existe des variabilités plus ou moins importantes inter individuelles selon toutes les caractéristiques mesurables. La prise en compte de la variabilité appelle en général l'utilisation des méthodes statistiques. Ainsi, les outils les plus fréquemment utilisés pour représenter les caractéristiques, pour synthétiser l'information qui définit les individus étudiés, sont la moyenne, la proportion, éventuellement la médiane, et les mesures de dispersions qui les accompagnent, comme l'écart-type ou l'étendue. Cette réduction de la complexité, nécessaire pour pouvoir induire ou déduire de la connaissance, la communiquer, repose sur un certain nombre de présupposés qu'il peut être bon de rappeler. En particulier, cette technique de neutralisation de la variabilité et ce mode de représentation de l'information amènent à développer un discours sur des grandeurs moyennes, sur des considérations de groupes, et non sur des individus – alors que l'effet visé est a priori le bénéfice individuel. Dans certaines disciplines, on constate plus aisément encore l'aspect normatif de ce type d'outil : par exemple en santé mentale, ou pathologie, personnalités et comportements se mêlent et interagissent.

Norme et normativité biologiques de Canguilhem

Le problème du normal et du pathologique n'est pas nouveau, à supposer qu'il existe une frontière nette entre normal et pathologique, ou encore qu'il puisse exister de manière universelle un caractère pathologique pour tout aspect qui relève de la santé. La porte de salut quant à la définition du pathologique par rapport au concept de santé est également barrée, dans le sens où la santé est définie comme « non uniquement l'absence de pathologie ». La question du normal et du pathologique est en fait une double question : celle de savoir si ces deux qualités sont bien définissables, et si elles sont nécessairement deux entités isolées et exclusives.

Depuis Canguilhem,[199] nous avons des arguments et des outils pour penser que la pathologie ne peut se résumer à un écart « significatif » à la moyenne ; la pathologie pouvant par ailleurs être vue ou vécue comme un « autre normal ». La vie est un agencement dynamique, un arrangement avec elle-même, qui permet de se maintenir dans un état plus ou moins stable, plus ou moins pérenne. Cette stabilité peut ne concerner qu'une partie des caractéristiques d'un individu, et même se maintenir au détriment d'autres caractéristiques – ce phénomène est connu de certains mécanismes d'adaptation à court ou moyen termes, comme en cardiologie, par les mécanismes compensatoires d'hypertrophie cardiaque, ou de dilatation cardiaque, qui sont à termes délétères.

Les approches statistiques classiques répondent mal, ou même ne permettent pas de transcrire ce type de représentation – en particulier, l'homéostasie et le caractère dynamique, intrinsèquement variable et incertain des organismes. Enfin, même si cela ne fera pas l'objet de nos travaux, il existe une confusion entre variabilité et aléatoire, et variabilité et imprédictibilité, de même qu'entre modèles et réalités. L'introduction de l'aléatoire dans les modèles mathématiques est avant tout un moyen pratique (non exclusif) de prendre en compte la variabilité ou l'incertitude – cela ne signifie pas que l'aléatoire règle nécessairement le système.[200–205] Inversement, il est parfaitement possible de modéliser un système complexe et peu prédictible à l'aide de techniques ne faisant pas intervenir de considérations aléatoires ou probabilistes.

Les techniques que nous présentons ici sont autant de pistes existantes que nous proposons d'utiliser, et qui ne fassent pas appel à la normativité et permettent à la fois de dégager des résultats de valeur scientifique, de réduire la complexité apparente et de raisonner en termes différents de ceux plus classiquement utilisés.

Techniques de clustering – techniques de partitionnement

L'identification de groupes homogènes ou de classes à partir de données peut s'effectuer grâce à des techniques dites de *clustering,* de reconnaissance de forme (*pattern recognition*), ou encore de partitionnement des données. Quelle que soit l'approche retenue, le but demeure le même : il s'agit de déterminer, dans un ensemble de données, quelle observation particulière peut être rapprochée d'une autre observation particulière, et ainsi, de proche en proche, de déterminer des groupes d'observations.

Nous décrivons ici les grands principes sous-tendant les techniques de *clustering*, puis des algorithmes implémentant ces techniques. Nous expliquons ensuite dans quel cadre il convient d'utiliser ces techniques. Enfin, nous donnerons des clés quant à l'interprétation qui peut être faite des résultats de ces techniques.

Identification de groupes homogènes : principes

Etant données un ensemble d'observations, qui peuvent être des observations de quantités continues, discrètes ordonnées ou catégorielles, le problème qui nous occupe est de 1) déterminer s'il existe des groupes homogènes, 2) s'ils existent, de les définir. Par observations, on entend la données de valeurs concernant plusieurs caractéristiques (âge, taille, poids…), pour n individus (personnes, animaux, voitures…). On parle de groupes homogènes uniquement par rapport à un certain nombre de ces caractéristiques – par exemple, selon la morphologie des personnes : existe-t-il des groupes plutôt bien définis constitués de personnes plutôt grandes et grosses, et de personnes plutôt petites et minces ? Il apparaît illusoire de vouloir chercher des groupes qui soient « infiniment » ou « absolument » séparés, c'est-à-dire, distinguant des ensembles d'individus qui, sans s'opposer, présenteraient toutes leurs caractéristiques à la fois en commun pour ceux d'un même groupe, et à la fois totalement différentes vis-à-vis des individus appartenant à d'autres groupes. En somme, et c'est une autre façon de poser la question : qu'est-ce qu'une forme ?

Enfin, il est intéressant de souligner que la recherche de groupes homogènes peut répondre à deux objectifs complémentaires : chercher à caractériser des ensembles selon ce qui en réunit les individus, ou bien chercher à caractériser ce qui distingue ces groupes. A ce titre, on peut souligner que parmi les mécanismes élémentaires permettant d'identifier des groupes homogènes, les algorithmes peuvent soit recourir à des mesures de similarité, soit à des mesures de dissimilarité.

Principes généraux des algorithmes de *clustering*

Il existe un grand nombre d'algorithmes de *clustering*,[94–96,98] mais la plupart, sinon tous, sont formulables en termes de géométrie. Certains sont plus spécifiquement dédiés à l'apprentissage de formes, dans le but d'une utilisation ultérieure comme moyens de classifications automatiques – on pourra penser aux réseaux de neurones, ou aux SVM (*support vector machines*). Puisque ces techniques nécessitent une phase d'apprentissage, il est nécessaire de disposer de références ou d'experts, qui soient capables de corriger les fautes des algorithmes au fur et à mesure, et ce afin qu'ils « apprennent » à reconnaître le mieux possibles les formes qu'on lui demande d'identifier. Nous ne nous intéresserons pas à ces classes d'algorithmes, même s'il en existe des variantes permettant de partitionner les données (par exemple, les SVC : *support vector clustering*),[206,207] puisque nous nous plaçons dans une situation « en aveugle », où l'on cherche à découvrir des objets dans les données, sans que l'on en ait la connaissance préalable.

Le prototype de l'algorithme de *clustering* est certainement l'algorithme des k-moyennes (*k-means*).[93,208,209] Il résume à lui seul les points clés typiques des différentes classes d'algorithmes.

L'algorithme des k-moyennes repose sur l'utilisation d'une distance, afin de quantifier les positions relatives des différentes observations, les unes par rapport aux autres, et d'un critère de regroupement, qui est un critère d'inspiration physique, un critère d'inertie. Son fonctionnement est donc éminemment géométrique.

Les observations sont plongées dans un espace de dimension égale au nombre de variables prises en compte, et la distance utilisées est la distance euclidienne. Puisque l'on mesure les distances entre observations prises deux à deux, il n'est pas nécessaire de spécifier une origine particulière, absolue. Il est impératif de spécifier à l'algorithme le nombre k de groupes qu'il est censé identifier. A partir de cette information, il définira k « centres de gravité », appelés à évoluer au fur et à mesure que les k groupes se constitueront, agrégeant autour des centres les individus, selon que l'ajout d'un individu à un groupe plutôt qu'à un autre minimisera l'inertie de ce groupe. Lorsque les groupes n'évoluent plus, au bout d'un certain nombre d'ajouts / retraits de points entre groupes, l'algorithme s'arrêtera, et considèrera avoir atteint une solution stable.

On peut remarquer dès à présent que plusieurs aspects sont critiques. Rares sont les algorithmes qui « savent » déterminer automatiquement le nombre de groupes à identifier : il faut en général spécifier à l'initialisation de l'algorithme le nombre de groupes. De fait, il faut d'une part balayer plusieurs nombres possibles, et d'autre part disposer d'un critère permettant d'affirmer qu'il existe plutôt k groupes que n groupes.

Par ailleurs, il n'existe pas qu'une seule distance, mais de nombreuses mesures satisfaisant la définition mathématique d'une distance. Le choix de la distance n'est pas anodin, et peut fortement influer sur les groupes identifiés. Considérons le choix entre la distance euclidienne, et la distance dite de Manhattan. La distance euclidienne est une distance « à vol d'oiseau » : elle calculera la plus courte mesure entre deux points sans prendre en compte aucune contrainte de « terrain ». La distance de Manhattan, elle, est au contraire une distance qui est contrainte par le « terrain » : elle doit son nom à la ville, et à la façon de calculer la plus petite mesure entre deux points, sachant qu'il n'est pas possible de survoler la ville, mais qu'il est obligatoire d'emprunter les routes qui sont des segments orthogonaux entre eux. Ainsi, la distance euclidienne est réputée adaptée aux calculs de distances entre deux points qui seraient plongés dans un espace de valeurs continues : par exemple, les tailles en cm. La distance de Manhattan, elle, serait à privilégier pour des distances entre deux points qui seraient plongés dans un espace de valeurs plutôt discrètes : par exemple, être de genre masculin ou de genre féminin. Dans les deux cas, le choix de la distance est censé rendre compte des spécificités des observations : de valeurs continues, ou de valeurs discrètes.

Ensuite, le choix du critère arrêtant l'algorithme – ici, un critère d'inertie – peut mener dans un certain nombre de cas à des solutions sous-optimales, c'est-à-dire à des extrema locaux. En effet, l'algorithme peut très bien se stabiliser autour d'une solution qui certes n'évolue plus lors de petits changements, sans que ce soit la « meilleure » solution. On peut atteindre des extrema locaux différents, s'ils existent, selon les conditions initiales, qui peuvent être différentes à l'initialisation de l'algorithme.

Enfin, selon les critères de formation des groupes, les formes de groupes qu'il est possible d'obtenir peuvent être particulièrement contraintes, empêchant ainsi d'identifier des formes de groupes parfois très intriqués ou très contournés. Dans le cas des k-moyennes, le critère d'inertie et le choix d'une distance euclidienne amène à n'identifier des groupes que de formes ellipsoïdes (respectivement, hyper-ellipsoïdes, en dimensions supérieures à 3).

Cadre général d'utilisation des techniques de *clustering*

Une fois ces principes connus, il faut être conscient du cadre dans lequel il est possible ou admissible d'utiliser les techniques de *clustering* dans le but d'identifier des groupes homogènes.

Lors de l'utilisation d'un algorithme de *clustering*, les données sont présentées à l'algorithme sous une forme particulière – entendre par là, qu'il existe une observation numéro 1, puis une observation numéro 2, jusque l'observation numéro n. Dans le cas des k-moyennes, il faut commencer à partir de k centres d'inertie, temporaires, souvent choisis arbitrairement ou aléatoirement. Le choix des centres initiaux, ainsi que l'ordre dans lequel seront prises les observations peuvent mener, comme nous venons de le voir, à des solutions « locales », sous-optimales, et l'interversion de l'ordre ou un choix initial alternatif des centres peuvent mener à l'identification de groupes différents selon les configurations. Il faut donc, on s'en aperçoit, s'affranchir au maximum de ces défauts, notamment de la sensibilité aux conditions initiales de l'algorithme, afin de proposer des objets qui soient les mieux définis ou les plus adaptés aux données. Une manière de procéder, dans son principe, tient à réitérer l'application d'un même algorithme aux données, en mélangeant aléatoirement l'ordre de présentation des données, et en allouant, dans le cas des k-moyennes, les centres initiaux aléatoirement. Il reste ensuite, pour un nombre « suffisant » d'itérations, à comparer les différents résultats obtenus, pour chaque itération. Nous verrons plus loin les critères sur lesquels il est possible de s'appuyer pour décider de la meilleure solution.

Le pendant de cette question de la sensibilité aux conditions initiales est le problème de l'échantillonnage des données, déjà un peu abordé précédemment. En effet, si la sensibilité aux conditions initiales est une problématique qui se pose une fois les données acquises, le problème de l'échantillonnage se pose avant le recueil des données. Afin d'éviter des configurations où certains individus soient surreprésentés par rapport à d'autres, entraînant des déformations des groupes, notamment par des déplacements des centres d'inertie, il faudrait pouvoir s'assurer que les situations d'intérêt ont été échantillonnées de manière équilibrée et représentative de leur existence « naturelle ». Or, dans un contexte où l'on cherche en aveugle à identifier des groupes, il paraît difficile a priori de viser un

échantillonnage parfait, à tout le moins adapté au mieux aux groupes à identifier. Les règles s'appliquant ici sont des règles d'usage plus large : éviter les sous-représentations de certains individus dont on sait par exemple qu'ils auront tendance à ne pas répondre, ou encore standardiser les observations, de telle sorte que certaines dimensions ne « pèsent » pas plus lourd que d'autres, et déforment à leur tour la géométrie des groupes.

Selon le type de données à traiter, la question du choix d'un algorithme plus adapté qu'un autre peut se poser. On peut se retrouver dans les situations suivantes : traiter des données continues, des données discrètes, ordonnées ou non, ou encore des données mixtes (continues et discrètes). Une grande majorité des algorithmes de *clustering* ont été conçus pour des données continues, une minorité pour des données discrètes, et encore moins pour des données mixtes. Néanmoins, il faut se souvenir que ce qui distingue généralement un algorithme conçu pour des données continues d'un algorithme conçu pour des données discrètes tient au type de distance ou de mesure de similarité / dissimilarité retenu. De fait, dans certains cas, plus que selon le type de données, c'est selon le type de problème que l'on a à traiter qu'il faut pouvoir choisir l'algorithme à employer. Il n'est pas aberrant, a priori, d'utiliser un algorithme pour valeurs continues sur des données discrètes si, structurellement, la notion de distances « à vol d'oiseau » a un sens. Un des algorithmes les plus simples et polyvalents est l'algorithme PAM – *Partitioning around medoïds*, qui est une variante de l'algorithme des k-moyennes, moins contraint, et s'appliquant à des données mixtes, continues, ou discrètes.[95] En outre, il existe une autre solution pour pouvoir utiliser des algorithmes destinés à des valeurs continues, lorsque l'on ne dispose que de valeurs discrètes : recourir à une méthode de transformation des données discrètes en données continues – en changeant le référentiel de représentation. On citera ainsi l'usage de l'analyse en correspondances multiples, qui permet de passer de données discrètes à des données continues. Théoriquement, cette transformation est conservative, pour peu que l'on garde le même nombre de dimensions en sortie qu'en entrée.

Le moment venu d'appliquer une ou des méthodes de *clustering* à un ensemble de données, il faut pouvoir s'assurer du caractère non artéfactuel des groupes identifiés. En effet, un algorithme, à moins qu'il ne rencontre un cas particulier qui le mène à une erreur d'exécution, rendra toujours un résultat. Ce résultat sera un nombre k demandé de groupes, quelle que soit leur réalité. S'il n'existe pas de critère unique ou parfait qui puisse trancher cet aspect, certains ont tout de même été proposés et sont utilisés. L'un des critères qui ont été proposés

très rapidement est le critère de la silhouette. Les valeurs que peut prendre la silhouette indiquent sur le degré d'artificialité potentielle des groupes identifiés – néanmoins, il s'agit de valeurs empiriques.[95]

En dehors des critères dont l'intérêt principal est d'attester le caractère artéfactuel ou non des groupes identifiés, il en existe qui renseignent sur plusieurs aspects à la fois. En particulier, le critère de robustesse des groupes informe à la fois sur la stabilité des groupes identifiés et sur le nombre optimal de groupes à découvrir.[210,211] Le principe en est le suivant : étant données des observations, présentées selon un ordre particulier, il va être appliqué le même algorithme un nombre n de fois ; entre chaque itération, une proportion p des observations sera aléatoirement mélangée (leur ordre de présentation). Pour chaque itération, il sera noté si 2 observations données sont classées ensemble dans le même groupe ou non. Si quel que soit le nombre d'itérations et pour une proportion suffisamment importante de « mélange », la plupart des couples sont classés ensemble – ou non – et n'oscillent pas d'une fois sur l'autre – ensemble, séparé, ensemble, séparé, etc. – alors on peut considérer que les groupes identifiés sont robustes pour cet algorithme, par rapport aux fluctuations aléatoires de présentation des données. En outre, si l'on applique cette procédure pour différents nombre k de groupes à identifier, la robustesse des différents groupes sera amenée à varier, ce d'autant qu'il existe des groupes « naturels ». Ainsi, en balayant plusieurs possibilités de k, et en examinant et visualisant les robustesses moyennes et leur dispersion selon k, il est possible de déterminer le nombre optimal de groupes à identifier au sens de la robustesse. Enfin, il devient possible de réaliser le même protocole pour différents algorithmes de *clustering* : si l'on tombe sur un consensus en termes de nombre optimal k, pour ces différents algorithmes, on dispose d'arguments forts pour se prononcer sur le nombre de groupes présents dans les données (au sens de la robustesse des groupes). En effet, si pour 10 algorithmes relativement différents, pour des ordres de présentations des données radicalement différents et aléatoirement choisis, la grande majorité des couples restent les mêmes pour toutes les itérations, il devient vraisemblable que l'on a mis à jour une structure particulière dans ces données.

Le critère de robustesse des groupes nous permet d'aborder le dernier aspect, non négligeable, du cadre d'utilisation des techniques de *clustering*. La vaste majorité des techniques ont des paramètres, des valeurs d'entrée qu'il est nécessaire de renseigner. En particulier, le paramètre le plus fréquemment rencontré est le nombre k de groupes à identifier. De rares algorithmes sont capables de déterminer le nombre optimal de groupes, c'est par exemple le cas des

algorithmes SOMs – *self organized maps* – qui sont des techniques dérivées des réseaux de neurones. Là encore, dans le cas où nous ne sommes pas dans un contexte d'apprentissage, pour lequel une source externe peut intervenir en tant que référence car connaissant a priori les groupes et leurs caractéristiques, des critères doivent être donnés pour obtenir les paramètres les plus optimaux possibles.

Interprétation en termes d'invariants et de profils

Une fois des groupes identifiés à partir des observations, quelles interprétations en donner ? La réalité, l'objectivité et donc la stabilité des groupes découverts sont des critères principaux dans notre démarche. De manière générale et idéalement, nous aimerions pouvoir considérer les groupes comme des invariants du système observé, comme des objets identifiables absolument par tout observateur. Nous avons vu que cette position est certainement utopique. En revanche, étant données une définition suffisamment précise des conditions d'observations et des modalités de mesure, d'autres observateurs devraient être en mesure, éventuellement et préférablement à partir d'autres approches ou algorithmes, d'identifier des groupes très sensiblement similaires.

La notion d'invariance est une notion géométrique et qui fait carrière en physique.[37,212] Elle doit cependant être considérée avec attention. Il n'est pas possible, ni souhaitable, d'identifier des groupes qui soient invariants dans l'absolu – invariants quel que soit l'algorithme utilisé, quelle que soit le type de transformation appliquée. Le pendant de l'invariance, c'est la définition de l'ensemble des transformations qui sont censées laisser les objets invariants. Cette question est cruciale, et ne peut, à ce jour et à notre avis, être évacuée dans le cas le plus général. Elle mérite néanmoins qu'on lui manifeste le plus grand intérêt.

Supposons que l'on ait identifié des groupes, qui sont invariants par toute procédure qui change l'ordre de présentation des données. Un groupe se définissant par les relations deux à deux – ou plus – des éléments le constituant, il n'est pas choquant de considérer ce type de transformation comme laissant nécessairement invariant les « vrais » groupes. Dans le cas contraire, cela impliquerait que l'ordre dans lequel on considère les personnes détermine la structure des groupes, là où l'on suppose que soit leurs interrelations, soit simplement leurs caractéristiques personnelles sont importantes. A ce stade, quelles peuvent être les autres transformations, portant sur les caractéristiques des personnes laissant invariants les groupes ? Précisément les transformations qui n'ont aucun impact sur la structure des groupes.

Inversement, les transformations portant sur des caractéristiques des personnes qui ne laisseraient pas les groupes invariants sont susceptibles d'être des facteurs causaux, structuraux de la typologie. Ainsi, si l'on partitionne l'ensemble des caractéristiques des personnes en un sous-ensemble de caractéristiques n'entraînant aucune déformation ou remaniement des groupes, et en un sous-ensemble complémentaire, alors nous disposons de l'ensemble des facteurs (observés) potentiellement causaux de la typologie. De fait, la recherche des invariances mène à la fois à la caractérisation des objets et à la caractérisation de leur mécanique.

Plus globalement, la recherche d'invariants et d'objets d'études qui soient qualifiables de "scientifiques" ramène à la question de la forme. Lorsque l'on veut identifier des groupes dans un ensemble de données, il faut se souvenir d'au moins deux éléments : on ne voit que ce que l'on est prêt à voir, et ce que l'œil (et le cerveau, en ligne directe) discrimine comme une forme n'est pas nécessairement isolable absolument et « démêlable » point par point.
On ne voit que ce que l'on est prêt à voir : quiconque a déjà rivé ses yeux à un microscope afin d'examiner des lames d'histologie, par exemple, se souviendra combien il a dû se convaincre qu'il voyait bien ce que l'on lui disait devoir distinguer. Il en va de même pour nombre de représentations (pour continuer dans les exemples médicaux : une radiographie, une échographie…), dont on apprend le vocabulaire, la sémantique concurremment de son interprétation, de sa compréhension. Il n'y a pas d'observations pures en elles-mêmes mais des observations intelligibles dans un cadre explicatif.
Une forme n'est pas nécessairement isolable absolument : ce serait même plutôt la règle. Si nous pensons distinguer certaines formes dans un ensemble, c'est bien parce que nous sommes convaincus d'y discerner un sens particulier. Lorsque l'on en arrive à programmer un automate, un ordinateur, afin qu'il identifie automatiquement, sans aide, ces mêmes formes, il en est bien souvent peu capable. Cette incapacité a plusieurs raisons, mais l'une d'elle est tout simplement liée au fait que l'on peut très bien, qualitativement, et même quantitativement jusqu'un certain degré, discourir de ce que le cerveau distingue dans un ensemble, qui conserve un degré de flou quant à la caractérisation exacte de la forme désignée. Enfin, l'attention se focalise sur une forme, un sens particulier au sein d'un ensemble : le complémentaire, ce qui resterait de l'ensemble si l'on ôtait la forme d'intérêt aurait certainement peu de sens. Pourtant, il reste vraisemblablement dans cet ensemble d'autres objets qui pourraient attirer l'attention, ou se définir de manière tout aussi valable. Il peut ainsi

exister des enchevêtrements de formes, qui ont des "valeurs" similaires lorsque l'on examine un ensemble de données. Où commence une forme, où finit-elle ?

Enfin, est-il légitime de considérer qu'il existe un seul nombre optimal de groupes à identifier dans un même ensemble ? La notion de sémantique joue un rôle, nous venons d'en convenir, mais il existe tout autant la notion d'échelle. Comme nous le verrons au travers d'exemples pratiques, étant donné un ensemble d'observations, il n'est pas ridicule de considérer qu'il y existe deux objets clairement définis, et de façon tout aussi valable, qu'il en existe quatre.
Comment cela est-il possible ? Tout simplement par un effet de résolution ou d'échelle : on peut très bien supposer que les organismes sont constitués de cellules, comme éléments de base. Néanmoins, ces cellules contiennent des organites, clairement identifiables également. Soit deux cellules sous un microscope : il existe bien deux formes, mais si l'on descend en échelle, on peut tout autant considérer qu'il existe quatre formes – les deux cellules, et les deux noyaux, par exemple.

Techniques d'apprentissage de variétés

La recherche de groupes homogènes est une tâche en soi intéressante et valide, et se pose d'autant plus que l'on dispose d'un grand nombre de variables : plus il existe de dimensions observées, plus il devient difficile à l'entendement d'identifier directement des groupes pertinents. Cependant, manipuler un grand nombre de variables nous fait entrer dans le domaine du big data,[213–217] qui présente des problématiques propres que l'on ne peut ignorer sous peine d'obtenir des résultats non valides. Une partie de ces problématiques est liée à la question de la distance utilisée dans l'espace des variables utilisées. La question de la distance est une question de géométrie, et la forme du problème qui se pose suggère également un certain nombre de solutions, à chercher du côté de considérations elles aussi d'inspirations géométriques. Ainsi, pour manipuler correctement les données de haute dimensionnalité, une solution reste de réduire la dimension de ces ensembles à sa plus petite valeur possible – c'est-à-dire, en en préservant la complexité et la structure. Le parti pris ici est celui d'une complexité topologique : une complexité des relations de voisinage. Il est possible de réduire la dimension des données en préservant la proximité, c'est-à-dire, la ressemblance entre les individus, en « apprenant » la forme géométrique du réseau de ces ressemblances, de ces proximités : en apprenant la forme de la variété, au sens mathématique du terme, qui sous-

tend les données. Les techniques d'apprentissage des variétés sont des techniques encore jeunes, peu mobilisées à ce jour ; nous en donnerons une revue brève, non exhaustive, et en comparerons plusieurs aux techniques plus classiques de réduction de la dimension.

Problématique des espaces de haute dimensionnalité et du *Big data*

Le big data n'a pas de définition univoque, ni d'origine spatiale et temporelle claire. La banalité des deux termes big et data n'en faciliteront jamais l'identification, si elle a un intérêt.[213] On rapporte néanmoins volontiers les bases de ce qui constitue les propriétés communes aux définitions du big data à un document du Meta Group, cabinet de conseil américain devenu Gartner, en 2001.[218] Ce document, très court, ne fait pas mention du terme big data, mais définit le "3D analytics", dont seront dérivés les 3, 4 ou 5V qui caractérisent actuellement le big data.

Communément, on s'attache aux minimum à définir le big data comme un objet présent des propriétés de grand Volume, de grande Variété, et nécessitant soit Vélocité de traitement, soit une bonne Véracité (le dernier terme de véracité étant particulièrement mal venu, en cela qu'il porte une forte charge d'idéologie).

Les caractéristiques de volume et de variété sont à considérer essentiellement dans le sens suivant, car source de problématiques intrinsèques au big data : un grand nombre de variables, hétérogènes de nature.

Le sens commun veut souvent entendre le "big" et le volume comme avant tout un grand nombre d'observations : un millier, un million, un milliard, bien plus encore d'observations ou de réalisations d'un même type de variable – par exemple, une glycémie, une taille. Fondamentalement, les considérations d'espace mémoire à part, traiter dix comme dix milliards d'observation ne changent que peu d'un point de vue technique. Ce qui entraîne de véritables difficultés est le nombre de variables différentes : traiter les informations de quinze ou vingt variables est déjà en soi beaucoup plus problématique que de traiter un milliards de tailles. Ainsi, par volume, il faut avant tout s'inquiéter du nombre de types de mesures effectuées. On définit la dimension ou dimensionnalité d'un ensemble de données par le nombre de variables différentes traitées – si possible, toutes indépendantes les unes des autres, tandis que la dimension intrinsèque est la plus petite dimension nécessaire et suffisante à décrire les données considérées.

Deux grands phénomènes, certainement liés, mais apparus distinctement, concourent à rendre particulière la manipulation de données de haute dimensionnalité.

- Le phénomène de l'espace vide : c'est un phénomène qui se comprend facilement, et qui est directement lié à la combinatoire des données.[97]
- La malédiction de la dimension : c'est un phénomène décrit par Bellman en 2003,[219] alors qu'il s'intéressait à des problèmes de programmation dynamique. Il énonce un paradoxe, un résultat contre-intuitif. L'idée naïve derrière l'augmentation du nombre de variables, donc de la dimension des données, est qu'en disposant de davantage de façons de caractériser une personne ou une situation, on devrait pouvoir mieux la cerner, mieux la définir, donc également mieux la séparer en tant qu'individualité des autres personnes plus ou moins semblables. Ces mesures discriminantes reposent pour une vaste majorité, à un moment ou un autre des calculs, sur l'utilisation d'une distance : on mesure les distances séparant deux à deux les individus dans un espace qui les caractérisent (couleur des yeux, taille, poids, genre...etc.). Or, on observe le phénomène inverse : plus la dimension croît, et plus les individus se "ressemblent" aux yeux de cette distance. Ils se concentrent tous dans une périphérie, et tendent à se rapprocher de valeurs semblables. De fait, la majorité des outils d'analyse classiques deviennent inopérants.

Ainsi, si l'on désire travailler sur des espaces de haute dimensionnalité, il est préférable de savoir les réduire à leur dimension intrinsèque – la dimension la plus petite, nécessaire et suffisante à représenter les données fidèlement.

Notion de complexité, complexité topologique et approche de la qualité

Que signifie réduire la dimension des données tout en en préservant la complexité ? De quelle complexité parlons-nous ? Sachant qu'il n'existe pas de définition consensuelle ni toujours très opérationnelle de la complexité, nous prenons le parti ici de traiter du cas de la complexité que nous appellerons « topologique ». Par complexité topologique, nous voulons signifier la complexité des relations de proximité, de voisinage entre individus, entre objets. Elle désigne donc la complexité sous-jacente à la question : dans quelles mesures nous ressemblons-nous ? À quelle distance suis-je de mon plus proche voisin, étant données un ensemble d'observations censées me caractériser ?

Ainsi, si les positions des uns et des autres dans un espace de description sont certes quantifiées (quantification effectuée soit par une distance stricto sensu, soit par une mesure de

similarité / dissimilarité), il faut rendre compte d'une part que ces distances ne sont pas des distances absolues, mais des distances relatives, définissant les positions des individus uniquement relativement les unes par rapport aux autres, d'autre part que cette étape de quantification permet d'aborder une étape qualitative, qui permet de caractériser la forme des relations, locales et globales, entre individus. Par analogie, on peut voir cette démarche comme celle utilisée en technologie numérique, où un échantillonnage adapté du signal en permet la reconstruction qualitative sans perte d'information.

En effet, la reconstruction du réseau des relations entre individus, sa géométrie, renseignent directement sur la dynamique des relations, sur les qualités de l'espace étudié : par l'identification des courbes géodésiques pour la dynamique, par l'analyse de stabilité structurelle pour la qualité de l'espace, entre autres.

La préservation de la complexité « topologique » dans toute opération de manipulation des données implique de facto la préservation des qualités et des dynamiques des données étudiées. L'application de techniques visant la réduction des données et préservant la complexité topologique de ces données impose de définir la notion de dimension intrinsèque d'un ensemble de données.

Notion de dimension intrinsèque ; techniques d'estimation

Le concept de dimension intrinsèque répond à la problématique suivante : partant de N variables, a-t-on besoin, sous l'hypothèse de préserver la topologie des données, c'est-à-dire les relations de voisinage, de N dimensions pour en décrire les qualités ? On peut, par exemple, songer au globe terrestre et à sa surface. Sa sphéricité impose qu'on le représente en 3 dimensions si l'on désire en visualiser la structure sans la déformer géométriquement – c'est un problème bien connu des cartographes qu'il n'est pas possible de représenter en 2 dimensions parfaitement la surface terrestre sans devoir lui faire subir des distorsions angulaires ou métriques. Néanmoins, a-t-on besoin de 3 dimensions pour placer sans ambiguïté possible deux personnes sur la surface terrestre ? Il apparaît rapidement que 2 dimensions sont suffisantes, par exemple les longitude et latitude. Ainsi, globalement et en dehors de cas extrêmes, seulement 2 dimensions sont nécessaires pour décrire des relations « de surface », et non 3.

Plus généralement, pour un ensemble de données quelconque, on a toutes les chances que ses N variables soient au moins partiellement corrélées, autrement dit, informationnellement

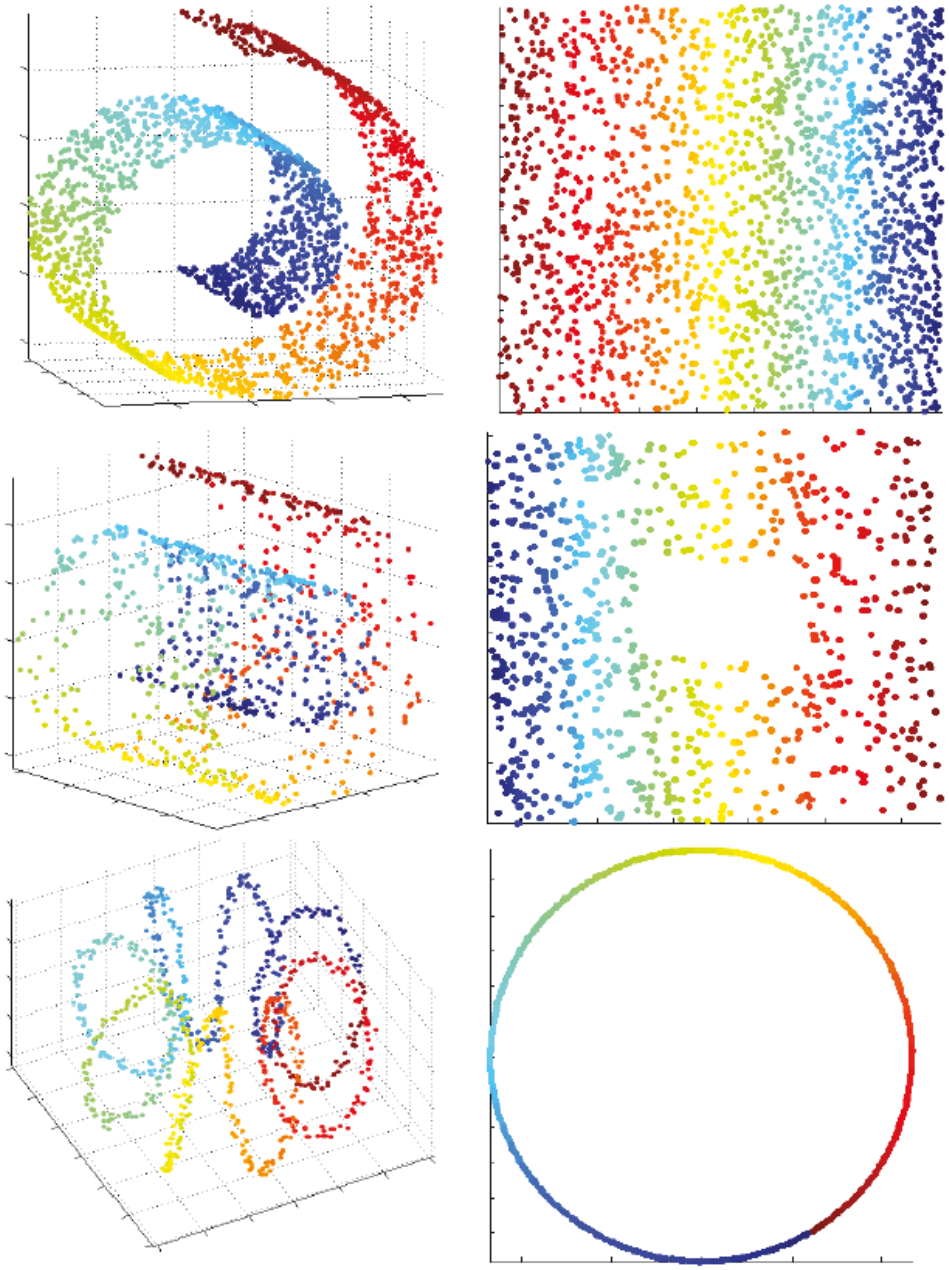
redondantes, et que moins de N dimensions soient nécessaires et suffisantes pour décrire la topologie des données étudiées. La plus petite dimension possible nécessaire et suffisante à caractériser les données est la dimension intrinsèque de ces données. Ce qui signifie que rien n'empêche de les observer dans un espace de plus grande dimension : on peut très bien visualiser une feuille dans un espace de 3 dimensions.

Ceci étant acquis, se pose le problème de la détermination de cette dimension intrinsèque. En pratique, étant donné un ensemble d'observations, comment en estimer la dimension intrinsèque ?

Actuellement, nous disposons de différentes techniques prétendant estimer la dimension intrinsèque des données. Une revue, qui date un peu désormais mais demeure tout à fait pertinente, en proposait notamment une typologie, selon le type d'approche sur lequel la technique repose. On distingue notamment les approches dites locales, globales et fractales.[100]

Parmi ces différentes techniques, nous en avons testé 6 d'entre elles sur 4 exemples classiques d'ensembles de données (rouleau suisse, hélice toroïdale, sommets jumeaux et symbole infini), réputées efficaces et fiables. Chacun de ces exemples tests ont leur intérêt particulier. Le rouleau suisse peut se voir comme une feuille enroulée sur elle-même dont les faces ne se touchent pas, et dont la dimension intrinsèque est 2 – il « suffit » de dérouler la feuille. Une variante du premier ensemble est le rouleau suisse percé, qui teste la robustesse des techniques face à une rupture topologique : le trou. De la même manière que dans le rouleau suisse, il s'agit de ne pas repérer de faux voisinage en « sautant » d'une face à l'autre, il s'agit en sus ici de ne pas « sauter » par-dessus le trou pour relier des points qui ne sont pas topologiquement proches. L'hélice toroïdale est un exemple d'ensemble de dimension intrinsèque 1 – un fil – s'enroulant sur une surface – un tore – tout en se bouclant sur lui-même. La technique d'apprentissage de la variété doit parvenir à dérouler sans briser la topologie, c'est-à-dire, à dérouler le fil sans le déboucler : il s'agit de retrouver un cercle. (Figure 1) Enfin, le symbole infini apporte une complexité supplémentaire, puisque il se recoupe, pouvant occasionner à l'intersection des sauts involontaires de voisinage, tandis que les sommets jumeaux ne sont ni plus ni moins que deux gaussiennes en 3 dimensions, plus ou moins entremêlées. Les résultats des tests concernant les estimateurs de dimension intrinsèque sont reportés en tableau 2.

Figure 1 : Trois ensembles de données tests, visualisés en 3 dimensions (à gauche), et dépliés correctement suivant leur dimension intrinsèque (visualisation en 2 dimensions, à droite) : rouleau suisse (haut), rouleau suisse avec un trou rectangulaire (milieu) et hélice toroïdale (bas)



Trois des estimateurs testés nécessitent des commentaires particuliers. L'estimateur dit PCA (pour *Principal Component Analysis*) est simplement l'application de la technique classique bien connue de l'analyse en composante principale. Sa pertinence n'est pas directement remise en cause, mais nous attirons l'attention sur une précaution d'usage impérative. Lorsque

l'on traite des données catégorielles, discrètes, et que l'on désire y appliquer des techniques conçues pour des données continues, il est fréquent de recourir à l'analyse en correspondances multiples (ACM), qui n'est rien moins qu'une variante de la PCA. Aussi, si l'ACM n'est qu'une étape de transformations des données, sans réduction de la dimension de ces données, il serait malavisé d'estimer la dimension de ces données transformées par la technique dite PCA : cette technique aura toute les chances de renvoyer une dimension égale au nombre de dimensions conservées en sortie de l'ACM.

Les 2 autres estimateurs qui retiendront notre attention sont les estimateurs dits de la « *correlation dimension* », hérité de la théorie de systèmes dynamiques, et du maximum de vraisemblance (*Maximum likelihood estimator* – MLE). Visiblement et empiriquement, ces 2 estimateurs semblent les plus fiables. En l'absence de consensus actuel sur le choix de l'estimateur, nous en recommanderons donc l'usage – de même que nous ne recommanderons pas de se limiter à un seul estimateur.

Tableau 2 : Estimateurs de la dimension intrinsèque pour 4 ensembles de données tests.

	Données tests			
	Rouleau Suisse	Hélice Toroïdale	Sommets jumeaux	Symbole infini
Dimension intrinsèque à retrouver	2	1	2	2
Estimateurs :				
Correlation dimension	1.94	1.47	2.02	2.29
Nearest neighbor dimension	0.56	0.7	0.51	0.44
GMST dimension	1.77	1.38	2.57	2.5
Packing numbers dimension	2.22	1.11	1.31	0.91
PCA eigenvalues dimension	2	2	2	2
Maximum likelihood dimension	1.94	1.5	2.14	2.53

Principes généraux des techniques d'apprentissage de variétés

L'idée même d'apprentissage de variétés repose sur l'hypothèse qu'il existe effectivement, sous-tendant la structure des données observées, une variété, c'est-à-dire un espace, un objet géométrique qui décrivent correctement ces données – plus précisément la structure de leurs relations. Cette variété a donc, au sens mathématique du terme, une dimension, que l'on suppose dans notre contexte plus petite ou égale au nombre de variables considérées, et que l'on assimilera à la dimension intrinsèque de ces données, ainsi que nous venons d'en discuter.

Etant données un ensemble d'observations et sa dimension intrinsèque, il s'agit d'apprendre ou de reconstituer la variété proprement dite. Pour y parvenir, nous disposons depuis le début des années 2000 de techniques dites non linéaires de réduction de la dimension (*nonlinear dimensionality reduction* – NLDR). Depuis les articles séminaux de Tenenbaum[99] et de Roweis[220] en 2000, un nombre sans cesse croissant de techniques d'apprentissage de variété a vu le jour. Beaucoup d'entre elles sont des variantes et des raffinements d'un nombre plus restreint de principes. On a coutume d'opposer les approches locales aux approches globales. Nous ne nous appesantirons pas sur cette classification qui n'apportera que peu d'informations pertinentes dans le contexte de ces travaux. Nous préférons ici présenter en quelques traits les principes généraux de certaines de ces techniques, afin d'en cerner les mécanismes forts et les idées directrices – ce en quoi elles s'opposent ou complètent les techniques classiques de réduction de dimension.

Nous prendrons deux exemples emblématiques : l'algorithme ISOMAP (Isometric feature mapping),[99] pionnier, et l'algorithme LTSA (*Linear tangent spaces alignement*).[221,222]

ISOMAP tire parti à la fois d'une idée extrêmement simple et d'une technique plus ancienne, à savoir la technique MDS (*multidimensional scaling*).[98] L'algorithme se déroule en 3 étapes : Il évalue d'abord la distance euclidienne de chaque paire de points (observations), et en constitue une matrice puis, plutôt que se contenter de cette distance, il va estimer les distances géodésiques entre chaque point – la géodésique entre deux points étant la plus courte distance entre ces deux points. Pour ce faire, il va considérer un certain voisinage pour chaque point, par exemple un voisinage de connexité 4. Une manière simple de se représenter ce type de voisinage, est de considérer un cadran d'horloge : le point d'intérêt est au centre, et les 4 voisins sont à 3h, 6h, 9h et 12h (un voisinage en connexité 6 serait ainsi défini par les voisins à 2h, 4h, 6h, 8h, 10h et 12h, etc.) Ce qui signifie qu'un voisin a 4 voisins directs, lesquels peuvent chacun avoir 4 voisins directs, et ainsi de suite. La distance entre un point et ses 4

voisins directs sera égale à la distance euclidienne reliant ces points. Pour tout autre point qui ne serait pas un voisin du point d'intérêt, la distance qui les sépare sera la plus petite somme des distances euclidiennes calculées entre chaque paire de points les séparant. Enfin, une fois la matrice des distances géodésiques obtenue, l'algorithme applique la technique MDS (*Multidimensional scaling*, ou : positionnement multidimensionnel) à cette matrice, afin d'obtenir la variété recherchée. MDS est une technique qui, sur la base de toutes les distances point à point (c'est-à-dire, toutes les distances entre 2 observations), va déterminer un ensemble de vecteurs dont la norme des différences correspond au mieux aux distances point à point. La donnée de ces vecteurs est ainsi un moyen de positionner les observations les unes par rapport aux autres dans un seul et même espace.

L'algorithme LTSA propose une approche différente, d'intuition très géométrique également. Autour de chaque point de l'ensemble de données, il va tenter de reconstruire l'espace tangent en ce point à l'espace des données. On peut songer, par exemple, à un ballon de football, qui est une sphère représentée par des facettes. Supposons chacune de ces facettes comme autant d'espaces tangents à une sphère véritable, en des points qui seraient les centres de chacune des facettes. Si l'on désire travailler uniquement sur la surface de cet espace, tout comme pour le globe terrestre, nullement besoin en général de 3 dimensions, mais de 2 uniquement. De fait, il s'agit alors de « déplier » le ballon, pour l'aplatir, même si cela occasionnera quelques distorsions – globalement, les relations de voisinage resteront les mêmes (l'exemple est plus simple et correct à visualiser si l'on considère un ballon coupé en son milieu en deux demies sphères). L'algorithme LTSA, estimant donc les espaces tangents en chaque point, s'appliquera à aligner tous ces espaces tangents les uns avec les autres, ce qui entraînera par conséquent l'aplatissement de l'espace – l'alignement des facettes du ballon, quand on l'écrase.

Il existe d'autres techniques, notamment dites « spectrales », qui utilisent les propriétés algébriques des espaces étudiés, ou les techniques dites à noyaux (à kernel), qui tirent parti d'une transformation non linéaire de l'espace initial vers un espace plus « maniable ». Mentionnons également des techniques d'inspiration physique, comme la technique des *diffusion maps* (cartes de diffusion).[223] Enfin, des techniques basées sur les réseaux de neurones existent (les auto-encodeurs,[224] par exemple). Nous n'en décrirons pas les principes plus en détails ici.

De la même manière qu'il existe un certain nombre de paramètres à fixer pour utiliser les techniques de *clustering*, il existe un nombre variable de paramètres à fixer pour utiliser les

techniques de réduction de dimension. Le plus évident, pendant du nombre de groupes à identifier, est la dimension intrinsèque estimée : il faut donner à l'algorithme la dimension de la variété à reconstruire. Autrement dit, il est nécessaire d'estimer avant tout la dimension intrinsèque D des données étudiées, afin de rentrer D en paramètre de l'algorithme de réduction de dimension : celui-ci cherchera la meilleure représentation des données dans un espace de dimension D. Selon les techniques, les paramètres à fixer diffèrent. Dans le cas d'ISOMAP, qui est une technique locale s'appuyant sur l'identification des voisins, il est nécessaire de spécifier le type de connexité – donc le nombre de voisins directs d'un point. De même qu'il est conseillé de ne pas se limiter à un seul estimateur pour déterminer la dimension intrinsèque des données, il apparaît nécessaire de tester différentes valeurs de connexité pour des algorithmes comme ISOMAP (par exemple, un point pouvant avoir 4, 6 ou 8 voisins directs). Enfin, selon la densité de points, selon l'échantillonnage qui peut avoir été fait de la population d'intérêt, il peut exister des zones moins densément peuplées, et imposer un trop grand nombre de voisins directs pour chaque point peut « forcer » de faux voisinages. Néanmoins, si cette critique a pu être faite à ISOMAP, il semble qu'il présente une bonne stabilité topologique.

Revue brève et comparaison des techniques d'apprentissage non linéaires vs techniques classiques (ACP et MDS)

Si nous avons annoncé ne pas vouloir nous lancer dans une revue systématique et exhaustive des techniques non linéaires de réduction de la dimension, c'est en partie en raison de leur nombre croissant, dont les représentants n'offrent par ailleurs qu'une variété qualitative restreinte. Nous préférons donc nous focaliser d'une part sur un petit nombre représentatif de ces algorithmes (6), et d'autre part en comparer les performances à 2 techniques classiquement utiliser en réduction de dimension : l'analyse en composantes principales (ACP) et le MDS (*multidimensional scaling*).[225]
Nous appliquons ces 8 techniques sur 3 ensembles tests usuellement utilisés dans ce type d'algorithmes : le rouleau suisse, le rouleau suisse percé, et l'hélice toroïdale, tous trois déjà présentés dans la section sur l'estimation de la dimension intrinsèque.

Avant d'aborder la comparaison à proprement parler, nous désirons brièvement rappeler le principe sous-tendant les techniques classiques de réduction de dimension, à savoir les

techniques linéaires basées sur l'analyse en composantes principales (ACP)[96,226,227] ou son équivalent pour variables discrètes, l'analyse en correspondances multiples (ACM).[96] L'ACP repose sur des considérations géométriques et physiques simples. Etant données un ensemble d'observations, la technique consiste à chercher les axes principaux de ces données, autrement dit, à chercher les axes dits d'inertie de ces données. Le concept d'inertie est hérité de la mécanique, et est utilisé ici comme un principe de maximisation. D'un point de vue statistique, l'ACP vise à trouver le système d'axes qui permette de décorréler au mieux les variables en utilisant la matrice des corrélations inter observations. Il s'agit de diagonaliser cette matrice, c'est-à-dire d'en trouver les vecteurs et valeurs propres (les valeurs invariantes pouvant former base de l'espace d'observation) et de transformer les variables initiales en des combinaisons linéaires de ces variables, qui s'ordonnent sur une base de décomposition obtenue en classant les valeurs propres associées selon leurs valeurs décroissantes. C'est ainsi que l'on obtient les axes dits principaux, avec le premier axe correspondant à l'axe qui présente la plus grande valeur propre, le deuxième, qui présente la deuxième plus grande valeur propre et ainsi de suite. L'ACP est une technique purement linéaire, et dite projective : l'identification des axes principaux mène à la projection des données sur chacun de ces axes. Ainsi, une voiture, prise comme objet tridimensionnel, serait analysée selon 3 axes principaux, le 1er étant un axe longitudinal d'avant en arrière du véhicule, le 2nd étant un axe transversal suivant la largeur du véhicule et le 3^{e} un axe selon la hauteur, de bas en haut. L'ACP fournit une grille de représentation des données, mais ne dit rien ou peu sur la structure des observations. L'ACM agit similairement pour des variables catégorielles, sur des tables de contingence généralisées plutôt que sur une matrice de corrélations.

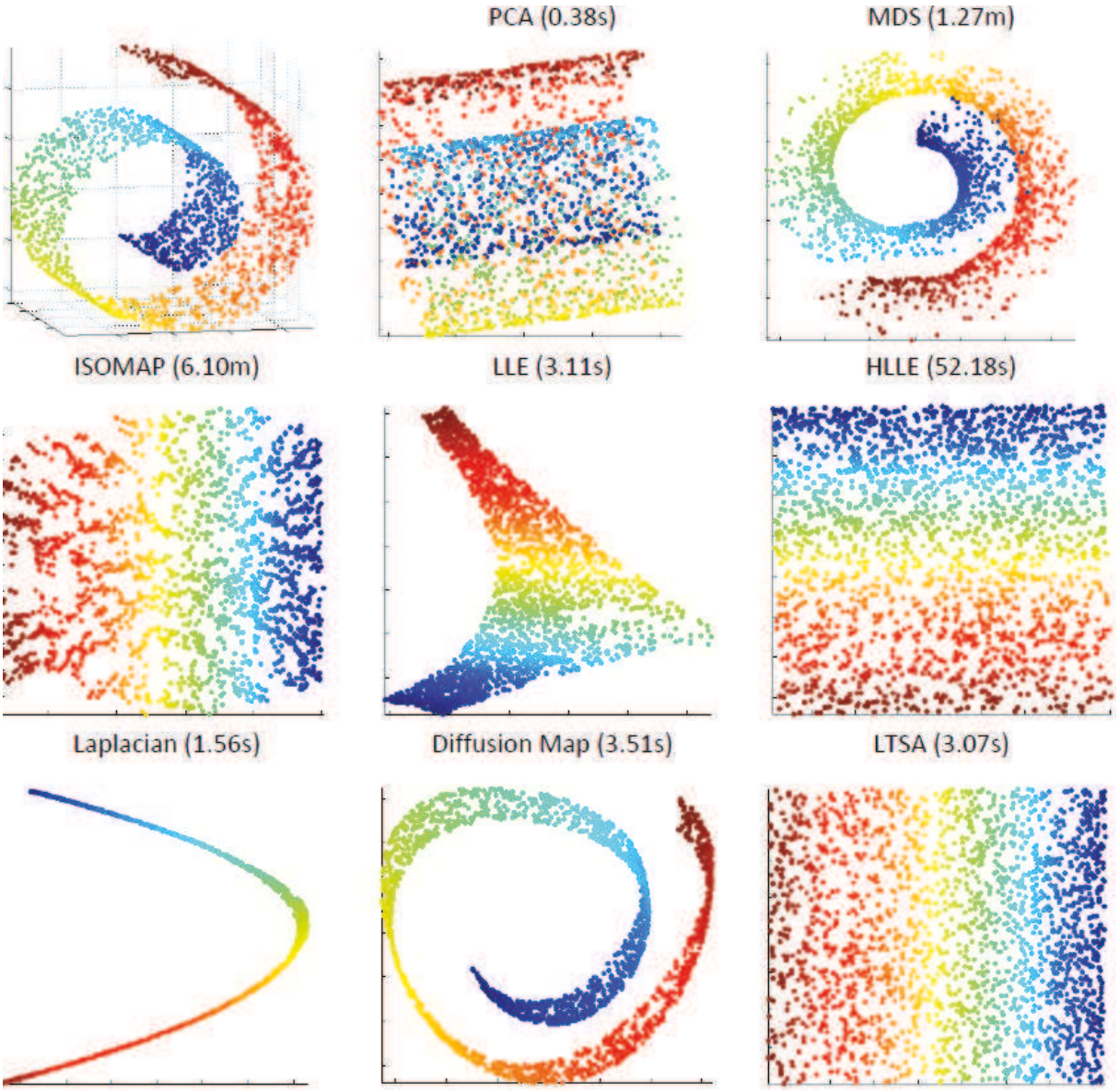
Nous comparons donc ici 8 techniques : l'ACP (PCA en anglais), le MDS, ISOMAP, LLE (*Locally linear embedding*),[220] HLLE (*Hessian locally linear embedding*),[228] les algorithmes *Laplacian eigenvalues* (valeurs propres du Laplacien, qui est un opérateur mathématique de dérivation),[97] *diffusion maps* et LTSA. Les résultats sont reportés aux Figures 2-3-4. En sus de la capacité des techniques testées à déployer correctement les espaces, le temps d'exécution (sur une machine de milieu de gamme, portable 4GB de mémoire et datant de 2010) sont donnés pour chaque algorithme (en secondes s et minutes m).

L'ACP bénéficie d'une technique simple et d'optimisation des techniques d'algèbres linéaires – c'est une technique rapide d'exécution. A l'opposé, ISOMAP est un algorithme en plusieurs étapes, l'une mettant en jeu la recherche de plus courtes distances dans un graphe (par l'algorithme de Dijkstra, par exemple), une autre utilisant MDS, qui est déjà une technique

relativement lente ; ISOMAP est un algorithme lent. En termes de résultats qualitatifs, Il est notable que l'ACP, MDS et *diffusion maps* agissent ou semblent agir comme des techniques projectives : les 2 premiers axes principaux sont un axe de révolution et un axe qui traverserait les couches de la feuille de bas en haut, par exemple. De fait, lorsque l'on commande à l'algorithme de déployer le rouleau en 2 dimensions, il projette bien les observations le long de ces 2 axes, ce qui procure cet aspect « écrasé » des observations, entremêlant des points rouges et bleus – autrement dit, des points qui ne sont aucunement voisins. A l'inverse, ISOMAP, LTSA et HLLE parviennent à déployer correctement le rouleau, ainsi que les deux autres ensembles tests. On notera par ailleurs la rapidité d'exécution de LTSA par rapport à ses concurrents (3s contre 6m pour ISOMAP et 52s pour HLLE).

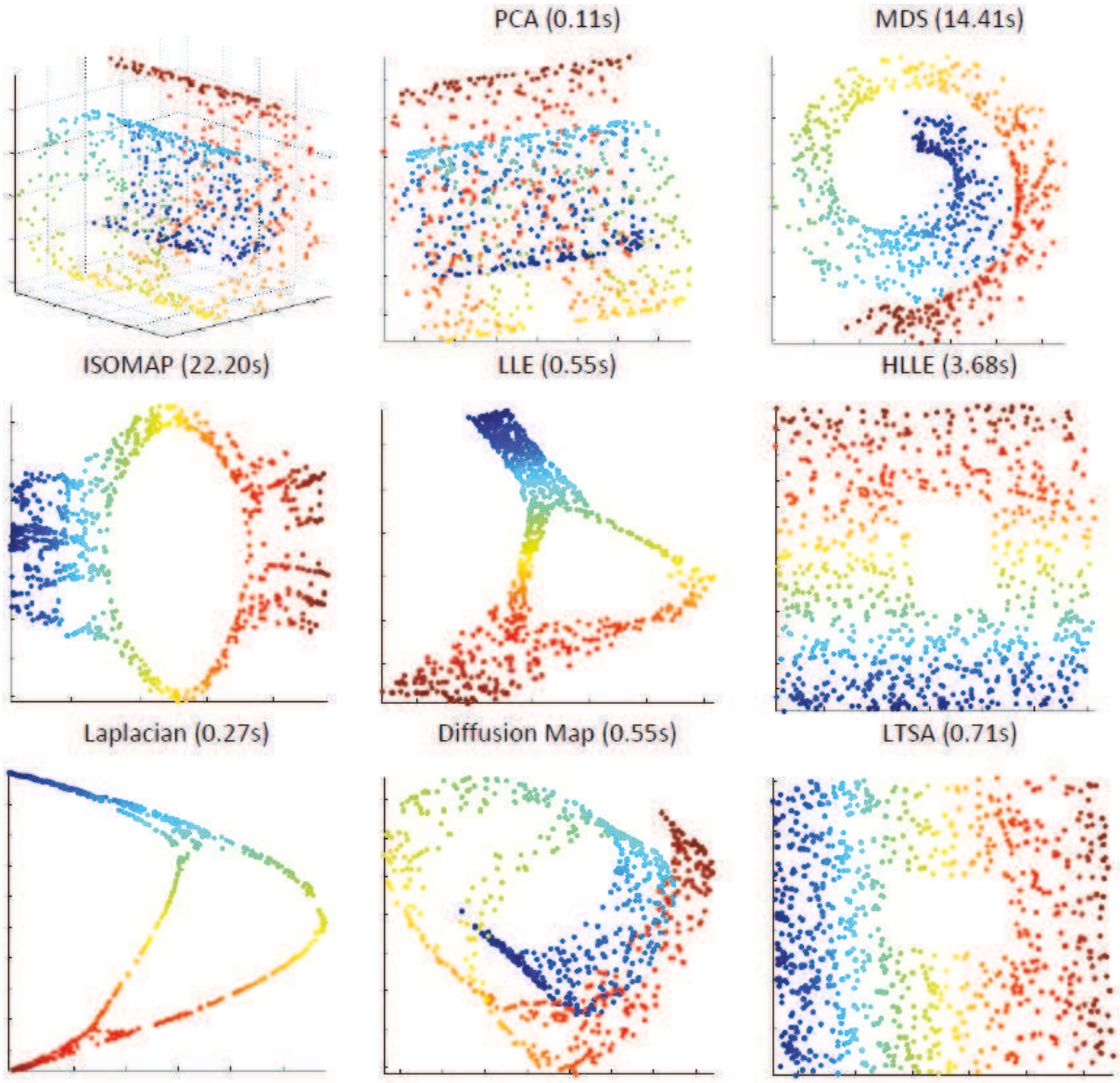
Ces résultats sont évidemment des résultats synthétiques, mais qui servent habituellement à tester tout nouvel algorithme, afin de pouvoir en comparer les performances. Ils n'exonèrent bien entendu pas d'une évaluation et d'une validation sur des données réelles. A ce jour, il n'existe que peu de publications permettant ce genre de tests, en partie parce qu'une telle validation implique que l'on connaisse par ailleurs la structure des données ; or ces techniques servent précisément à explorer et reconstruire des structures inconnues jusque présent, ou difficilement accessibles. Si les techniques non linéaires paraissent au moins aussi performantes que les techniques classiques types ACP, il semble que ce ne soit pas systématique. A l'heure actuelle, nous ne sommes pas en mesure de donner un critère clair qui ferait préférer un type à l'autre.[222,229–232] Plutôt qu'une fin de non-recevoir, cet état est à voir comme un encouragement et une nécessité à tester ces nouvelles approches.

Figure 2 : L'ensemble du rouleau suisse, sur lequel sont appliqués 8 techniques de réduction de dimension. Le nom de l'algorithme utilisé est inscrit au-dessus de son résultat. Le temps d'exécution de chaque technique est indiqué en secondes (s) ou minutes (m).



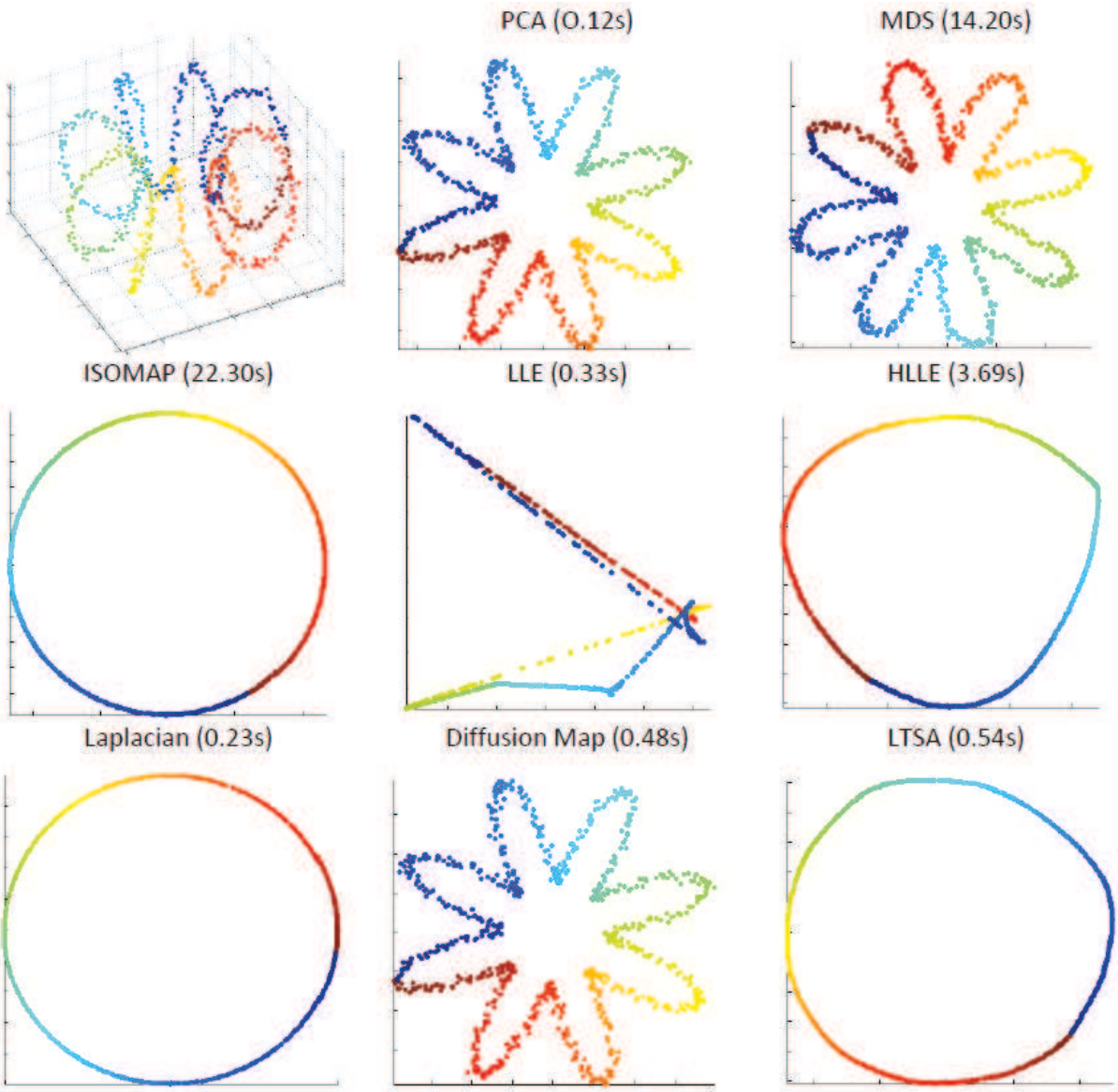
PCA : principal compenent analysis. MDS : multidimensional scaling. LLE : locally linear embedding. HLLE : Hessian LLE. LTSA : local tangent space alignment.

Figure 3 : L'ensemble du rouleau suisse avec un trou rectangulaire, sur lequel sont appliqués 8 techniques de réduction de dimension. Le nom de l'algorithme utilisé est inscrit au-dessus de son résultat. Le temps d'exécution de chaque technique est indiqué en seconds (s) ou minutes (m)



PCA : principal compenent analysis. MDS : multidimensional scaling. LLE : locally linear embedding. HLLE : Hessian LLE. LTSA : local tangent space alignment.

Figure 4 : L'ensemble de l'hélice toroïdale, sur lequel sont appliqués 8 techniques de réduction de dimension. Le nom de l'algorithme utilisé est inscrit au-dessus de son résultat. Le temps d'exécution de chaque technique est indiqué en secondes (s) ou minutes (m)



PCA : principal compenent analysis. MDS : multidimensional scaling. LLE : locally linear embedding. HLLE : Hessian LLE. LTSA : local tangent space alignment.

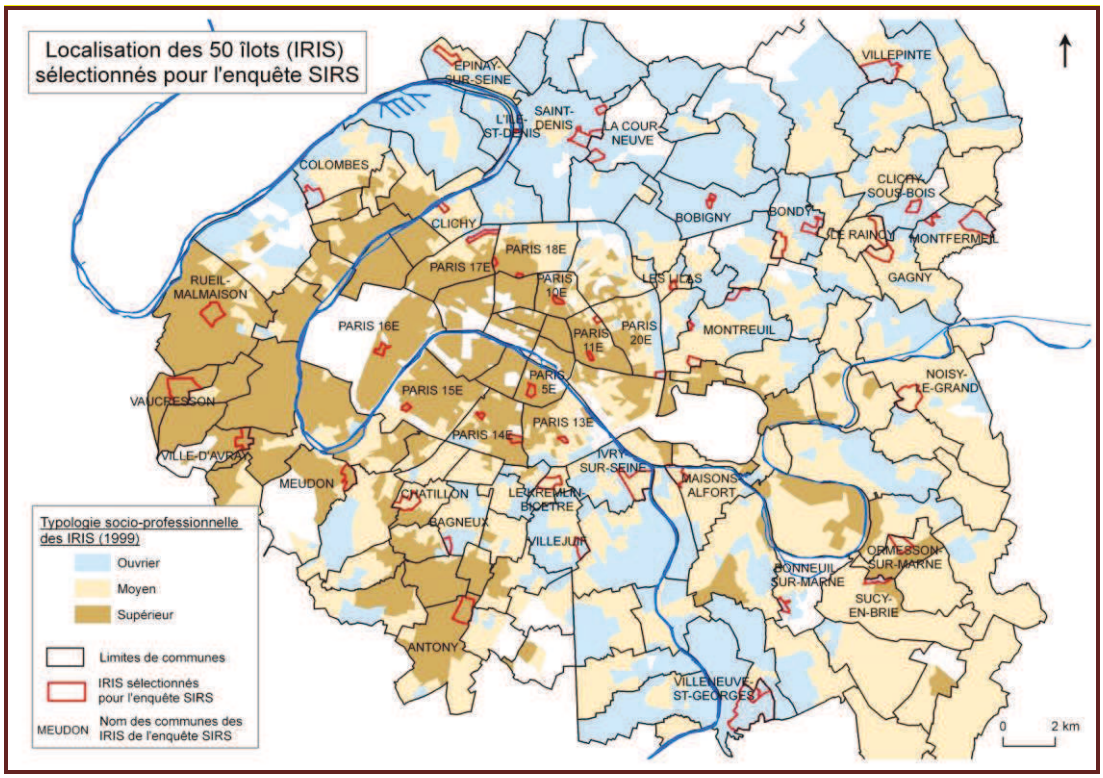
Recherche d'une typologie de recours aux soins : utilisation du cadre général d'utilisation des techniques de *clustering* aux données SIRS

Présentation de la cohorte SIRS, vague 2010

La cohorte SIRS (Santé, inégalités et ruptures sociales) est une cohorte nominative, représentative de la population majeure francophone et vivant en ménage ordinaire dans l'agglomération parisienne (départements 75, 92, 93 et 94), comprenant 3000 ménages inclus à l'automne 2005 par un échantillonnage aléatoire à trois degrés. Le premier niveau de tirage est constitué d'IRIS (unités INSEE comptant environ 2 000 habitants) et stratifié sur la typologie socioprofessionnelle de l'espace francilien de E. Préteceille[233] et le classement en Zone Urbaine Sensible (ZUS) : 50 IRIS ont été tirés au sort parmi les 2 595 IRIS éligibles de l'agglomération, les IRIS en ZUS étant surreprésentés tout comme les quartiers de type ouvrier hors ZUS de la typologie. Au deuxième niveau, 60 logements ont été tirés aléatoirement dans chaque IRIS. Enfin, un adulte fut tiré au sort par logement et inclus dans la cohorte. La première vague d'enquête s'est déroulée en face à face, avec un questionnaire d'une durée moyenne de 90 minutes à l'automne 2005. Une seconde vague (courte) a été réalisée par téléphone en mars 2007, portant essentiellement sur les changements dans la situation professionnelle et familiale des personnes.

Une troisième vague a été réalisée en face-à-face, avec un questionnaire approfondi comme en 2005, au cours de l'automne et de l'hiver 2009-2010. Cette cohorte socio-épidémiologique est conduite par l'équipe ERES de l'Inserm (qui héberge les données nominatives) dans le cadre d'un programme de recherche associant en outre l'équipe de recherche sur les inégalités sociales (ERIS) du Centre Maurice Halbwachs (CNRS-EHESS-ENS). En 2005, le projet a bénéficié du soutien (notamment pour le recueil des données et le suivi de la cohorte) de l'Inserm, de l'IRESP, du Secrétariat interministériel à la Ville (ex–DIV), de la Mairie de Paris et du Fonds Social Européen. Depuis 2009, les recherches conduites sur les données recueillies sont financées sur projet, notamment par l'ANR, l'IRESP, l'ANRS et l'INCa.

Figure 5 : Echantillon d'enquête – Cohorte SIRS



Cette cohorte constitue une première française dans le champ de l'épidémiologie sociale. Il s'agit en effet, à notre connaissance, de la première cohorte représentative de la population générale, constituée ad hoc pour l'étude des déterminants sociaux de la santé, surreprésentant les quartiers en difficulté, géocodant les personnes interrogées et recueillant, en face à face, un nombre important de caractéristiques sociales et sanitaires – y compris dans leur dimension subjective et concernant des caractéristiques rarement interrogées en population générale (insertions et identités sociales, capital social, représentations de santé, raisons de non recours aux soins…). La dimension régionale permet d'étudier l'ensemble du continuum social et certaines dimensions territoriales (en étudiant simultanément les caractéristiques des individus – recueillies dans la cohorte – et les caractéristiques de leur IRIS de résidence – compilées à partir de sources de données extérieures).

Les données de la première vague d'enquête (2005) constituent un ensemble conséquent et unique de caractéristiques sociales et sanitaires. Plus de 400 variables renseignent les dimensions suivantes : statut socioéconomique, conditions de vie, insertions et ruptures

sociales et événements biographiques, rapport au quartier de résidence, histoire migratoire et origine sociale, capital psychologique, santé ressentie, maladies chroniques, IMC, santé mentale, santé des femmes, attitudes et représentations vis-à-vis de la santé et de la médecine, certains comportements liés à la santé (alcool, tabac, activité physique, consommation de fruits et légumes, de viande, de poisson), modalités et fréquence des recours aux soins (curatifs et préventifs), etc.[234,235]

Les données de la troisième vague d'enquête (2010) reprennent de nombreuses dimensions communes, en particulier en ce qui concerne le statut socio-économique des personnes et leurs conditions de vie, auxquelles ont été ajoutées de nouvelles (notamment sur l'alimentation, le dépistage des cancers féminins, l'utilisation de l'offre de soins et les mobilités quotidiennes). Sans rentrer dans les détails du suivi de cohorte et de la méthodologie utilisée lors de cette troisième vague, on donnera néanmoins ici quelques données de cadrage. En 2010, 47% des adultes inclus en 2005 ont pu être réinterrogés (2,6% étaient décédés, 1,8% trop malades pour participer, 13,9% avaient déménagé en dehors des IRIS sélectionnés, 2,7% étaient absents à la période d'enquête, 18,4% ont refusé de répondre et 13,4% ont été perdus de vue sans nouvelle). Leur sexe ratio et leur âge moyen étaient identiques à ceux des non réinterrogés. Les perdus de vue étaient sensiblement plus jeunes et plus aisés que les autres mais le type de leur IRIS de résidence et leur état de santé n'étaient pas différents. Au contraire, les absents au moment de l'enquête étaient, eux, d'un statut socioéconomique plus bas et plus souvent des immigrés. Les personnes incluses en 2005 non réinterrogées en 2010 ont donc été remplacées par tirage au sort selon une méthode identique au sein des 50 IRIS de la cohorte, afin d'obtenir un effectif final de 60 adultes interrogés par IRIS. Le taux de refus des nouveaux enquêtés était, en 2010 comme en 2005, de 29%.

L'échantillon final a été redressé pour prendre en compte la stratégie d'échantillonnage, puis recalé par âge et sexe d'après le recensement de la population de 2006.

Notre étude se base sur l'analyse transversale des données de cette troisième vague d'enquête de la cohorte SIRS d'après le recensement de la population de 2006. L'échantillon final de 2010 comporte 3000 adultes.

Questions retenues

Les questions SIRS retenues pour notre étude concernent d'une part les questions de recours aux soins : soins primaires, accès aux spécialistes, utilisation des différents lieux et modes de pratiques de la médecine, médecines parallèles et alternatives, recours aux soins non programmés et aux services d'urgences (questions D9-F4-F10-F11-F12-F16A-F17-F18) ; d'autre part les questions ayant trait aux potentielles variables explicatives du recours aux soins : variables socio-démographiques, relation aux soins et à la santé, lien social ou encore santé perçue (questions A1 à A7, B1-D2-D4-D14-D24-E1-E2A-E4B-F1-F2-O4A-O11).
Les questions correspondantes sont reportées en annexe, et sont des extraits du questionnaire SIRS.

Gestion des données recueillies et des corrections

Données de recours aux soins

L'espace à partitionner est donc constitué des variables de recours aux soins, toutes catégorielles.
Les variables sont les suivantes :
Soins primaires – 6 variables :

- Date de la dernière visite simple chez le dentiste (4 niveaux – d9 : <2 ans, 2-3 ans, >3 ans, jamais)
- A un médecin généraliste référent déclaré (O/N – f4)
- Fréquence de consultation d'un spécialiste d'accès direct dans les 12 derniers mois (6 niveaux : 0-1-2-3 à 5-6 fois et plus)
- A pratiqué un bilan de santé dans un centre de la sécurité sociale dans les 12 derniers mois (O/N – f10)
- A sollicité l'avis d'une personne du monde de la santé dans son entourage dans les 12 derniers mois (4 niveaux – f17 : 0-1 à 2-3 à 10-11 fois et plus)

Médecine complémentaire ou alternative – 2 variables :

- A consulté un ostéopathe ou un acupuncteur dans les 12 derniers mois (O/N – f18_1)
- A consulté en médecine alternative / parallèle dans les 12 derniers mois (O/N – f18_2)

Recours aux spécialistes d'accès indirect – 1 variable :

- Fréquence de consultation d'un spécialiste d'accès indirect dans les 12 derniers mois (6 niveaux : 0-1-2-3 à 5-6 fois et plus)

Lieux de consultation – 6 variables :

- A consulté un généraliste en secteur public dans les 12 derniers mois (recouvre l'hôpital public de secteur public, le dispensaire) (O/N)
- A consulté un spécialiste en secteur public dans les 12 derniers mois (idem supra) (O/N)
- A consulté un généraliste en secteur privé dans les 12 derniers mois (recouvre l'hôpital public de secteur privé, la clinique) (O/N)
- A consulté un spécialiste en secteur privé dans les 12 derniers mois (idem supra) (O/N)
- A consulté un généraliste en cabinet dans les 12 derniers mois (O/N)
- A consulté un spécialiste en cabinet dans les 12 derniers mois (O/N)

Recours aux urgences – 2 variables :

- Fréquences de consultation pour un motif d'urgence à domicile dans les 12 derniers mois (6 niveaux : 0-1-2-3 à 5-6 fois et plus)
- Fréquences de consultation dans un service d'accueil des urgences dans les 12 derniers mois (6 niveaux : 0-1-2-3 à 5-6 fois et plus)

Facteurs associés

Les variables potentiellement explicatives sont elles aussi catégorielles. Elles sont listées ci-dessous :

Données démographiques – 3 variables :

- Genre (b1)
- Age (5 niveaux – a4 : 18-29 / 30-44 / 45-59 / 60-74 / 75 ans et plus)
- Nationalité / nationalité des parents (3 niveaux – origin_natio : Français, né de parents Français / Français, né d'au moins un parent étranger / étranger)

Données socio économiques – 4 variables :

- Occupation actuelle (3 niveaux : employé / chômeur / inactif)
- Niveau d'étude (3 niveaux – niv_etude : 3e cycle / secondaire / primaire ou aucun)
- Revenu imputé par unité de consommation (5 niveaux : 5 quintiles)
- Couverture maladie (4 niveaux : Sécurité sociale + complémentaire / CMU / Sécurité sociale seule / aucune)

Données rapport à la médecine – 3 variables :

- A un proche atteint d'une pathologie grave (O/N)
- Rapport à la médecine : consulte un médecin quand… (2 niveaux : si ne peut faire autrement / dès que ne se sent pas bien)
- A un proche de l'entourage exerçant dans le (para)médical (O/N)

Données lien social – 3 variables :

- Sentiment d'isolement (4 niveaux : très entouré / plutôt entouré / plutôt isolé / très isolé)
- Soutien social (3 niveaux : élevé / moyen / faible)
- Fréquence des contacts sociaux (4 niveaux : en quartiles)

Données de santé – 3 variables :

- Pathologie chronique ou à caractère durable (O/N – d14)
- A été limité au moins 6 moins dans des activités que d'autres font (O/N)
- Etat de santé perçu (3 niveaux : bon / médiocre / mauvais)

Corrections apportées à la base

La détection d'erreurs potentielles s'est faite en 2 temps.
Dans un premier temps, par examen exploratoire et descriptif des différentes variables étudiées, éventuellement par recoupements entre variables partiellement redondantes, ou avec des proxys laissant penser à des discordances. Par exemple, parmi les fréquences de consultation du généraliste, l'existence d'une seule personne ayant consulté 99 fois dans les 12 derniers mois, alors que la deuxième fréquence la plus élevée tourne autour de 20 fois

l'année, attire l'attention. D'une manière un peu différente, la présence de questions filtres peut amener des réponses aux questions qui soient variables d'une personne interrogée à l'autre, ou d'un enquêteur à l'autre. Par exemple : une personne déclarant avoir consulté un gynécologue (question filtre), mais qui ne donne aucune fréquence, ou une fréquence nulle, de consultation quel que soit le lieu possible de consultation (hôpital, ville etc.) ; inversement, une personne déclarant n'avoir vu aucun gynécologue, mais possédant des fréquences de consultation non nulle en hôpital pour le gynécologue, etc.

Dans un deuxième temps, pour l'ensemble des anomalies potentielles détectées, sauf flagrantes et « sûres », nous sommes retournés interroger les dossiers d'enquêtes, identifiant par identifiant. Les erreurs, discordances, valeurs aberrantes possibles ont alors pu être examinées plus sûrement, avec davantage d'informations. Le cas échéant, la base de données a été corrigée.

Dans la suite, pour plus de facilité, nous classons les types d'anomalies rencontrées en 3 catégories : les valeurs aberrantes possibles, les valeurs incohérentes, et les erreurs de classement. Beaucoup d'anomalies n'ont pu être suspectées que par l'examen de variables intermédiaires, construites sur les variables du questionnaire SIRS 2010. La construction précise n'est pas donnée ici pour toutes les variables intermédiaires, le code les concernant est répertorié dans un autre document (code STATA®). Enfin, la liste des identifiants avec le type d'erreur associée, et le type de correction proposée est également disponible dans un autre document Excel®.

Valeurs aberrantes potentielles.

Parmi les questions examinées, 5 individus ont été identifiés comme présentant potentiellement des valeurs aberrantes par rapport aux autres personnes.

2 concernait la fréquence de consultation du généraliste en cabinet (50 et 70 fois dans les 12 derniers mois). Après examen des dossiers, il s'est avéré que les deux présentaient des pathologies chroniques associées lourdes (HTA, diabète, dépression, troubles du sommeil), et que l'une des deux personnes a en sus présenté une pathologie ayant nécessité une intervention chirurgicale, avec complications post opératoires. Ces valeurs n'ont donc pas été corrigées.

Une personne présentait une fréquence de consultation de spécialiste en cabinet et du psychiatre de 101 et 99 fois l'année. Les vérifications ont montré une personne avec dépression a priori chronique et antécédent de rechute. Aucune correction n'a été apportée.

Une autre personne a rapporté une fréquence de consultation en hôpital public, secteur public (autre qu'une hospitalisation, normalement) de 100 dans les 12 derniers mois, avec consultation de 90 fois d'un spécialiste. Cette personne a déclaré un cancer broncho-pulmonaire et du pancréas. Les valeurs n'ont pas été modifiées.

Enfin, la dernière personne concernée avait rapporté une fréquence de consultation du dermatologue de 60 fois l'année ; elle a déclaré être traitée pour psoriasis. La valeur n'a pas été changée.

Les valeurs incohérentes dont nous traitons ici ne concernent que des incohérences internes au tableau F16 des consultations de spécialistes. Elles sont dues à des discordances entre les valeurs des variables filtres (de la forme : « avez-vous consulté un spécialiste X ? », réponse oui/non – 0/1), et les valeurs données par sous catégories (de la forme : consultation du spécialiste X en cabinet de ville, la réponse possible étant un nombre de consultation, ou une valeur manquante).

Ces valeurs incohérentes ont été détectées par la création de variables sommatoires intermédiaires, qui correspondent globalement à la sommation soit en ligne, soit en colonne du tableau F16. Par exemple, afin de connaître la fréquence de consultation du cardiologue par la personne Y dans les 12 derniers mois, quelque soit le lieu de consultation (sommation en ligne), ou bien, la fréquence de consultation d'un type de lieu de soins particuliers (exemple : le dispensaire) par une personne Y dans les 12 derniers mois (sommation en colonne).

De fait, nous avons relevé 128 incohérences, certaines concernant la même personne, suggérant une attitude de remplissage particulière de la part de l'enquêteur dans la plupart des cas.

Nous ne détaillerons pas ici en détails l'ensemble des incohérences et de leurs corrections, mais allons présenter les cas de figures type rencontrés, et le principe de leur correction.

- La variable filtre annonce une consultation, mais aucune fréquence n'est donnée (valeurs manquantes) : dans ce cas, on considère qu'il y a eu consultation, mais que les valeurs de fréquences sont effectivement manquantes.
- La variable filtre n'annonce pas aucune consultation (valeur manquante ou à 1), mais les fréquences sont à 0 : on considère que la mise à 0 des fréquences est un acte positif, et l'on modifie donc la variable filtre en la passant à 0.
- La variable filtre n'annonce pas de consultation (valeur manquante, ou à 0), mais les fréquences associées sont non nulles et non manquantes : la variable filtre est mise à 1.

Le questionnaire SIRS 2010 contient un tableau (question F16) de consultation des spécialistes, pré identifiés pour certains (gynécologue, cardiologue, psychiatre, dermatologue, rhumatologue, diabétologue et ORL), avec la possibilité de rapporter 4 autres spécialistes non listés. Par ailleurs, d'autres questions font références à certains spécialistes, ou au généraliste, et il est possible que certaines réponses aient donc été « mal classées », c'est-à-dire, données à la « mauvaise question ». C'est notamment le cas avec la médecine du travail, de l'acupuncteur / ostéopathe et du généraliste.

10 personnes ont ainsi déclaré un spécialiste diversement relié aux problèmes d'articulation, d'os, de calcium, codé par ailleurs « 65 », c'est-à-dire rhumatologue, mais qui n'ont pas été cochés comme rhumatologue en tant que spécialiste pré codé (ils ont été inscrits parmi les 4 autres spécialistes possibles). Tous ces individus ont donc été recodés comme ayant consulté un rhumatologue, et les lignes correspondant à « l'autre spécialiste » avec le code 65 du rhumatologue ont été mises à 0.

1 personne a déclaré avoir consulté un « chirurgien gynécologue », sans que le gynécologue soit coché comme spécialiste pré codé. Il a été décidé de laisser les valeurs telles quelles.

33 personnes (respectivement 7 et 36) ont déclaré dans le tableau F16 un « autre spécialiste », soit un acupuncteur, soit un ostéopathe, sans avoir précisé à la question suivante F18_1 (« avez-vous consulté un acupuncteur ou un ostéopathe ? » Réponse : oui/non/déjà mentionné dans le tableau F16) qu'ils l'avaient déjà inclus dans leur réponse.

Ces valeurs ont été laissées dans le tableau F16, mais la valeur de la question F18_1 a été mise à 3 (« déjà mentionné dans le tableau »).

1 personne a déclaré avoir consulté un généraliste parmi les « autres spécialistes ». Les valeurs de cette ligne ont été mises à 0.

1 personne a déclaré avoir consulté un médecin du travail parmi les « autres spécialistes ». Les valeurs de cette ligne ont été mises à 0.

Utilisation de l'algorithme PAM et évaluation de la robustesse des groupes

Nous avons utilisé les données SIRS corrigées telles que présentées dans les sections précédentes, et leur avons appliqué des techniques de *clustering* associées à une approche par la robustesse des groupes. Le *clustering* a été mené sur les données de recours aux soins. L'analyse principale a été menée avec l'algorithme PAM (*Partitioning around medoïds*), qui est une variante de l'algorithme des k-moyennes. Cet algorithme a été choisi pour sa conception qui prend en compte tant les données catégorielles, continues que mixtes. La

distance initiale est la distance Euclidienne. Nous ne sommes donc pas passés par l'approche traditionnelle qui consiste à effectuer une analyse en correspondances multiples sur des données catégorielles pour les transformer en variables continues, puis à appliquer une technique de *clustering* hiérarchique, le plus souvent.

Le nombre optimal de classes a été cherché en recourant à l'approche par la robustesse des groupes. Nous avons cherché ce nombre entre 2 et 6, prévoyant d'augmenter notre fourchette selon l'allure de la courbe de robustesses obtenue en première intention. Le calcul étant intensif, nous avons d'abord estimé les robustesses sur un échantillon de la population, tiré aléatoirement, plutôt que sur l'ensemble de la population (500 personnes). Le partitionnement étant une opération qui fournit des classes qualitatives avant tout, et le nombre de classes recherché demeurant faible, l'étude d'un sous-ensemble de cette taille ne pose pas problème. Dans un rapport technique non publié, nous avions étudié l'influence de la taille de l'échantillon sur l'estimation de la robustesse des groupes, à partir de la population de SIRS, les tailles d'échantillons variant de 250 personnes à 1500 personnes. Aucune différence évidente n'a été constatée. En revanche, les ressources et temps nécessaires aux calculs sur des échantillons plus grands semblent croître exponentiellement, justifiant donc l'utilisation d'un sous-ensemble de taille raisonnable. Une fois le nombre optimal connu, les algorithmes de *clustering* ont été appliqués sur l'ensemble des données.

Des analyses de sensibilité ont été menées en faisant varier la mesure de similarité : distance Euclidienne, de Manhattan, mesure de Gower, mais également en utilisant d'autres méthodes de *clustering* : k-moyennes, k-moyennes dans sa version logique floue (« *fuzzy c-means* »), *clustering* hiérarchique descendant, et enfin par cartes auto-organisatrices (SOM, *self-organizing map,* en l'occurrence, la version SOTA : *self-organizing tree algorithm*).

Le but de cette première étape était donc double : d'une part identifier le nombre de classes présentes dans les données en s'assurant de leur robustesse sous un certain nombre de transformations, et d'autre part caractériser ces classes en termes de recours aux soins. Une fois cette tâche effectuée, nous nous sommes intéressés aux facteurs associés – ou « explicatifs » - de ces classes.

Les analyses ont été conduites sous R 2.1 64 bits, avec le package clusterCons.

Analyses multinomiales complémentaires des facteurs associés aux profils

En toute rigueur, les techniques d'identification de groupes produisent des résultats dont on peut discuter les avantages et les limites, dont on connaît les domaines de validité. Il n'est pas nécessaire de les appuyer sur des techniques statistiques systématiquement. Dans notre travail, nous avons tenu à utiliser des analyses statistiques à deux titres : d'une part pour permettre plus facilement l'acceptation des techniques de *clustering* par les épidémiologistes, en les rattachant à des approches dont ils sont plus familiers, ce qui permet également de souligner les avantages et défauts des deux types d'approches, d'autre part pour non pas neutraliser les variations intrinsèques des caractéristiques de chaque classe, mais pour contrôler partiellement ce qui peut être dû à des fluctuations d'échantillonnage, des effets de taille d'échantillon : les différentes classes ne sont pas nécessairement peuplées également, et la taille des populations de chaque groupe permet, selon les variables, une plus ou moins grande précision. L'utilisation de tests statistiques usuels de comparaison de moyennes ou de proportions avait essentiellement pour but de montrer que l'identification de classes bien séparées se répercute logiquement, si la puissance le permet, en différences « significatives » selon la plupart sinon toutes les caractéristiques.

De la sorte, nous avons construits des modèles multinomiaux, avec en variable dépendante l'appartenance à une classe ou une autre, et en variables explicatives les variables décrites dans les sections précédentes, par sous-ensembles. Etant dans un contexte observationnel, nous ne spécifions pas de valeur particulière du « petit p » comme étant spécifiquement significatif – étant entendu que plus la valeur est faible, plus la confiance en une différence qui ne soit pas attribuable au hasard est importante.

Les analyses ont été conduites sous Stata 11.

Une autre façon de réduire la complexité : application des techniques d'apprentissage de variétés à un cas d'utilisation du système de santé à des fins judiciaires

Position du problème

Ainsi que nous avons pu en faire état dans les sections précédentes, l'autorité judiciaire, un magistrat, peut demander à un médecin, par réquisition, que soit estimé l'âge d'une personne dont on ne peut établir avec une certitude satisfaisante la date de naissance ou l'âge civil, chronologique. En pratique, si tout médecin peut être réquisitionné, cette tâche est plutôt dévolue aux médecins légistes, éventuellement aux pédiatres et radiologues.

Il s'agit donc d'une utilisation du système de santé à des fins judiciaires. Comme nous l'avons souligné, il n'existe pas de normes en la matière, que ce soit dans les techniques mobilisées pour répondre à la question, ou en termes de formulation de la réponse. Parmi les tendances actuelles, nous assistons à la multiplication des modalités d'estimation d'âge, et les sources se multipliant, à l'intégration de ces différentes sources de données, dans une logique d'améliorer la précision des résultats – ou de diminution de l'incertitude. La question que nous nous sommes posée a été de savoir si une telle intégration pouvait en effet apporter une amélioration des estimations. Simultanément, nous nous sommes interrogés sur le gain informationnel d'une telle intégration : est-ce qu'informationnellement, considérer les sources classiques d'estimation d'âge apportait un véritable gain ?

Données utilisées

Nous avons exploité des données qui ont fait l'objet d'une publication antérieure, en 2010.[166] Ces données ont été recueillies entre le 1r janvier 2007 et le 31 décembre 2007. Les adolescents migrants inclus étaient inclus prospectivement et consécutivement lors de leur présentation par les autorités judiciaires dans le service de médecine légale de l'hôpital Jean Verdier (AP-HP, Bondy, France). Les demandes d'estimation d'âge émanaient du procureur de la République, dans le cadre soit d'une demande d'asile, soit d'une garde à vue pour infraction pénale, quelle qu'elle soit. Les examens étaient effectués avec l'accord de la personne concernée, en présence d'un interprète si besoin (ou le cas échéant, avec l'assistance d'un service d'interprétariat par téléphone).

Les données recueillies concernaient des informations cliniques, dentaires et radiologiques.

Les données cliniques consistaient en le sexe, l'origine géographique et l'âge allégué. Les diverses origines géographiques ont été catégorisées en 8 classes : Afrique, Asie, Europe de l'Ouest, Europe de l'Est, Moyen-Orient, Océanie, Amériques du sud et centrale. L'âge allégué est l'âge déclaré par l'adolescent.

Les données dentaires consistaient en la présence ou l'absence de chaque deuxième ou troisième molaire (dent de sagesse), soit 8 variables binaires au total (présence : O/N).

Les données radiologiques consistaient en l'appréciation par un médecin légiste et par un radiologue, séparément, du caractère fusionné ou non des épiphyses distales du radius et de l'ulna de la main non dominante (soit en général, la main gauche). L'âge radiologique était

l'âge estimé à partir de l'atlas de Greulich et Pyle, qui attribue à chaque aspect radiologique des clichés de mains gauches un âge chronologique, par sexe.
Enfin, il était rendu un âge estimé, par le médecin légiste, sur la base de toutes les informations disponibles et mentionnées ci-dessus.
Au total, 14 variables ont été utilisées.

Analyses et utilisation d'algorithmes NLDR

Nous avons procédé à la description de toutes les variables mentionnées, et notamment calculé les âges médians avec leurs 10^{e} et 90^{e} percentiles en fonction de la présence de chaque deuxième ou troisième molaire et en fonction du sexe. Les comparaisons ont été calculées à l'aide du test de Kruskal-Wallis. Les corrélations entre les différents âges (allégué, radiologique et estimé) ont été calculés (coefficient de Spearman).
Les données étant catégorielles et afin également de comparer nos résultats à ceux obtenus avec des techniques classiques, nous avons procédé à une transformation par analyse en correspondances multiples (ACM), en conservant toutes les dimensions initiales.
Afin d'estimer le contenu informationnel de notre ensemble de données, nous avons utilisé une technique d'estimation de la dimension intrinsèque des données (« *correlation dimension* »), appliquée aux données transformées par l'ACM. La dimension étant estimée, nous avons ensuite utilisé des techniques non linéaires de réduction de la dimension. Le choix étant relativement vaste, nous avons retenu un algorithme non conservateur bien connu (ISOMAP) et un algorithme conservateur, un autoencodeur. Par conservateur nous entendons que toutes les observations sont conservées et réarrangées par l'algorithme. A contrario, un algorithme non conservateur comme ISOMAP, lors de l'étape qui consiste à déterminer le graphe des connexités, peut considérer qu'un certain nombre d'observations ne sont pas rattachables à d'autres : elles sont donc écartées (comme des outliers). Nous avons travaillé sur des données complètes ; aucune imputation n'a été faite.

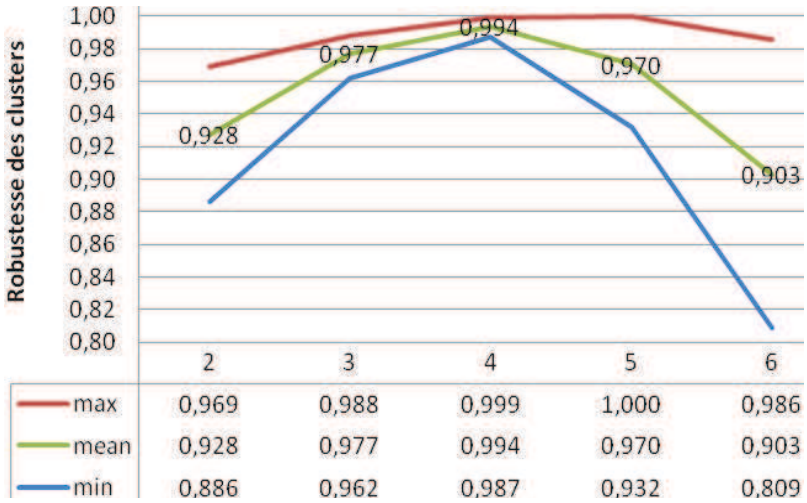
Résultats

Présentation de la typologie de recours aux soins : résultats généraux

En appliquant l'algorithme PAM couplé à l'approche par la robustesse des classes, nous avons trouvé un nombre optimal de groupes égal à 4, sans grande ambiguïté ; 4 profils de recours aux soins ont donc été isolés dans nos données, contenant les 3000 personnes de la vague 2010 de la cohorte SIRS.

Ces 4 groupes ont montré une excellente stabilité à l'égard du ré-échantillonnage des personnes et des permutations aléatoires de ceux-ci dans les sous-échantillons (voir Figure 6 : Robustesse des clusters vs nombre de clusters).

Figure 6 : Robustesse des clusters vs nombre de clusters



Pour chaque nombre de classes à identifier, la robustesse de chaque classe est calculée : pour 3 clusters, la robustesse des 3 clusters est calculée, pour 4 clusters, 4 sont calculées etc. Afin de comparer les résultats issus, par construction, de nombres de classes différents, on calcule la robustesse moyenne, la robustesse maximale et la robustesse minimale. Le nombre de clusters présentant un pic de robustesse pour les 3 valeurs simultanément est considéré comme le nombre optimal (dispersion minimale autour d'une robustesse moyenne maximale).

Comme nous le verrons en abordant la description détaillée de la typologie, les 4 profils sont relativement bien équilibré en effectifs (30.0%-21.0%-25.7%-23.3%), et présentent une certaine polarité, si l'on rapproche en couple des profils qualitativement proches, mais quantitativement distinguables, et de caractéristiques explicatives sous-jacentes qualitativement différenciées.

Les résultats sont donnés en accord avec la méthode générale annoncée : d'abord la description des types en termes de variables de recours aux soins, puis ensuite en termes de variables « explicatives ». La description en termes de variables « explicatives » est dans un premier temps présentée en bi-variée (type vs variable d'intérêt), puis par des modèles multinomiaux, par groupes de 3 variables (modèle données démographiques, modèle données socio-économiques…etc.).
Les résultats apparaissent tous un minimum contrastés, des différences non significatives étant en général cependant sous-tendues par une tendance intéressante à relever, étant entendu que la « significativité » n'est pas notre but ici.

Nous présentons ci-après la description des 4 types de recours aux soins découverts dans les données SIRS ; *la p-value* est précisée uniquement si elle est au-dessus du seuil de 0.001.
La description est donnée selon les 6 groupes :

- Recours aux soins primaires
- Recours à une médecine alternative
- Recours à des soins spécialisés, d'accès direct
- Recours à des soins spécialisés, d'accès indirect
- Lieux de recours aux soins
- Recours aux soins pour motif d'urgence

Et est présentée dans les
Tableau 3 : Quatre types de profil de recours aux soins dans la région d'Ile de France, 2010. Un tiret «-» indique un effectif nul.

Tableau 3 : Quatre types de profil de recours aux soins dans la région d'Ile de France, 2010.

	Type 1 (30.0%)	Type 2 (21.0%)	Type 3 (25.7%)	Type 4 (23.3%)	All types N=3006	p
Soins primaires						
Date de la dernière consultation dentiste						<0.001
< 2 ans	84.3	68.1	72.9	87.0	78.2	
2-3 ans	4.0	9.8	9.2	4.3	6.8	
> 3 ans	7.8	16.8	13.1	6.2	10.9	
jamais	3.9	5.3	4.8	2.5	4.1	
A un médecin généraliste référent						<0.001
	96.0	71.5	93.2	90.6	88.1	
Fréquence de consultation du généraliste						<0.001
0	2.4	60.4	-	6.7	16.9	
1	9.0	39.6	-	23.7	17.7	
2	17.0	-	31.8	23.9	18.1	
3-5	38.6	-	47.9	34.6	30.5	
6 +	33.0	-	20.3	11.1	16.8	
Fréquence de consultation d'un SAD						<0.001
0	49.3	87.7	93.6	-	57.8	
1	22.3	12.2	6.4	40.2	20.2	
2	11.4	0.1	-	25.1	9.1	
3-5	9.4	-	-	16.3	6.5	
6 +	7.6	-	-	18.3	6.4	
A effectué un check-up en centre de sécurité sociale						0.127
	7.0	5.1	5.7	3.6	5.4	
Fréquence des avis médicaux demandés aux proches						0.0057
0	78.1	85.3	84.4	78.3	81.4	
1-2	14.2	10.7	12.3	14.7	13.0	
3-10	6.1	3.8	3.4	6.3	4.9	
11 +	1.6	0.2	-	0.7	0.6	
Médecines parallèles et médecines alternatives						
A consulté un acupuncteur / ostéopathe						<0.001
	22.7	6.4	9.4	17.1	17.1	
A consulté un professionnel de santé non conventionnel / médecine traditionnelle						<0.001
	7.2	2.1	3.3	6.9	4.9	
Spécialistes d'accès indirect						
Fréquence de consultation d'un SAI						<0.001
0	-	81.1	78.4	68.0	54.8	
1	1.3	12.2	21.6	28.6	15.3	
2	29.3	1.8	-	3.3	9.4	
3-5	40.8	-	-	0.1	12.6	

6 +	28.7	-	-	-	0.8	
Lieu de recours aux soins						
Hôpital/Clinique publique - MG	12.3	3.1	14.1	7.6	9.4	<0.001
Hôpital/Clinique publique - spécialiste	42.5	5.5	7.2	19.9	19.6	<0.001
Hôpital/Clinique privés - MG	14.2	1.8	10.5	6.3	8.4	<0.001
Hôpital/Clinique privés - spécialiste	21.1	1.5	3.5	9.3	9.3	<0.001
Ambulatoire – MG	91.0	34.9	90.9	88.6	76.8	<0.001
Ambulatoire – spécialiste	83.0	22.3	16.3	88.8	53.4	<0.001
Soins d'urgence et non programmés						
Fréquences des consultations à domicile						0.002
0	-	81.1	78.4	68.0	54.8	
1	1.3	12.2	21.6	28.6	15.3	
2	29.3	1.8	-	3.3	9.4	
3-5	40.7	4.9	-	-	19.7	
6 +	28.7	-	-	-	0.8	
Fréquence des consultations dans un SAU						<0.001
0	77.5	89.8	80.7	83.0	82.6	
1	15.9	9.7	15.7	13.0	13.7	
2	4.0	0.5	2.2	2.1	2.3	
3-5	2.2	-	1.2	1.6	1.3	
6 +	0.4	-	0.2	0.3	0.2	

MG : médecin généraliste. SAD : spécialiste d'accès direct. SAI : spécialiste d'accès indirect. SAU : service d'accueil des urgences. Les résultats sont exprimés en pourcentages.

4 profils de recours

Profil 1 : le recours intensif, partout, de toutes les façons possibles et le plus souvent

Quel type de recours aux soins ?

- Soins primaires : le type 1 est le premier consommateur, sinon le 2^{e} immédiat, de soins primaires, tant en qualité qu'en quantité et fréquence. Il est celui qui présente le plus souvent une visite récente chez le dentiste avec le type 4, à déclarer un généraliste de référence (96%) et à le fréquenter (33% des consultants le consultent 6 fois et plus par an). Il est également le premier à consulter son entourage pour avis médical (7.7% ont « consulté » au moins 3 fois dans l'année). Il est cependant en recul (2^{e}) concernant la consultation (50.7%) et la fréquence de consultation du spécialiste d'accès direct (44% des consultants consultent une seule fois), notamment vis-à-vis du type 4, qui sont 100% à consulter un tel professionnel.

- Médecines parallèles et médecines alternatives : de la même façon, il est celui qui requiert le plus souvent à un ostéopathe, un acupuncteur (2.7%), ou à toute médecine alternative (7.2%).
- Spécialistes d'accès indirect : si le type 1 requiert beaucoup (plus de 50%) au spécialiste d'accès direct, mais moins que le type 4 (100%), il présente cependant un recours uniforme au spécialiste d'accès indirect (100%), et le consulte extensivement.
- Lieux de recours aux soins : là encore, tous les lieux sont utilisés, avec une prédilection pour le public en ce qui concerne le spécialiste (42.5%), et le cabinet (91%) ou le privé (14.2%) pour le généraliste. Il est le premier à recourir au secteur privé, hors cabinet de ville. Il est celui qui utilise le plus l'hôpital public pour la consultation de spécialistes.
- Soins d'urgence et non programmés : le type 1 est enfin celui qui recourt aux urgences le plus souvent, globalement, notamment à domicile, puisque cela concerne 100% d'entre eux. Près de 70% des consultants consultent par ailleurs au moins 3 fois. Quant au SAU, il est également le premier utilisateur, avec 22.5% de consultants.

En résumé, le type 1 est un grand consommateur de soins, et bénéficie de toutes les opportunités offertes, en général extensivement.

Profil 2 : le recours le moins fréquent, et le moins diversifié

Quel type de recours aux soins ?

Soins primaires : le type 2 se partage avec le type 3 la 3^{e} et la 4^{e} place en termes de recours aux soins primaires. Concernant le dentiste, 22.1% rapportent une visite datant d'au moins 3 ans, sinon aucune. Il est celui qui déclare le moins de généraliste référent, de loin (près de 30% n'en ont pas). Et lorsqu'ils consultent le généraliste, 100% des type 2 ne le consultent qu'une seule fois. Il consulte rarement le spécialiste d'accès direct (12.3%), et le cas échéant, presque qu'une seule fois (99%, jamais plus de 2). Enfin, que ce soit le bilan de sécurité sociale, ou la demande d'avis à l'entourage, il est avant-dernier.

- Médecines parallèles et médecines alternatives : le type 2 est celui qui a le moins recours à ce type de soins (6.4% pour l'ostéopathe ou l'acupuncteur).
- Spécialistes d'accès indirect : En termes de consultants, il est aussi bon dernier (18.9%, lorsque 100% des type 1 en consultent un), mais parmi ceux-ci, il est le

deuxième en fréquence de consultation, puisque 40.8% d'entre eux consultent 3 à 5 fois l'année.

- Lieux de recours aux soins : globalement, quel que soit le lieu, le taux de fréquentation du type 2 est au moins de 2 à 3 fois inférieur à celui du type 1, faisant de lui également le dernier à consulter. La différence est encore plus marquée lorsqu'il s'agit du secteur privé, hors cabinet. Il consulte néanmoins davantage en cabinet que partout ailleurs, à l'instar de tous les autres types.
- Soins d'urgence et non programmés : le type 2 a très peu recours aux urgences, que ce soit à domicile ou au SAU : 18.9% à domicile, pour une consultation souvent unique, 10.2% au SAU, pour 92.6% des consultants ne consultant qu'une seule fois, au maximum 2 fois en un an.

En résumé, le type 2 représente essentiellement un usager qui va requérir le système de soins en fréquences et en diversité.

Profil 3 : le recours au généraliste de ville ou aux urgences non programmées

Quel type de recours aux soins ?

- Soins primaires : le type 3 est l'avant-dernier à consulter fréquemment le dentiste pour contrôle (4.8% ne l'ont jamais fait). Il est le 2^{e} à déclarer avoir un généraliste référent, et parmi les consultants, il est celui qui consulte le plus fréquemment avec le type 1 (au moins 2 fois, et 20.3% consultent plus de 6 fois). Concernant les spécialistes d'accès direct, il est le dernier à consulter (6.4%), et parmi les consultants, ceux-ci les consultent au maximum 1 fois. Il est le dernier à solliciter l'entourage pour avis médicaux, et le 2^{e} à recourir aux centres de sécurité sociale.
- Médecines parallèles et médecines alternatives : Il est le 2^{e} à moins recourir à la médecine alternative ou aux médecines parallèles.
- Spécialistes d'accès indirect : il est le 2^{e} type à consulter le moins ce genre de spécialiste (21.6%), et parmi les consultants, le dernier en termes de fréquence de consultation : 100% des consultants ne consultent qu'une seule fois.
- Lieux de recours aux soins : il est le premier à recourir au secteur public pour le généraliste. Concernant le privé, il est le 2^{e} à l'utiliser, cabinet inclus, pour le généraliste ; il est avant-dernier ou dernier pour le spécialiste (dernier pour la consultation en cabinet de ville).

- Soins d'urgence et non programmés : En revanche, il a recours assez facilement aux urgences, que ce soit à domicile ou au SAU (respectivement 21.6% et 19.3%).

En résumé, le type 3 est celui qui a un recours plutôt binaire : soit le généraliste de ville qu'il consulte fréquemment, soit les urgences ou soins non programmés, qu'il ne requiert pas forcément le plus, mais souvent s'il consulte.

Profil 4 : le recours aux spécialistes de cabinet et aux soins alternatifs paramédicaux

Quel type de recours aux soins ?

- Soins primaires : le type 4 est celui qui fréquente le plus souvent le dentiste pour simple contrôle (87% pour une visite de moins de 2 ans). Il est le troisième à déclarer avoir un généraliste, en tenant compte que les différences sont minimes entre types, à l'exception du type 2 ; parmi les consultants, il est cependant aussi l'avant-dernier. A l'inverse, il est le premier à consulter les spécialistes d'accès direct, puisque 100% des type 4 le font ; parmi les consultants, il est également celui qui consulte le plus souvent. Il est également le deuxième à recourir à un avis de l'entourage. Il est le dernier à recourir aux bilans de santé en centre de sécurité sociale.
- Médecines parallèles et médecines alternatives : Il est le 2^{e} à recourir à la médecine alternative (6.9%), à l'ostéopathe ou à l'acupuncteur (17.1%).
- Spécialistes d'accès indirect : le type 4 est le deuxième à rapporter consulter les spécialistes d'accès indirect (32%), mais relativement peu fréquemment (99.5% ne consulteront pas plus de 2 fois dans l'année).
- Lieux de recours aux soins : il a le même rang de consultation du spécialiste (2^{e}, avec 29.9% de consultants dans le public) et du généraliste (il consulte peu) tant pour le public que pour le privé. Cependant, pour la consultation de cabinet, il est le premier à recourir au spécialiste (88.8%).
- Soins d'urgence et non programmés : le type 4 consulte relativement peu les urgences, que ce soit à domicile ou au SAU. Néanmoins, pour le SAU, parmi les consultants, il est le deuxième en fréquences de consultation.

En résumé, le type 4 caractérise un usager qui consomme de façon assez importante, et plus spécifiquement le spécialiste (notamment d'accès direct), en cabinet, et les médecines alternatives.

Facteurs associés aux profils de recours

Comme annoncé, les facteurs associés à la typologie sont présentés en deux temps : d'abord selon des analyses bi-variées (type vs variable d'intérêt), puis selon des modèles multinomiaux, par groupes de variables faisant sens a priori. Il n'y a pas de sélection de variable pour inclusion dans les modèles multinomiaux : elles sont toutes incluses au sein de leur groupe d'appartenance.

Le tableau d'analyses bi-variées (Tableau 4 : Caractéristiques des personnes relevant des quatre types de profils de recours aux soins en région d'Ile de France, 2010.) se lit en fréquences relatives, avec sommation à 100% en colonne. Par exemple, le type 1 se répartit dans les 5 classes d'âge de la façon suivante : 11.6% dans les 18-29ans, 22.8% dans les 30-44, 25.5% dans les 45-59, 24.2% dans les 60-74 et 15.9% dans les 75 et +. La p value reportée est celle issue en général d'un chi-2 (il n'y a pas de classes vides pour ces variables).

Le tableau d'analyses multivariées (Tableau 5 : Caractéristiques associées aux types de profils de recours aux soins : modèles de régression multinomiale (type 4 pris en référence). Région d'Ile de France, 2010.) rapporte les odd-ratios OR, et doivent se comprendre de la façon suivante. Le type de référence est le type 4 : on compare donc les types 1, 2 ou 3 au groupe 4. La deuxième référence de comparaison dépend de la variable d'intérêt. Prenons l'exemple de la classe d'âge : l'OR est de 11.14, et compare la probabilité de faire partie de la classe 75 ans et +, dans le type 1, par rapport à faire partie de la classe d'âge 18-29 ans, dans le type 4. L'intervalle de confiance à 95% est donné entre parenthèse (ici : 5.67-21.89).

Tableau 4 : Caractéristiques des personnes relevant des quatre types de profils de recours aux soins en région d'Ile de France, 2010.

	Type 1 (30.0%)	Type 2 (21.0%)	Type 3 (25.7%)	Type 4 (23.3%)	Tous types N=3006	p
Modèle 1 : démographique						
Age (ans)						<0.001
18-29	11.6	32.1	21.2	23.7	21.8	
30-44	22.8	32.0	35.0	39.5	31.9	
45-59	25.5	23.4	20.2	23.7	23.3	
60-74	24.2	9.3	14.1	9.6	14.7	
75 +	15.9	3.3	9.5	3.5	8.3	
Genre						<0.001
Masculin	42.6	64.3	59.6	21.0	47.0	
Féminin	57.4	35.7	40.4	79.0	53.0	
Origines						0.006
Français, né de parents Français	72.2	63.3	61.1	69.1	66.6	
Français, né d'au moins un parent étranger	18.3	22.0	23.2	20.1	20.8	
Etranger	9.5	14.7	15.7	10.8	12.6	
Modèle 2 : statut socioéconomique						
Niveau d'éducation						<0.001
3e cycle	59.3	54.5	48.7	63.1	56.5	
Secondaire	32.5	38.0	42.3	32.2	36.2	
Primaire ou aucun	8.2	7.5	9.0	4.7	7.3	
Activité professionnelle						<0.001
Employé	47.9	61.6	55.5	63.1	56.7	
Chômeur	5.7	8.6	9.3	7.1	7.6	
Inactif	46.4	29.8	35.2	29.8	35.7	
Revenu par UC (quintiles)						
1r	16.5	22.8	24.6	18.9	20.6	<0.001
2e	14.2	22.4	23.0	18.5	19.3	
3e	22.3	21.8	19.7	23.2	21.7	
4e	20.0	18.2	15.6	17.8	18.0	
5e	27.0	14.8	17.1	21.6	20.4	
Couverture maladie						<0.001
Sécurité sociale + complémentaire	94.1	78.5	85.3	89.6	87.1	
CMU	1.7	4.1	2.1	3.0	2.6	
Sécurité sociale seule	3.6	16.6	12.4	6.8	9.7	
Aucune	0.6	0.8	0.2	0.6	0.6	
Modèle 3 : comportement vis-à-vis de la médecine et de la santé						
Comportement général envers le fait de consulter						<0.001
Si ne peut faire	33.0	68.7	38.7	43.2	45.4	

autrement						
Dès que ne se sent pas bien						
	67.0	31.3	61.3	56.8	54.6	
A un proche ou ami souffrant d'une maladie grave						0.002
Oui	51.5	42.7	41.8	53.0	47.3	
Non	48.5	57.3	58.2	47	52.3	
A des professionnels de santé parmi ses proches						0.215
Non	55.7	59.3	61.8	55.8	58.1	
Oui	44.3	40.7	38.2	44.2	44.2	
Modèle 4 : intégration sociale						
Sentiment d'isolement						0.251
Très entouré	29.7	35.2	32.0	34.5	32.7	
Plutôt entouré	55.6	54.1	53.7	52.5	54.1	
Plutôt isolé	12.5	9.6	12.1	12.3	11.6	
Très isolé	2.2	0.1	0.2	0.7	1.6	
Niveau de soutien social						<0.001
Élevé	81.9	90.7	87.4	91.3	87.6	
Moyen	13.8	6.5	9.1	6.2	9.1	
Bas	4.3	2.8	3.5	2.5	3.3	
Fréquence des contacts sociaux (quartiles)						0.054
1^{r}	25.0	23.3	22.5	18.7	22.5	
2^{e}	24.2	24.9	27.1	21.1	24.4	
3^{e}	26.3	21.6	22.0	30.4	25.1	
4^{e}	24.5	30.2	28.4	29.7	28.0	
Modèle 5 : état santé						
État de santé perçu						<0.001
Bon	62.3	92.2	77.4	82.4	78.0	
Médiocre	28.5	6.8	20.0	15.2	18.0	
Mauvais	9.2	1.0	2.6	2.4	4.0	
Problème de santé chronique ou durable						<0.001
Non	40.2	86.4	64.4	72.2	64.8	
Oui	59.8	13.6	35.6	27.8	35.2	
Limitation fonctionnelle durable des activités quotidiennes						<0.001
Non	63.1	94.4	81.1	86.5	80.6	
Oui	36.9	5.6	18.9	13.5	19.4	

Tous les résultats sont exprimés en pourcentages. UC : unité de consommation.

Tableau 5 : Caractéristiques associées aux types de profils de recours aux soins : modèles de régression multinomiale (type 4 pris en référence). Région d'Ile de France, 2010.

	Type 1 OR [IC 95%]	Type 2 OR [IC 95%]	Type 3 OR [IC 95%]	p*
Modèle 1 : démographique				
Age (ans)				<0.001
18-29	Ref	Ref	Ref	
30-44	1.18 [0.72-1.93]	0.60 [0.37-0.96]	0.99 [0.67-1.45]	
45-59	2.24 [1.35-3.73]	0.77 [0.48-1.31]	1.01 [0.67-1.55]	
60-74	5.44 [3.28-9.03]	0.80 [0.48-1.31]	1.87 [1.17-2.97]	
75 +	11.14 [5.67-21.89]	1.00 [0.47-2.12]	4.45 [2.49-7.94]	
Genre				<0.001
Masculin	Ref	Ref	Ref	
Féminin	0.32 [0.22-0.46]	0.14 [0.11-0.20]	0.17 [0.11-0.24]	
Origines				0.0375
Français, né de parents Français	Ref	Ref	Ref	
Français, né d'au moins un parent étranger	1.10 [0.79-1.54]	1.32 [0.89-1.95]	1.54 [1.11-2.17]	
Étranger	1.19 [0.72-1.99]	1.67 [1.06-2.61]	1.95 [1.30-2.93]	
Modèle 2 : statut socioéconomique				
Niveau d'éducation				0.0119
3e cycle	Ref	Ref	Ref	
Secondaire	1.23 [0.89-1.69]	1.25 [0.91-1.71]	1.61 [1.24-2.10]	
Primaire ou aucun	1.94 [1.09-3.48]	1.67 [0.92-3.01]	2.26 [1.39-3.69]	
Activité professionnelle				<0.001
Employé	Ref	Ref	Ref	
Chômeur	1.40 [0.65-3.00]	0.87 [0.43-1.76]	1.20 [0.61-3.36]	
Inactif	2.08 [1.59-2.71]	1.01 [0.77-1.34]	1.25 [0.98-1.61]	
Revenu par UC (quintiles)				<0.001
1r	Ref	Ref	Ref	
2e	0.96 [0.61-1.51]	1.11 [0.70-1.76]	1.05 [0.68-1.61]	
3e	1.39 [0.87-2.20]	0.92 [0.53-1.60]	0.82 [0.51-1.29]	
4e	1.63 [0.96-2.77]	1.11 [0.62-1.99]	0.91 [0.57-1.47]	
5e	1.80 [1.20-2.70]	0.79 [0.51-1.21]	0.90 [0.56-1.43]	
Couverture maladie				0.0058
Sécurité sociale + complémentaire	Ref	Ref	Ref	
CMU	0.55 [0.24-1.22]	1.45 [0.69-3.05]	0.59 [0.26-1.34]	
Sécurité sociale seule	0.63 [0.29-1.37]	2.72 [1.39-5.35]	1.78 [0.89-3.56]	
Aucune	1.01 [0.19-5.41]	1.60 [0.41-6.32]	0.28 [0.04-2.03]	

Modèle 3 : comportement vis-à-vis de la médecine et de la santé				
Comportement général envers le fait de consulter				<0.001
Si ne peut faire autrement	Ref	Ref	Ref	
Dès que ne se sent pas bien	1.55 [1.19-2.01]	0.33 [0.23-0.47]	1.15 [0.84-1.58]	
A un proche ou ami souffrant d'une maladie grave				0.0058
Oui	Ref	Ref	Ref	
Non	0.97 [0.71-1.32]	0.62 [0.44-0.85]	0.67 [0.48-0.90]	
A des professionnels de santé parmi ses proches				0.3301
Non	0.96 [0.77-1.20]	1.19 [0.90-1.59]	1.20 [0.94-1.55]	
Oui	Ref	Ref	Ref	
Modèle 4 : intégration sociale				
Sentiment d'isolement				0.1295
Très entouré	Ref	Ref	Ref	
Plutôt entouré	1.11 [0.81-1.52]	0.94 [0.69-1.29]	1.01 [0.72-1.41]	
Plutôt isolé	0.88 [0.60-1.31]	0.68 [0.43-1.09]	0.90 [0.66-1.22]	
Très isolé	2.84 [1.08-7.50]	1.45 [0.46-4.58]	3.06 [1.06-8.84]	
Niveau de soutien social				0.0384
Élevé	Ref	Ref	Ref	
Moyen	2.32 [1.40-3.84]	1.03 [0.62-1.72]	1.43 [0.90-2.28]	
Bas	1.72 [0.84-3.56]	1.11 [0.59-2.11]	1.28 [0.66-2.49]	
Fréquence des contacts sociaux (quartiles)				0.0370
1^{r}	1.42 [0.96-2.13]	1.28 [0.87-1.90]	1.21 [0.80-1.82]	
2^{e}	1.29 [0.90-1.87]	1.20 [0.87-1.64]	1.32 [0.92-1.89]	
3^{e}	1.03 [0.70-1.53]	0.71 [0.51-1.00]	0.75 [0.46-1.22]	
4^{e}	Ref	Ref	Ref	
Modèle 5 : état santé				
État de santé perçu				<0.001
Bon	Ref	Ref	Ref	
Médiocre	1.40 [0.97-2.03]	0.53 [0.33-0.84]	1.18 [0.84-1.67]	
Mauvais	1.65 [0.71-3.84]	0.75 [0.24-2.36]	0.78 [0.33-1.87]	
Problème de santé chronique ou durable				<0.001
Non	Ref	Ref	Ref	
Oui	2.91 [2.27-3.71]	0.48 [0.34-0.67]	1.34 [1.02-1.76]	
Limitation fonctionnelle durable des activités quotidiennes				<0.001
Non	Ref	Ref	Ref	
Oui	2.24 [1.44-3.48]	0.56 [0.36-0.88]	1.34 [0.90-2.00]	

*: p value de tendance globale. Le type 4 est le type de reference pour tous les OR. UC : unité de consommation.

Profil 1 : le français âgé, aisé et très bien couvert, socialement isolé, inactif et en mauvaise santé

Profils sociodémographique, économique, et sanitaire :

Le type 1 est globalement le plus représenté, avec 30% de la population d'étude.

- Profil démographique : le type 1 est le plus vieux, avec 40 % ayant au moins 60 ans. Il est également le 2e plus féminin, avec plus de la moitié de ses effectifs étant des femmes. Il est également le plus souvent français né de parents français, et le moins souvent étranger, avec le type 4.
- Profil socioéconomique : En termes de niveau d'éducation, sa situation est contrastée : il est le 2e plus éduqué, globalement, mais également le 2e type présentant le plus fort taux de personnes non scolarisées ou de niveau primaire (8.2%). Il est le plus inactif (46.4%), tout en possédant le moins de chômeurs (5.7%). Le type 1 est le plus aisé économiquement, puisque 27% d'entre eux sont dans le 5e quintile, et 47% dans le 4e ou 5e quintile. Il est celui qui présente la meilleure couverture sociale, avec la plus forte propension à être couvert à 100%, mais également avec le moins de chance d'être bénéficiaire de l'AME, ou encore d'être couvert sans complémentaire santé.
- Profil rapport à la médecine : il est celui qui consulte le plus souvent son généraliste dès que quelque chose lui paraît anormal. Avec le type 4, il est celui qui connaît le plus de personnes dans son entourage à souffrir d'une pathologie grave, et à connaître le plus souvent des médecins ou professionnels de santé dans son entourage.
- Profil lien social : le sentiment d'isolement en tant que tel ne présente pas de variations claires ou significatives entre les différents types. En revanche, le type 1 annonce un soutien social plutôt moins bon, tandis qu'il accuse une fréquence des contacts sociaux la plus basse, toute proportions gardées, les différents cas de figure étant relativement équilibrés (mais tendant vers des fréquences moindres que les autres types).
- Profil santé perçue et rapportée : il est celui qui a le plus de chances de présenter une pathologie chronique (59.8%), ainsi que de rapporter une limitation dans les actes de la vie quotidienne au cours des 6 derniers mois (36.9%). Enfin, il est celui qui rapporte le plus fréquemment une santé perçue comme mauvaise ou médiocre (37.7%).

En résumé, le type 1 se caractérise par une personne plus souvent âgée, français de parents français, parmi les plus aisés, plutôt féminin et très bien couvert. Il est volontiers inactif mais non chômeur. Il est entouré de personnes plus fréquemment malades, connaît des professionnels de santé et consulte dès que quelque chose ne va pas. Il dit bénéficier d'un soutien social moyen, avec relativement peu de contacts. Il se ressent comme une personne souvent limitée dans ses actes et en mauvaise santé.

Profil 2 : le jeune homme d'origines étrangères, chômeur ou peu aisé, à la mauvaise couverture maladie mais en bonne santé

Profils sociodémographique, économique, et sanitaire :

Le type 2 représente la plus petite proportion de la population, avec néanmoins 21% de celle-ci.

- Profil démographique : le type 2 représente la population la plus jeune, avec notamment 3.2% seulement ayant 75 ans et plus, et 32.1% ayant entre 18 et 29 ans. Avec le type 3, il est plus volontiers masculin (64.3%), ainsi qu'étranger, avec 22% de français né de parent étranger et 14.7% d'étrangers.
- Profil socioéconomique : le type 2 est celui a reçu le plus souvent une éducation de niveau secondaire avec le type 3 (38%), avec peu de non scolarisés ou de niveau primaire. En termes d'occupation actuelle, il tend à être le plus fréquemment chômeur (8.6%) toujours avec le type 3, et le 2^{e} plus souvent actif (61.6%). Par rapport au type 4, ces différences n'apparaissaient cependant pas significatives. Enfin, concernant la couverture maladie, il est de loin le moins bien couvert : seulement 78.5% sont couverts par la sécurité sociale de base plus une complémentaire santé, tandis que 4.1% sont bénéficiaires de l'AME ou de la CMUc, et 16.6% d'entre eux n'ont aucune complémentaire santé.
- Profil rapport à la médecine : il est avec le type 3 celui qui connaît le moins dans son entourage de personnes atteintes de pathologie grave. Plus des deux tiers d'entre eux ne consultent leur médecin que s'ils s'y sentent contraints et ne peuvent faire autrement. 41% ne connaissent aucun professionnel de la santé dans leur entourage.
- Profil lien social : le sentiment d'isolement en soi ne permet pas ici de mettre en évidence des différences ou des tendances entre types. Le soutien social, s'il n'est exhibé là aussi aucune différence significative en comparaison du type 4, apparaît

tantôt assez pauvre chez le type 2 (2.8% disent n'avoir aucun soutien), tantôt important (90.7% annoncent un très bon soutien social). En revanche, en termes de fréquence des contacts sociaux, il est celui qui est le plus volontiers dans le 1r quartile (donc les contacts les moins fréquents), avec le type 1 : 23.3%.

- Profil santé perçue et rapportée : le type 2 est celui qui rapporte le moins souvent de pathologie chronique (13.6%), ou de limitation fonctionnelle dans les actes de la vie quotidienne, au cours des 6 derniers mois (5.6%). Sa santé perçue et rapportée est la meilleure : 92.2% la disent bonne, contre 1% qui la considèrent mauvaise.

En résumé, le type 2 caractérise une personne jeune, plus souvent masculine, plus fréquemment étrangère ou française d'ascendance étrangère. Il a bénéficié d'une éducation moyenne et est plus fréquemment chômeur. Il est mal couvert contre le risque maladie. Il désigne peu de proches atteints de pathologie grave, ne consulte qu'en dernier recours. Le soutien social est contrasté, mais la fréquence des contacts sociaux tend à être plus pauvre. Il se considère comme celui en meilleur santé.

Profil 3 : le représentant d'une population d'âge moyen, d'origines étrangères au niveau socioéconomique faible, couvert médiocrement, plutôt isolé et en mauvaise santé

Profils sociodémographique, économique, et sanitaire :

Le type 3 est le deuxième plus représenté parmi les effectifs (25.7%)

- Profil démographique : le type 3 présente un profil d'âge assez étalé et équilibré, moins représenté aux extrêmes : 35% ont entre 30 et 44 ans, 20.2% entre 45 et 59 et 14.1% entre 60 et 74 ans. Il est plus volontiers masculin (59.6%). De façon analogue au type 2, les étrangers ou les personnes d'ascendance étrangère sont plus représentées que dans les types 1 et 4, avec 15.7% d'étrangers, et 23.2% de parent étranger.
- Profil socioéconomique : le type 3 est celui qui a le plus de risques de n'avoir reçu aucune éducation scolaire, ou de n'avoir qu'un niveau d'études primaires (9.0%), tout en étant celui qui a le moins de chances d'avoir atteint un niveau d'études supérieures (48.7%). Il est le 1r à accuser le plus de chômage parmi ses représentants (9.3%), et il atteint 35.2% d'inactifs. Il est également le plus pauvre des 4 types, avec 24.6% appartenant au premier quintile de revenu, pour seulement 17.1% dans le 5e. Enfin, il

rapporte la 2e moins bonne couverture maladie, avec 85.3% de couverture de base plus complémentaire santé, 12.4% ne possédant pas de complémentaire, et 2.1% de bénéficiaires de l'AME ou de la CMUc.

- Profil rapport à la médecine : le type 3 est le dernier à connaître des proches atteints d'une pathologie grave (moins de la moitié), mais le 2e, à l'instar du type 1 à consulter son médecin dès qu'il en ressent le besoin (61.3%). En outre, il est celui à connaître le moins de médecins ou autres professionnels de santé parmi son entourage (61.8% n'en connaissent pas).
- Profil lien social : en termes de sentiment d'isolement, aucune différence significative entre types n'est trouvée. Cependant, notons que 0.2% se sentent très seuls, 12.1% plutôt seuls. Quant au soutien social, il est le 2e à le qualifier de bas et de peu fréquent (3.5%), et le 3e à en avoir un très bon (87.4%). Les fréquences de contacts sociaux mettent en évidence une tendance à une plus grande parcimonie dans le nombre de relations, à l'instar du type 1, même si là également, statistiquement, aucune significativité n'est exhibée.
- Profil santé perçue et rapportée : le type 3 est le 2e à se plaindre de pathologie chronique (35.8%), comme d'une limitation fonctionnelle dans les actes de la vie courante, au cours des 6 derniers mois (18.9%). De même, il est le 2e à rapporter une santé impactée, avec 2.6% de mauvaise santé perçue, et 77.4% de bonne santé.

En résumé, le type 3 représente une personne issue d'une population de tout âge, et de sex-ratio proche de 1. Il est cependant plus volontiers étranger ou d'ascendance étrangère. Il est également celui qui a le niveau d'éducation le plus bas, le plus souvent inactif ou chômeur, et le plus pauvre. Il a une mauvaise couverture maladie, connaît peu de professionnel de santé, et a recours dès que besoin au généraliste. Il présente un soutien social, une fréquence des contacts plus souvent faibles. Enfin, il est le 2e à se plaindre d'une santé médiocre ou mauvaise.

Profil 4 : la jeune française active, de haut niveau socioéconomique, très bien couverte, bien entourée et en bonne santé

Profils sociodémographique, économique, et sanitaire :

Le type 4 est le deuxième moins représenté en effectifs, avec 23.3% de la population totale.

- Profil démographique : le type 4 est le 2e plus jeune, avec 63.2% de ses effectifs en dessous de 45 ans, 23.7% de moins de 30 ans et 3.5% e 75 ans et plus. Il est également de loin le plus féminin (79%). De façon semblable au type 1, il est plus susceptible d'être français de parents français (69.1%), que français de parent étranger ou étranger (10.8%).
- Profil socioéconomique : le type 4 possède le profil de niveau d'étude le plus élevé, avec 63.1% de niveau supérieur, et le dernier rang concernant les personnes sans scolarité ou de niveau primaire (4.7%). Ils sont également les plus actifs, et les 2e moins chômeurs avec le type 1 (7.1%). Ils sont les 2e plus aisés économiquement, très proches des type 1 notamment pour le 1er quintile (18.9%), et possède plus d'un tiers de ses effectifs dans les 2 derniers quintiles (21.6% dans le 5e quintile et 17.8% dans le 4e). Le type 4 a la 2e meilleure couverture maladie, avec 89.6% d'entre eux possédant à la fois couverture de base et couverture complémentaire, tandis que seuls 6.8% ne présentent qu'une couverture simple sans complémentaire.
- Profil rapport à la médecine : il est le 2e à rapporter des proches présentant une pathologie grave (53%). Son rapport à la médecine est assez partagé : 56.8% d'entre eux consultent dès que quelque chose ne va pas. Enfin, il est celui qui compte parmi ses proches le plus souvent un professionnel de santé, avec le type 1 (44.2%).
- Profil lien social : si le sentiment d'isolement ne présente pas de différence significative entre type, on peut néanmoins relever une tendance à un faible isolement social ressenti (0.7% se sentent très seuls) ; il rapporte le meilleur soutien social avec 91.3% disant bénéficier d'un très bon soutien, 2.5% de très peu de soutien. De même, en termes de fréquences de contacts sociaux, il est avec le type 3 celui qui rapporte des contacts fréquents avec 29.7% dans le 4e quartile, 18.7% dans le 1r.
- Profil santé perçue et rapportée : il est le 2e à souffrir le moins souvent d'une pathologie chronique (27.8%), de même qu'à présenter une limitation fonctionnelle dans les actes de la vie quotidienne au cours des 6 derniers mois (13.5%). Enfin, en

termes de santé perçue, le type 4 rapporte la 2e meilleure santé perçue, avec 82.4% de bonne santé déclarée, contre 2.4% de mauvaise.

En résumé, le type 4 est une personne plutôt jeune, très souvent une femme, française de parents français, à l'excellente éducation, et à la situation financière en général avantageuse. Elle possède une très bonne couverture, connaît plus fréquemment des professionnels de santé dans son entourage et bénéficie d'un très bon réseau social. Elle se déclare globalement en bonne santé.

L'utilisation du système de santé à des buts judiciaires : demande d'estimation de l'âge chez les adolescents migrants

Au total, 499 adolescents ont été inclus. Les données étaient complètes pour 233 d'entre eux (47%). Les trois quarts d'entre eux étaient de sexe masculin. Les origines géographiques étaient réparties de la manière suivante : 96 venaient d'un pays d'Afrique (41%), 70 du Moyen-Orient (30%), 40 d'Asie (17%), 18 d'Europe de l'Est (8%), 5 d'Amérique du sud, 3 d'Europe de l'Ouest et 1 d'Océanie.

L'âge allégué moyen était de 16,4 ans (écart-type : 2,1), l'âge allégué médian de 16,5 ans (min-max : [9,0-36,0]). L'âge radiologique moyen, obtenu par l'atlas de Greulich et Pyle, était de 17,8 ans (1,6), et médian de 18,0 [9,0-19,5]. Enfin, l'âge estimé moyen et rendu comme conclusion par le médecin était de 17,9 ans (2,1) et l'âge médian de 18,5 [9,0-36]. Les autres résultats descriptifs sont consignés dans le Tableau 6 : Ages médians estimés par le médecin légiste, selon chaque variable d'intérêt (genre, présence des 2e et 3e molaires). Prises individuellement, toutes les variables étaient associées à des différences statistiquement significatives en termes d'âge médian estimé et rendu par le médecin. C'était particulièrement notable dans le cas de la présence des troisièmes molaires. Les âges estimés par le médecin étaient bien corrélés aux âges allégués (coefficient de Spearman de 0,73 $p<0,001$) et aux âges radiologiques (0,92 $p<0,001$).

Tableau 6 : Ages médians estimés par le médecin légiste, selon chaque variable d'intérêt (genre, présence des 2e et 3e molaires)

2e molaires	17**	27**	37*	47*
Présente	18.5 [15.5; 19.0]	18.5 [15.5; 19.0]	18.5 [15.5; 19.0]	18.0 [15.5; 19.0]
Absente	15.8 [10.5; 19.0]	15.5 [10.0; 17.5]	17.0 [11.0; 19.0]	16.0 [10.0; 19.0]

3e molaires	18**	28**	38**	48**
Présente	19.0 [17.5; 19.0]	19.0 [17.5; 19.0]	19.0 [17.5; 19.0]	19.0 [17.5; 19.0]
Absente	17.5 [14.5; 19.0]	17.5 [14.5; 19.0]	17.5 [14.5; 19.0]	17.3 [14.5; 19.0]

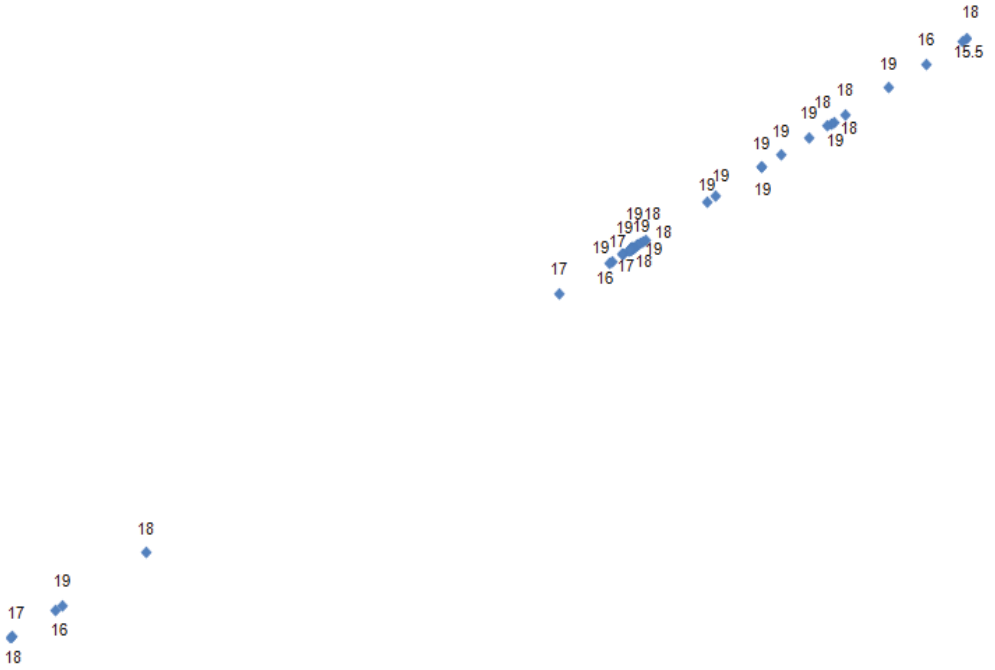
Genre**	
Masculin	19.0 [15.5; 19.0]
Féminin	17.5 [15.0; 19.0]

*: $p<0.01$, **: $p<0.001$. L'âge médian (en années) est donné pour chacun des cas où une des quatre 2e ou 3e molaires est présente ou absente, les 10e et 90e percentiles étant consignés entre crochets. Les 2e molaires sont numérotées 17-27-37-47. Les 4e molaires sont numérotées 18-28-38-48.

L'ACM rendait deux premiers axes principaux expliquant respectivement 16% et 11% de la variance totale. Les différents types d'âge (allégué, radiologique et estimé) semblaient répartis aléatoirement sur le plan défini par ces deux premiers axes principaux. Les âges inférieurs et supérieurs à 18 ans semblaient entremêlés les uns avec les autres, sans séparateurs linéaires évidents.

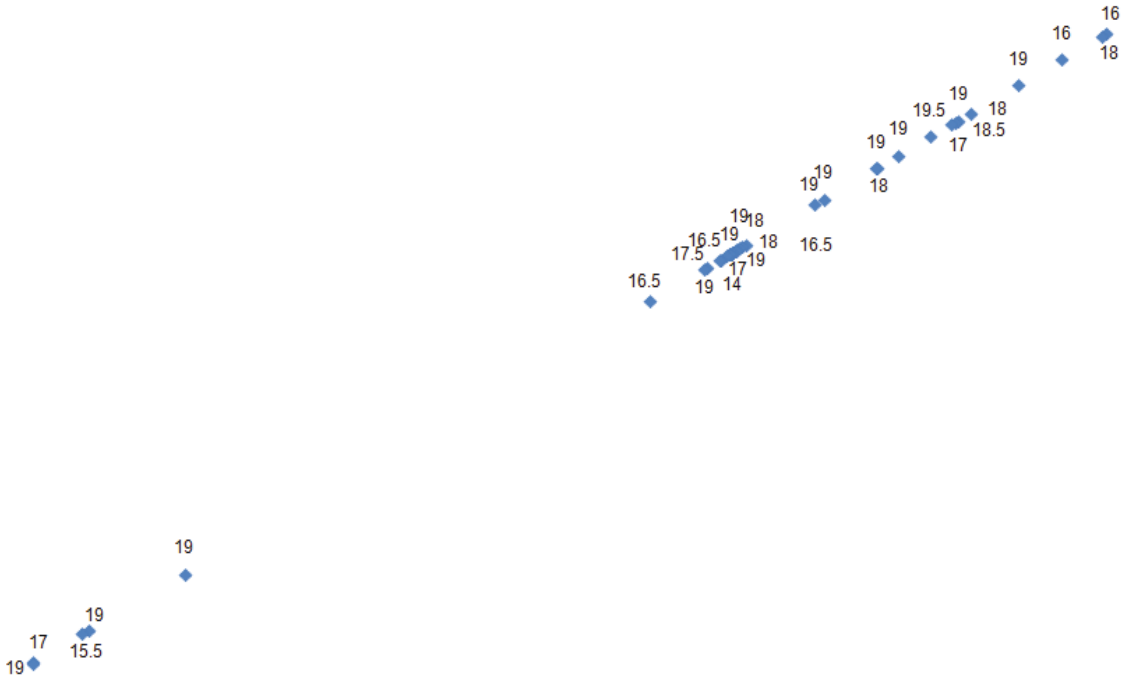
La dimension intrinsèque des données utilisées était estimée à 2, suggérant qu'il est possible de réduire significativement la taille des données considérées. L'algorithme ISOMAP n'a pris en compte que 101 observations, soit 43% des données entrées. Pour ces 101 personnes, une seule dimension était en fait suffisante, car ainsi qu'on le voit sur les Figures 7 et 8, ces 101 personnes sont parfaitement alignées les unes avec les autres (ce qui n'est pas que visuel : les coordonnées s'appuient parfaitement sur une droite). Dans le cas de l'autoencodeur, qui est conservateur, on retrouve une portion de l'espace présentant une telle propriété, à savoir un alignement des personnes dans le plan. En revanche, les autres personnes sont réparties un peu partout sur le plan. Les Figures 9-10-11 ne représentent pas toutes les observations pour des raisons de lisibilité, mais un tirage aléatoire de 40 observations. Tant les âges estimés, ou que les âges allégués ou radiologiques semblaient répartis aléatoirement sur l'ensemble du plan. On ne distinguait aucune région évidente où se seraient concentrés les âges inférieurs ou supérieurs à 18 ans.

Figure 7 : Application de l'algorithme ISOMAP à l'estimation de l'âge chez les adolescents migrants : échantillon aléatoire de la distribution des âges radiologiques après réduction de l'espace de description.



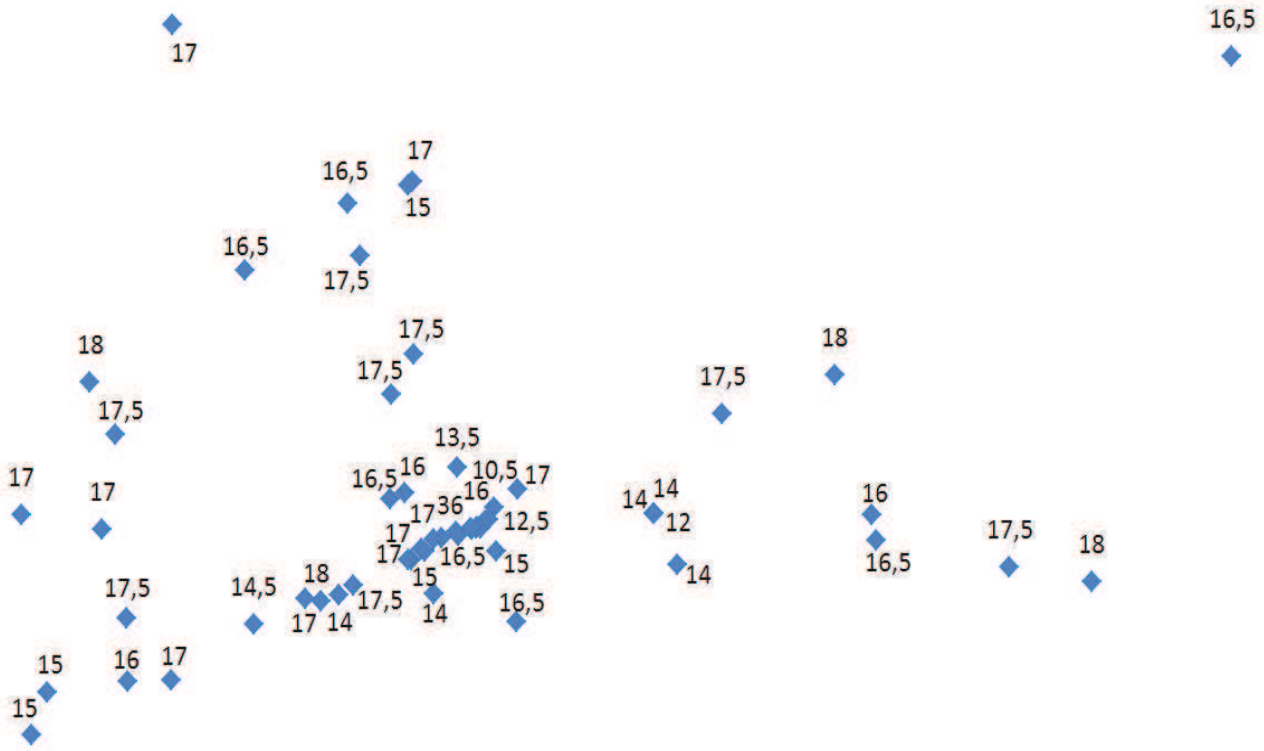
L'algorithme ISOMAP a identifié 101 individus suffisamment connectés les uns avec les autres sur un total de 233. Plus les individus sont similaires en termes de caractéristiques cliniques, dentaires et radiologiques, plus ils sont spatialement proches les uns des autres. Ici, nous voyons que deux individus proches spatialement peuvent présenter des âges radiologiques totalement différents. Les âges semblent être distribués aléatoirement le long de la droite. Aucun axe n'est représenté, car ceux-ci seraient totalement arbitraires et non informatifs. Seules les positions relatives des points dans le plan a ici un sens.

Figure 8 : Application de l'algorithme ISOMAP à l'estimation de l'âge chez les adolescents migrants : échantillon aléatoire de la distribution des âges estimés après réduction de l'espace de description.



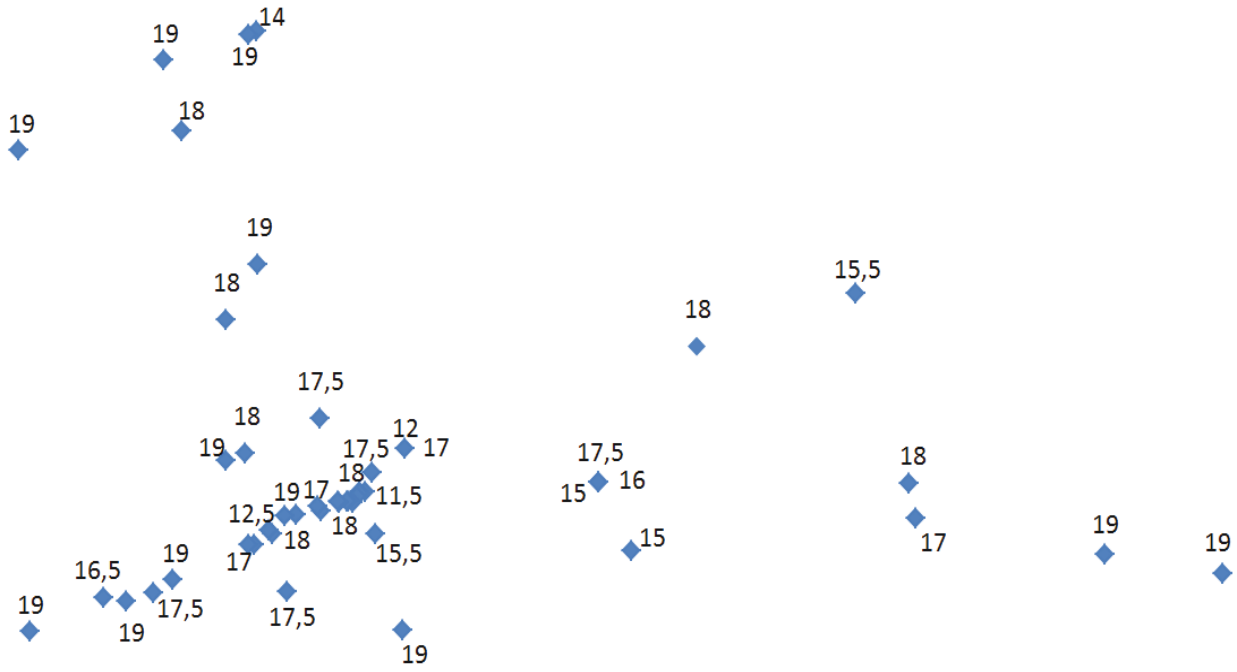
L'algorithme ISOMAP a identifié 101 individus suffisamment connectés les uns avec les autres sur un total de 233. Plus les individus sont similaires en termes de caractéristiques cliniques, dentaires et radiologiques, plus ils sont spatialement proches les uns des autres. Ici, nous voyons que deux individus proches spatialement peuvent présenter des âges estimés par le légiste totalement différents. Les âges semblent être distribués aléatoirement le long de la droite. Aucun axe n'est représenté, car ceux-ci seraient totalement arbitraires et non informatifs. Seules les positions relatives des points dans le plan a ici un sens.

Figure 9 : Application de l'algorithme autoencoder à l'estimation de l'âge chez les adolescents migrants : échantillon aléatoire de la distribution des âges allégués après réduction de l'espace de description.



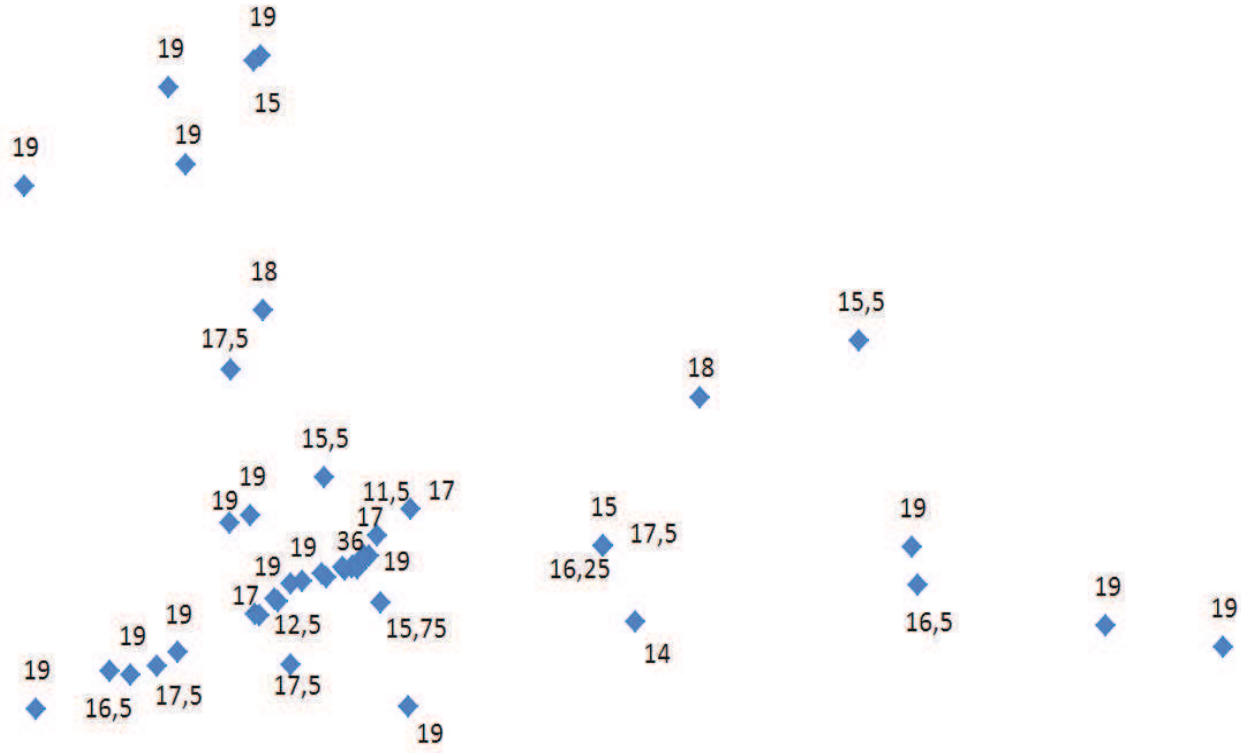
L'algorithme autoencoder est un algorithme conservative, qui prend ici en compte la totalité des 233 individus. Il les répartit sur l'ensemble du plan, tout en en préservant les relations topologiques (relations de voisinage, ou de ressemblance). Plus les individus sont similaires en termes de caractéristiques cliniques, dentaires et radiologiques, plus ils sont spatialement proches les uns des autres. Ici, nous voyons que deux individus proches spatialement peuvent présenter des âges allégués totalement différents. Il ne semble pas exister de région clairement ou linéairement identifiées pour lesquelles ces âges soient similaires ou très proches. 50 individus ont été tirés au sort pour être représentés sur ce schéma, afin qu'il demeure lisible. Aucun axe n'est représenté, car ceux-ci seraient totalement arbitraires et non informatifs. Seules les positions relatives des points dans le plan a ici un sens.

Figure 10 : Application de l'algorithme autoencoder à l'estimation de l'âge chez les adolescents migrants : échantillon aléatoire de la distribution des âges radiologiques après réduction de l'espace de description.



L'algorithme autoencoder est un algorithme conservative, qui prend ici en compte la totalité des 233 individus. Il les répartit sur l'ensemble du plan, tout en en préservant les relations topologiques (relations de voisinage, ou de ressemblance). Plus les individus sont similaires en termes de caractéristiques cliniques, dentaires et radiologiques, plus ils sont spatialement proches les uns des autres. Ici, nous voyons que deux individus proches spatialement peuvent présenter des âges radiologiques totalement différents. Il ne semble pas exister de région clairement ou linéairement identifiées pour lesquelles ces âges soient similaires ou très proches. 50 individus ont été tirés au sort pour être représentés sur ce schéma, afin qu'il demeure lisible. Aucun axe n'est représenté, car ceux-ci seraient totalement arbitraires et non informatifs. Seules les positions relatives des points dans le plan a ici un sens.

Figure 11 : Application de l'algorithme autoencoder à l'estimation de l'âge chez les adolescents migrants : échantillon aléatoire de la distribution des âges estimés après réduction de l'espace de description.



L'algorithme autoencoder est un algorithme conservative, qui prend ici en compte la totalité des 233 individus. Il les répartit sur l'ensemble du plan, tout en en préservant les relations topologiques (relations de voisinage, ou de ressemblance). Plus les individus sont similaires en termes de caractéristiques cliniques, dentaires et radiologiques, plus ils sont spatialement proches les uns des autres. Ici, nous voyons que deux individus proches spatialement peuvent présenter des âges estimés par le légiste totalement différents. Il ne semble pas exister de région clairement ou linéairement identifiées pour lesquelles ces âges soient similaires ou très proches. 50 individus ont été tirés au sort pour être représentés sur ce schéma, afin qu'il demeure lisible. Aucun axe n'est représenté, car ceux-ci seraient totalement arbitraires et non informatifs. Seules les positions relatives des points dans le plan a ici un sens.

Discussion et perspectives

Interprétation de la typologie de recours aux soins à la lumière du modèle d'Andersen

Du point de vue des inégalités de santé, les résultats que nous avons obtenus peuvent aider à clarifier certains aspects du recours, ou du non recours aux soins, certains comportements vis-à-vis du système de soins. Si l'on part de l'hypothèse que l'essentiel des inégalités sociales de santé repose sur des inégalités d'ordre économiques,[129,236,237] comme les inégalités de revenus, l'introduction en France, en 1999, d'une couverture maladie universelle aurait dû abaisser sinon éliminer cette barrière à l'accès aux soins.[238] En réalité, l'effet escompté n'est pas observé, du moins à la hauteur de ce que l'on pouvait espérer.[239–241] Les causes, les processus sociaux sous-tendant les inégalités de recours aux soins sont complexes, et dépassent les facteurs purement économiques ou matériels ; ils impliquent différents facteurs d'ordre psychosocial et comportemental.[4,104] Nous pensons que nos travaux contribuent, ou peuvent contribuer à dévoiler et représenter fidèlement une partie de cette complexité, de cette diversité.

D'un côté, nous avons identifié plusieurs des associations connues de la littérature entre un type de recours aux soins donné et des facteurs sociaux. Nous pensons par exemple à l'état de santé perçu ou mesuré[129,242] ou plus généralement encore, aux besoins de santé perçus,[115] au capital social,[112] ou à l'intégration sociale.[243] Tous ces déterminants coexistent et contribuent dans des proportions variées à la structure des 4 profils de recours aux soins que nous avons dégagés. Par exemple, l'état de santé, mesuré au travers de l'existence d'une maladie chronique, de limitations fonctionnelles, était effectivement associé à une utilisation plus importante du système de soins, de même qu'un niveau d'éducation ou de revenu plus élevés. Nous avons également identifié les différences bien connues de recours aux soins en lien avec le genre.

Parmi l'ensemble des variables considérées, seul le fait d'avoir des professionnels de santé dans son entourage n'était pas discriminant pour la typologie ; la question est en fait peut-être trop générale pour être d'une grande utilité informationnelle. Pour ce qui est de l'intégration sociale, la fréquence des contacts apparaissait aussi peu discriminante. La donnée brute des fréquences des contacts sociaux, sans aucun autre détail sur leur qualité, leur contexte ou leur contenu, a sans doute moins de valeur que d'interroger directement les personnes sur leur perception d'un sentiment d'isolement social. C'est le sens que nous donnons aux différences

observées entre les types 1 et 3. Pris ensembles, ces résultats, tout en étant cohérents avec la littérature, confèrent du sens et donc de la validité à notre typologie.

Ainsi, nous pouvons examiner la forme de la typologie de recours aux soins à la lumière du cadre d'interprétation des modèles comportementaux d'Andersen, notamment dans une optique de mesure de l'équité d'accès au système de soins. Si l'on s'en tient aux variables discriminant le mieux les différents types – nous prenons ici arbitrairement le critère de variables pour lesquelles on mesure des odds-ratios ≥ 2 ou $\leq 0{,}5$ – les facteurs prédisposant étaient l'âge, le genre, les origines, le niveau d'éducation, le sentiment d'isolement et les habitudes de consultation (pour les types 2 et 4, en ce qui concerne ce dernier facteur). Les *enabling factors* les plus importants étant pour leur part représentés par le sentiment d'isolement et le type de couverture maladie (ici aussi, pour les types 2 et 4).

Les personnes appartenant au type 1 semblent avoir besoin d'accéder au système de soins, en raison de la forte prévalence des maladies chroniques et des limitations fonctionnelles qu'elles rapportent. De manière attendue, elles utilisent effectivement le système, à la fois parce qu'elles peuvent se le permettre, financièrement et en termes de temps, et que leurs habitudes et perceptions les rendent à même d'y recourir. Notre étude ne fournit pas d'information sur le fait que leur accès aux soins corresponde ou non à leurs besoins, mais il est remarquable de constater que leur mode de recours ne semble pas s'expliquer en termes de facteurs prédisposant tels que le genre ou le niveau d'éducation.

Les personnes appartenant au type 2 sont majoritairement des hommes jeunes, avec un niveau de recours aux soins qui peut refléter le faible niveau de leurs besoins perçus. Ils sont également les plus susceptibles de ne présenter qu'une couverture maladie « de base », sans complémentaire. Cette caractéristique est connue pour être un obstacle à l'accès aux soins ; cette situation est préoccupante quant à l'équité d'accès aux soins, spécifiquement en raison d'une sous-estimation probable des besoins réels par les jeunes adultes.

Les personnes relevant du type 3 ont montré de faibles taux d'utilisation du système, et de multiples facteurs prédisposant négatifs. Ainsi, ils sont plus volontiers d'origines étrangères, avec de bas niveaux d'éducation et de revenus, alors qu'ils sont simultanément plus susceptibles de souffrir de maladie chronique et de limitations fonctionnelles. Il semble que le système de soins soit le plus inéquitable pour les personnes de ce type. Ils préfèrent ou doivent recourir essentiellement aux généralistes ou aux services d'urgences, loin devant aux services des spécialistes.

Enfin, les personnes du type 4 expriment peu de besoin, mais ont facilement recours au système, notamment aux généralistes et aux spécialistes d'accès direct. Elles disposent des ressources appropriées à un recours aisé au système de soins, tant en termes économiques – par leurs revenus et leur couverture maladie – qu'en termes culturels – majoritairement de genre féminin, d'un haut niveau d'éducation et fortement soutenues socialement.
Si l'on veut se résumer, au moins un des quatre types, éventuellement deux, présentent des caractéristiques qui, selon le modèle d'Andersen, soulignent l'iniquité du système de soins français en termes d'accès.

Comprendre la structure de la typologie : interprétation par états propres, ou interprétation par états déformables

Parmi les 3006 patients de la cohorte francilienne SIRS, nous avons identifié 4 grands groupes correspondant à autant de comportements vis-à-vis du système de soins, dans un espace du recours aux soins décrits par une petite vingtaine de variables. Ce nombre correspondait à un optimum marqué en termes de robustesse, c'est-à-dire en termes d'invariance par changement d'algorithme, de mesure de similarité et de présentations des données. Cette bonne délimitation des groupes était confirmée par plusieurs analyses multinomiales subséquentes, mettant en évidence des différences significatives pour la quasi-totalité des variables étudiées ; les fluctuations aléatoires semblaient donc bien circonscrites, même au sein des groupes. En outre, les analyses multinomiales reliant l'appartenance à un des 4 groupes en fonction de variables « explicatives » montraient elles aussi une bonne séparabilité des caractéristiques sociodémographiques, économiques, de lien social ou de représentations médicales. Cette typologie en 4 classes était donc porteuse de sens.
Comme nous en avions discuté en début de ces travaux, l'interprétation de l'existence de ces 4 groupes n'est pas nécessairement directe, et dépend jusque ce jour de la sensibilité ou des préférences de celles ou ceux qui désirent l'interpréter.
Nous nous en expliquons en donnant deux points de vue qui paraissent a priori peu conciliables, et qui n'excluent ni l'un ni l'autre une confirmation expérimentale. L'existence de ces 4 classes peut être vue comme l'existence de 4 « états propres » de la structure de l'espace de recours aux soins. Ce type d'interprétation est compatible avec soit une vision quantique des mécanismes à l'œuvre dans la structuration du recours aux soins, soit avec une vision type théorie des jeux, qu'évoque volontiers Bourdieu (dans Questions de sociologie)[244]

quand on lui demande si les types que l'on peut observer sont des types fixes ou fixés. Dans le cas quantique, l'espace de description est un espace « absolu », sa structure est globalement fixe, et il existe des repères, invariants par un certain type d'opérateur (c'est-à-dire, une opération d'observation du système, un mesure de sa structure) : ce sont ses états propres. Ainsi, l'identification de groupes qui soient invariants sous certaines transformations peut aller en ce sens. Cette interprétation impliquerait que l'espace de recours aux soins, si on en mesure l'état à un moment quelconque, c'est-à-dire, si l'on mesure la position d'une personne dans l'espace du recours aux soins à un moment quelconque, sa position sera une combinaison linéaire des 4 groupes identifiés, avec la possibilité évidemment d'appartenir strictement à un groupe précis. Cependant, la donnée des 4 groupes détermine totalement la structure de l'espace de recours aux soins : c'est une propriété intrinsèque de cet espace. Dans la perspective de Bourdieu, on considérerait que l'existence de ces 4 groupes est le résultat d'un arbitrage entre comportements admissibles dans cet espace, et que l'on est dans un état stable et nécessaire de cet espace : on ne saurait finalement en bouger, et là également, la donnée de ces groupes fige totalement la structure de l'espace – ce qui n'empêche pas une personne de passer de l'un à l'autre dans le temps ou sous l'action d'un ou de plusieurs facteurs particuliers.

L'existence de ces 4 groupes peut également être vue comme la photographie instantanée de l'état de l'espace de recours aux soins, qui peut présenter différentes formes ou différents états, lesquels sont déformables, scindables ou résorbables dans le temps ou sous l'action d'un ou plusieurs facteurs. Ainsi, l'invariance et la robustesse dont font preuve les 4 groupes identifiés n'en sont pas pour autant artificielles, et renseignent sur de « véritables » états du système, mais ne dit rien, ou peu de manière univoque, sur d'une part la forme générale, la structuration de cet espace per se, sur sa dynamique particulière. Il n'est pas exclu que l'observation faite de ces 4 groupes corresponde à un état le « plus probable », ou le plus stable des états de ce système, que ce soit dans le temps ou sous la perturbation raisonnable par un ou plusieurs facteurs. Nous sommes alors dans une optique davantage analogue soit à un point de vue relativiste, soit un point de vue « catastrophiste » - au sens de la théorie des catastrophes. En effet, dans le cas relativiste, il peut être fait des mesures spécifiques d'un espace, identifier des trajectoires spécifiques qui sont modelées par un ou plusieurs facteurs, car c'est tout l'espace des recours aux soins qui est alors un continuum ou quasi-continuum pouvant se déformer sous l'influence de facteurs. La photographie à un instant donné des personnes prenant part à ces trajectoires ne renseignent pas suffisamment sur la dynamique de l'espace, a priori, pour en inférer par des méthodes classiques la géométrie globale. Les 4

groupes observés peuvent donc être en soi des groupes déformables, interdépendants. Dans le cas de l'interprétation catastrophiste, il faut se souvenir que les catastrophes permettent de décrire « le discret à partir du continu, la qualité à partir de la quantité », par de brusques changements de topologie (de géométrie). Aussi, les 4 groupes observés peuvent être le résultat de changements brusques de topologie de l'espace de recours aux soins, qui lui est finalement un continuum soumis au contrôle d'un ou plusieurs facteurs. La variation de ces facteurs peut faire apparaître ces changements brusques de topologie, et donc révéler l'existence de groupes apparemment séparés. La théorie des catastrophes a depuis été intégrée au sein de la théorie des systèmes dynamiques. Dans ce cadre plus vaste et général, on peut également interpréter les groupes comme des états stables du système, voire des états attractants : l'espace de recours aux soins n'est pas parcourable en pratique dans toute son intégralité, mais confine en des régions spécifiques de son espace une infinité de trajectoires, qui « orbitent » autour de points d'accrétion (c'est-à-dire, de points attirant les trajectoires). Ainsi, les groupes sont des objets dynamiques mais stables, sans cesse parcourus par des trajectoires ressemblantes mais uniques, des personnes gravitant dans cet espace du recours aux soins : l'objet en lui-même existe, mais n'existe qu'à la fois par la structure de l'espace et par l'existence des personnes qui y sont décrites – il n'y a pas de groupe sans trajectoires, et les trajectoires ne peuvent que respecter les conditions d'existence imposées par le groupe.

L'expérience, si elle est possible, permet d'approcher des caractéristiques de ces 2 interprétations opposées, mais ne permettrait pas nécessairement de trancher en faveur de l'une ou l'autre sans ambiguïté. Le choix le plus raisonnable pourrait l'emporter si l'on révélait par exemple que la deuxième interprétation est en théorie vraie et vérifiée, mais ne représente des variations inexpliquées par la première interprétation que dans des cas très hypothétiques ou extrêmes (par exemple : disparition du système de soins, passage à un régime anarchique).

Cette incertitude entre deux modèles n'est pas forcément rédhibitoire, et n'empêcherait pas, si les techniques se développaient pour en exploiter les conceptions, d'en étudier les dynamiques propres ; seulement le formalisme, et les implications en termes de mécanismes d'action, de politique ne seraient peut-être pas les mêmes. Dans le cadre de la seconde interprétation, il serait possible que l'énergie à investir pour déformer ou faire apparaître / disparaître certains groupes soit hors de portée, ou en contradiction avec d'autres considérations (éthiques, économiques). Dans le cadre de la première interprétation, les états sont constitutifs de l'espace de recours aux soins, et il n'est pas question de pouvoir soit en créer d'autres, soit les

faire disparaître – au mieux, il peut être envisagé d'en augmenter ou diminuer les effectifs. Là aussi, tout dépend des types de facteurs à manipuler, et de l'énergie à mobiliser pour contraindre le système : en effet, si l'état naturel du système est de comporter 4 groupes, cela signifie qu'en réduire la représentation d'un ou de plusieurs est « contraire à la nature » de ce système, et nécessitera forcément une énergie plus importante, et à maintenir dans le temps – donc des efforts à entreprendre et à conserver.

Un certain déterminisme social – de la notion de déterminisme

Les groupes que nous avons identifiés existent d'une part parce que les personnes qui les constituent existent et peuvent les représenter, et d'autre part parce qu'ils s'imposent à ces personnes, en tant qu'objets sociaux, produits de l'espace de recours aux soins. En cela, il existe un déterminisme qui est celui de l'existence de ces groupes : on y appartient ou pas. Notons que dans le cas d'une interprétation quantique, ce déterminisme est finalement moins fort : à un moment quelconque, une personne peut appartenir à une « combinaison linéaire » de ces états, donc ne pas appartenir exclusivement à un groupe, en tant qu'individu particulier. Ce type de déterminisme est un déterminisme structurel : il est nécessaire, car les choses existent et s'organisent en formes déterminées.

Selon le point de vue dynamique, la notion de déterminisme social est tout aussi intéressante à étudier. En effet, les systèmes dynamiques sont en général définis par des relations purement déterministes – il n'y a pas d'interventions de l'aléatoire ou de l'incertitude a priori. Pourtant, les systèmes présentant des propriétés chaotiques, des attracteurs dits étranges comme ceux qui peuvent décrire l'existence de groupes dynamiquement stables et bien délimités, ont cette propriété spéciale qui fait que les trajectoires individuelles ne sont pas prévisibles avec précision, à plus ou moins moyen terme. Simultanément, la forme des régions attractantes peut être parfaitement décrite et circonscrite, définissant un groupe à part entière. De la sorte, alors que le système présente un déterminisme là aussi structurel – sa configuration induit de facto des groupes particuliers – ni l'appartenance d'une personne à un groupe ni sa trajectoire particulière au sein de ce groupe, ne sont prévisibles ; cette dernière étant par ailleurs unique.

On le voit, le déterminisme social peut coexister avec les notions d'incertitude, d'imprédictibilité et de changement. Nous sommes loin d'un lien unicausal monogénique déterminant une position sociale – sans parler de trajectoire.

Limites de l'étude de recherche d'une typologie de recours aux soins

Certaines limites dont souffre notre étude sont des limites classiques, mais doivent néanmoins être soulignées. Premièrement, les données SIRS sont des données déclaratives, sans liens à l'heure actuelle avec des bases de données de consommation de soins, médicales ou d'assurance maladie ; nous ne disposons donc pas de mesures objectives. Par ailleurs, des biais de rappel peuvent exister. Evidemment, les données de recours aux soins sont intimement liées au système de soins, comme il l'a été présenté. De fait, il est difficile sinon impossible d'extrapoler tout ou partie des résultats dégagés ici pour d'autres systèmes de soins – notamment parce que les politiques de régulation et de financement varient d'un pays à l'autre, ce alors qu'un système comme le système de soins présente intrinsèquement une grande complexité, sur différentes échelles. Un exemple concret de spécificité peut être illustré si l'on considère le profil de type 4 : ce profil concerne très volontiers des femmes, qui par ailleurs ont une utilisation importante des spécialistes d'accès directs, en majorité les gynécologues. En France, les gynécologues peuvent être vus comme des « généralistes pour femmes », dans le sens où ils se chargent de tous les aspects de la santé des femmes, depuis la contraception jusque le suivi de grossesse, en passant par le dépistage du cancer cervical. Les gynécologues échappent au système français de *gate-keeping*, puisque les femmes peuvent consulter sans pénalité leur gynécologue directement. Cet aspect du système de soins demeure assez atypique par rapport aux autres systèmes connus à l'international.

S'il existe désormais une couverture maladie universelle, prenant en charge une partie des frais de santé, le revenu et le type de couverture, l'existence d'une complémentaire santé peuvent influer sur l'accès aux soins. Ce peut être par exemple le cas du profil de type 3. En effet, si l'on écarte les personnes les plus défavorisées et ceux souffrant de maladie chronique, en l'absence de complémentaire, environ 30% des frais demeurent à la charge des patients. En outre, selon le type de contrat supplémentaire souscrit, tous les types de reste à charge ne sont pas nécessairement remboursés, que ce soit en partie ou en totalité. Enfin, dans de nombreux cas de figure, quel que soit le montant pris en charge par les couvertures maladie, il s'agit en général d'un mécanisme de remboursement ; ce qui signifie que les frais doivent être réglés par la personne avant tout, qui sera ensuite remboursée. L'instauration d'un système de *gate-keeping* à la française peut également être source de distorsions dans l'accès et le recours aux soins, puisqu'il peut amener soit un surcoût direct, soit un remboursement moins important des frais.

Si les résultats ne peuvent être directement extrapolés aux autres systèmes de santé, il faut se garder de les étendre trop rapidement au reste de la France également. La région parisienne ne saurait être tenue pour représentative du territoire, s'agissant d'une région très urbanisée, en moyenne plus riche, et présentant une offre de soins plus riche et dense spatialement. Ces considérations moyennes ne doivent pas occulter d'importantes hétérogénéités, que l'on parle d'inégalités sociales ou bien de répartition spatiale de l'offre de soins.[239]

D'un point de vue plus technique, soulignons le fait que la particulière stabilité, la robustesse exceptionnelle des classes identifiées peuvent résulter en partie d'hypothèses assez contraignantes sus lesquelles fonctionnent les algorithmes de *clustering*. Sans revenir sur le caractère partiellement construit des groupes, il est fort possible que les groupes correspondent à une réalité, mais que leurs frontières manquent de finesse et de précision. Les méthodes types k-moyennes ou PAM reposent sur des considérations géométriques et ont été appliquées sur un espace « plat », de telle sorte que la forme générale des groupes identifiables et identifiés ne peut être très différente d'un ellipsoïde. Cela signifie que si l'espace de représentation utilisé n'est pas un « bon » espace (il fait se mélanger de beaucoup les différents objets, selon la dimension observée), ou simplement que les objets ne sont pas linéairement séparables ou très isolés, alors les algorithmes utilisés ne sauront pas démêler des formes très intriquées, interpénétrées, et couperont au mieux au milieu des observations dont on peut difficilement dire si elles appartiennent plus à un groupe qu'à un autre.
Enfin, selon certaines dimensions, comme l'utilisation des services d'urgence pour des soins non programmés, nous avons pu manquer de puissance statistique, pour ce qui concerne les analyses multinomiales.

Application des techniques d'apprentissage de variétés aux données SIRS

Pour cette première étude de recherche de typologie, nous avons préféré attaquer les données directement, sans manipulation ou transformation de l'espace brut. Nous avons donc encouru le risque que nous venons de décrire, à savoir travailler dans un espace qui n'est pas le mieux configurer pour les algorithmes de recherche de forme, qui ne savent au final identifier que des formes relativement régulières et excluantes. Par ailleurs, nous avons travaillé sur un espace à une vingtaine de dimensions, ce qui peut poser des problèmes quant à la validité des outils statistiques (« malédiction de la dimension »).

Pour ces deux raisons au moins, il serait intéressant de reprendre le même espace, et d'en estimer la dimension intrinsèque, puis de lui appliquer une technique de réduction de la dimension. A partir de cet espace, utiliser de nouveau des techniques de *clustering*, à la recherche de groupes homogènes. Il serait instructif alors, d'une part de vérifier si l'on obtient le même nombre optimal de groupes, avec la même robustesse, et d'autre part si les groupes obtenus sont qualitativement semblables à ceux identifiés sans réduction préalable de la dimension de l'espace de travail. D'une manière générale, ce type d'approche paraît de toute façon approprié à la manipulation de données telles que celles obtenues dans des cohortes à nombreuses dimensions comme SIRS.

Application à d'autres aspects : recherche de meal patterns *(J. Riou)*

Sur le modèle d'analyses proposé dans le cadre d'une recherche de typologie de recours aux soins, Julien Riou a pu au sein d'ERES entreprendre l'étude des façons dont les franciliens de SIRS prennent actuellement leurs repas.[245]

D'un côté, les données caractérisant les prises alimentaires recueillies dans le cadre de SIRS ont été utilisées : nombre de prises alimentaires déclarées comme « repas » par les participants (sans référence au modèle des 3 repas quotidiens, pour éviter toute approche trop normative), les lieux des prises, les plages horaires, si ces repas sont pris seuls ou accompagnés – le cas échéant, de qui ? – et toute activité concomitante du repas (discussion, regarder la télévision, travail…). De l'autre, plusieurs variables explicatives ont été extraites, comme les variables sociodémographiques, l'activité professionnelle, le revenu ou encore l'environnement social de la personne (par exemple, la constitution de sa cellule familiale, de son foyer).

La recherche de profils distincts s'est effectuée sur le premier ensemble de données, dont le nombre optimal a été donné par l'approche par la robustesse des groupes. Des modèles multinomiaux ont été construits pour mettre en relation appartenance à un profil et groupes de variables explicatives.

Au final, 5 profils distincts ont été identifiés, dont 2 ne respectaient pas le schéma traditionnel des 3 repas par jour. Ces deux profils concernaient ensemble un quart des participants (13 et 12%).

Insuffisance et inadéquation des techniques actuelles pour la détermination de l'âge chez les adolescents migrants

De notre point de vue, aucune technique actuelle ne permet de répondre correctement à la demande judiciaire. Aucune méthode combinée n'y parvient non plus à l'heure actuelle. Notre étude montre en outre que l'intégration de modalités différentes, dans la logique d'une amélioration de la précision des estimations, ne paraît pas des plus pertinentes. Par ailleurs, le simple constat d'une corrélation entre une modalité et un âge chronologique ne saurait régler la question de la prédiction de l'âge la plus certaine et précise. Enfin, il semble que selon la question à laquelle on désire répondre, il existe deux voies méthodologiques susceptibles de traiter de manière plus appropriée le problème.

Soit l'on désire estimer un âge en prenant en compte au mieux toute l'incertitude qui imprègne et le contexte de l'estimation et les modalités d'estimation, et l'approche conceptuelle qui paraît la plus adaptée est celle offerte par le cadre Bayésien.[3,163,246,247]

Soit l'on désire classer au mieux les personnes selon un seuil d'âge, correspondant ici au critère de majorité légale, auquel cas il semble opportun de recourir aux meilleures techniques prédictives que l'on connaisse, à savoir les SVM (*support vector machines*),[92,206] les réseaux de neurones[224,248] ou encore les classificateurs Bayésiens[249–251] – en bref, les approches dites de *machine learning*.[92,252,253]

L'un dans l'autre, l'identification d'une question claire de recherche et le recours à une méthode adaptée à la question ne résoudra pas à eux seuls, si tant est que les techniques évoquées puissent faire preuve d'excellentes performances, le problème principal qui est celui de pouvoir valider ces approches sur des populations qui sont concernées en pratique, et non sur des populations qui, a priori, ne ressemblent que peu à celles des adolescents migrants.

Limites de l'étude portant sur l'intégration d'informations multiples pour l'estimation de l'âge chez les adolescents migrants

Cette étude est à notre connaissance la seule en médecine légale, et vraisemblablement l'une des rares en biomédecine à s'intéresser directement à l'intérêt, aux avantages et aux limites liés à l'intégration de multiples informations, qui ne soit pas orientée par l'implicite que « *bigger is better* ».

Néanmoins, elle fait office d'étude prototype dans sa démarche, étant donné plusieurs de ses limites. La limite la plus embarrassante reste l'absence d'âge chronologique connu : en

l'absence de référence, il demeure difficile d'être catégorique quant aux résultats obtenus. Comme nous l'avons vu, le problème de l'âge contrôlé est une limite commune à toutes les études d'estimation d'âge, dans un sens ou dans un autre.

Il est également gênant de ne disposer de données complètes que pour la moitié des participants, même s'il nous a été possible de travailler sur presque 250 personnes, d'horizons divers.

Au regard de la production actuelle en médecine légale sur le sujet, il est dommage de ne pas avoir disposé de données plus riches encore, prenant notamment en compte soit des données d'examen physique comme le poids et la taille, le stade pubertaire (toujours d'actualité, en principe), soit des données de modalités plus récentes, comme les scanners de clavicule. Ces informations auraient donné plus de poids et de pertinence à notre propos, de portée à notre critique.

Néanmoins, notre intention première était de sensibiliser, sur un exemple concret et polémique, aux problèmes qu'induit la montée en dimensions, aux outils disponibles pour les prendre en compte et enfin, au fait que l'accumulation de données n'entraînent pas nécessairement un gain de performance. Enfin, nous espérons, dans un futur proche, pouvoir illustrer l'intérêt de formuler explicitement et précisément la question de recherche dans le cadre de l'estimation de l'âge chez les adolescents migrants, en particulier si l'on prend le parti de la prédiction, de la classification selon le critère de majorité. Nous appliquerons des techniques dédiées à cette tâche sur un petit groupe de migrants dont l'âge est par ailleurs documenté.

Recherche des variables de contrôle sous-tendant la variété : apport des réseaux bayésiens et des SEM

Quelle que soit l'interprétation préférée quant à l'existence des groupes identifiés dans un espace observé, la question des variables contrôlant la dynamique de passage entre groupes ou simplement la dynamique globale de l'espace déformable – ici, de la variété – se pose. En l'absence d'approche expérimentale, plusieurs possibilités peuvent s'offrir à nous.

D'une part, il est possible de recourir aux réseaux « causaux », comme les réseaux bayésiens.[3,254] La causalité dont il y est question est avant tout une causalité « informationnelle », à savoir que l'on cherche à déterminer si la connaissance de telle ou telle information influe sur la connaissance de telle ou telle variable. Par ailleurs, l'approche par

réseaux bayésiens permet de simplifier la structure causale de l'espace, en en présentant le squelette minimal. Ainsi, pour les variables descriptives de l'espace étudié, on peut chercher à étudier les relations de causalité entre ces variables et un autre ensemble de variables, classiquement envisagées comme « explicatives ». La connaissance de ces relations permet d'avoir une idée de ce qui permet de régler de manière la plus directe possible la dynamique des groupes – leur constituants ou leur structure-même.

D'autre part, s'il s'agit d'explorer des relations structurelles entre variables, et si l'on désire tester des hypothèses complexes, des hypothèses structurelles, il est possible de recourir à la modélisation par équations structurelles (SEM).[255] Ici, il n'est pas question d'explorer à l'aveugle l'espace des réseaux de causalité pour en dégager la forme la plus probable, mais bien de tester différents scénarii de relations entre variables.

De la recherche d'une typologie par techniques multivariées à l'utilisation du Big Data

Ces travaux nous ont permis de présenter des méthodes qui sont par nature des méthodes multivariées, ou plus spécifiquement, des méthodes multivariables. Elles sont plus ou moins inédites, selon que l'on parle des techniques de *clustering* ou des techniques d'apprentissage de variétés, et selon que l'on s'intéresse à un domaine plus qu'à un autre – on pensera en particulier au domaine de l'intelligence artificielle dont dérivent ces méthodes, dont le *clustering* depuis une cinquantaine d'années, et de la génétique, où ces techniques ont été sans doute les plus utilisées en biomédecine jusque présent. Ces méthodes représentent une avancée dans le sens où elles permettent de considérer plusieurs aspects d'un même problème simultanément, et non simplement des associations deux à deux, par exemple.

La recherche de typologies au sein de données plus ou moins vastes, dont on ignore la structure particulière, en est une illustration particulière : partant d'un espace à 20 variables, dont plusieurs possédant plusieurs niveaux et renseignées pour 3000 personnes, il semble difficile soit pour un opérateur humain d'y distinguer aisément des groupes ou des trajectoires, soit d'examiner toutes les comparaisons « à la main » de caractéristiques des personnes – comparaisons deux à deux dont rien ne nous dit qu'elles persistent si l'on considèrent une tierce variable.

En outre, nous avons vu que l'utilisation de vastes ensembles de données – le big data – présente des caractéristiques propres, inattendues et contre-intuitives. Plus que jamais, la caractérisation d'objets au sein de grands ensembles de données nécessite une approche

soignée, rigoureuse et adaptée. Il reste a contrario à la charge du big data de prouver sa réelle pertinence, que ce soit en termes prédictifs ou explicatifs. Il est également un défi à la découverte de modes de représentation synthétique de l'information.

L'échelle et l'utilisation de moyennes en méthodes classiques

La question des groupes, des typologies, la question sociale, implique la question de l'échelle (sociétés et individualités), et du sens des indices utilisés en routine pour représenter l'information. En épidémiologie sociale, l'analyse multiniveaux a été introduite afin de comparer des résultats concernant deux échelles ou plus, par exemple des propriétés au niveau individuel, au niveau d'un foyer et au niveau d'un quartier. Plus spécifiquement, de vérifier s'il n'existe pas des effets spécifiques d'un niveau, sinon totalement indépendants de ces composantes. Cette approche nous permet de revisiter l'interprétation d'un des indices les plus anciens et les plus utilisés : la moyenne.

Il est connu qu'une moyenne peut ne représenter aucune information individuelle pertinente, ne correspondant par exemple à aucune valeur existante prise par aucun des individus à partir desquels cette moyenne est calculée : pensons à une classe où les notes à un examen sont également réparties entre la note 5 et la note 15 (50% ont 5, 50% ont 15) : la moyenne de la classe est de 10 – note que personne n'a obtenue.

En revanche, il n'est pas impossible que la moyenne ait un sens « indépendant » des individus : elle peut avoir un sens en tant qu'elle caractérise un groupe. Ainsi, l'identification de groupes, de typologies peut très bien mettre en évidence des structures à une échelle donnée, présentant un sens particulier dans un espace social, alors qu'aucun constituant de ce groupe ne possède lui-même cette caractéristique : il bénéficie d'une caractérisation duale, en tant qu'individu et en tant que partie d'un groupe. Bien entendu, cette interprétation n'a de sens que si les groupes en question ont eux-mêmes un sens, une opérationnalité propre dans l'espace social.

Un tel point de vue met en lumière le dilemme d'une conception de la santé publique ou des politiques qui soient exclusivement adossées sur des considérations de groupes, qui peuvent ne pas refléter les réalités individuelles. Par ailleurs, si les individus n'ont pas un intérêt ou la conscience d'une appartenance à un groupe spécifique, il n'y a a priori que peu de chances pour que des politiques visant à ne déformer que les caractéristiques de groupe entraînent une adhésion suffisante des individus. Cela pose également la question des interventions personnalisées, ciblant les personnes dans leurs spécificités, puisque si cela modifie leurs

comportements ou états singuliers, il n'est pas évident d'en inférer la répercussion sur les propriétés de groupe – l'idéal étant qu'à l'intervention personnalisée corresponde un mouvement de l'individu d'un groupe à un autre, désigné comme préférable, sans en altérer la forme.

Pourtant, certains exemples tirés de situations connues ne sont pas nécessairement si évidents. Si l'on considère l'approche par facteurs de risque, notion purement statistique et non nécessairement causale, et par le « nombre de personnes à traiter pour éviter un évènement » (par exemple, un décès), on s'aperçoit que la mobilisation des individualités n'est pas nulle – même si elle est loin d'être totale ou homogène, uniforme. Traiter une personne contre un état réputé morbide, disons l'hypertension artérielle, n'implique pas que ce traitement prémunira fatalement cette personne contre un évènement indésirable : a priori, avec ou sans, il est probable – très probable, même – que rien de fâcheux ne lui advienne. En revanche, en termes de groupes, il est montré que traiter suffisamment de personnes présentant une hypertension rendra un service à quelques individus de ce groupe. Le bénéfice incertain à prendre un traitement, ou de manière complémentaire, le désavantage incertain de ne pas le prendre, n'empêche pas un nombre non négligeable de personnes de suivre le traitement prescrit. Bien entendu, on peut se représenter que la situation n'est pas clairement présentée ou vécue de la sorte par les individus : l'état morbide d'une part, et la prise d'un traitement censé en contrer les conséquences incertaines, sont vraisemblablement associées à son échelle singulière, dans l'esprit de la personne. Autrement dit, à l'échelle individuelle, la relation « à risque » entre état morbide et conséquences indésirables, établie en termes statistiques, s'instancie et se concrétise. On peut ici invoquer l'aversion au risque, mais il n'est pas évident que cela suffise à expliquer suffisamment les trajectoires et les mobilisations individuelles ou de groupes.

Un exemple d'application du big data en épidémiologie sociale

La cohorte SIRS est une cohorte reconduite depuis 2005, et présentant un grand nombre de questions, dont une bonne partie porte sur différentes dimensions sociales. Les personnes interviewées sont notamment interrogées sur leur exposition personnelle à différents types de violence. Cette spécificité permet d'articuler là encore épidémiologie sociale et médecine légale – médecine des situations de violence.

Les catégories sociales, les classes, en bref, la possible structure sociale de la population francilienne est échantillonnée à la fois par un nombre important de variables et par une participation représentative des personnes vivant en Ile-de-France.

Actuellement, dans le prolongement de nos travaux, déjà continués par Julien Riou au sein d'ERES, une étude porte sur la recherche de la structure sociale représentée dans SIRS par une approche massivement multivariée, relevant du big data (nombre important de dimensions, hétérogénéité des informations).

Ainsi, à partir d'une soixantaine de variables, il est recherché la dimensionnalité intrinsèque – la plus petite dimension nécessaire et suffisante à la description des données – de la structure sociale de SIRS, et cette dimension acquise, la forme de cette structure : le nombre de groupes, et leur composition.

En ce sens, la démarche s'inscrit dans le prolongement des travaux de Bourdieu lorsqu'il a proposé son espace social et sa structuration en termes de capitaux culturel et économique (La Distinction).[256]

L'obtention de cette structure sociale pourrait ensuite se raffiner en procédant à l'analyse des dimensions de cette structure, mais également à l'analyse des structures interne et externe des classes sociales identifiées. Nous appelons analyse de la structure interne l'étude des interrelations entre variables caractérisant chaque groupe, ou simplement entre leurs dimensions, et analyse de la structure externe des classes, l'étude des interrelations entre les classes elles-mêmes et possiblement avec d'autres variables externes à la caractérisation des classes. Cette étude peut s'envisager par l'utilisation de réseaux bayésiens ; alternativement, étant données des hypothèses de structures, ces hypothèses pourraient s'explorer par l'emploi de modèles structuraux.

Il est également possible de recourir à deux autres approches complémentaires, afin de cerner au mieux la robustesse et les dimensions de ces structures sociales : par le tests de techniques dites de *co-clustering*, où il s'agit d'identifier conjointement des groupes définis par les dimensions sociales et définis par des variables dites « explicatives » ; par l'utilisation enfin de techniques de *clustering* spatial, tirant ainsi parti des données géolocalisées de SIRS. A considérer que sans aller jusque le recours à ces techniques, il est déjà possible d'adjoindre les coordonnées spatiales à chaque personne de SIRS, afin d'en observer la répartition géographique des classes sociales.

Un exemple d'application du big data en médecine légale

Le service de médecine légale de l'hôpital Jean Verdier (Bondy – 93) s'est doté depuis 2008 d'une base de données alimentée en continu par les informations recueillis lors de chaque consultation effectuée, que ce soit des consultations pour violences volontaires ou involontaires, physiques, psychologiques ou sexuelles, pour maltraitance ou négligences lourdes, pour enfin un examen médical de personnes placées en garde à vue.

Ces informations sont recueillies à partir de certificats standardisés, dont certains ont été publiés en anglais et en français.

En pratique, cela signifie que sont renseignées quotidiennement environ 300 variables, pour un volume annuel de consultations oscillant entre 25 000 et 30 000. Là également, nous sommes dans un contexte de big data. Les données sont variées et hétérogènes en nature. Elles concernent aussi bien les données sociodémographiques que les circonstances des violences, ou encore les données de santé des personnes.

Si une partie des certificats sont enregistrés directement sous forme électronique, ce n'est pas le cas de tous, les certificats pour la garde à vue étant encore sous format papier carbone. La saisie des informations se fait donc par des opérateurs humains, à partir des certificats rédigés par les médecins.

Toutes les informations ne sont pas codées, ni même toute l'activité : c'est par exemple le cas de l'activité de consultation des psychologues. Nous sommes actuellement en train de formaliser le recueil des informations issues des consultations psychologues.

Néanmoins, nous désirons optimiser la chaîne de saisie, et l'exhaustivité du recueil d'information, tout en atténuant la pénibilité du travail de saisie actuel. Pour ce faire, nous nous rapprochons du laboratoire LIMICS (Paris 6 - Paris 13), qui est un laboratoire d'informatique médicale. L'idée est de bénéficier des avancées en *text mining* et en analyse sémantique, et d'automatiser le renseignement de la base de données à partir des certificats standardisés, tout en l'enrichissant.

Cette nouvelle étape nous permettrait d'améliorer la qualité des données, de les enrichir, de décharger le personnel accueillant d'une grande partie de saisie fastidieuse, de séparer la saisie administrative de la saisie à but de recherche, et enfin d'exploiter des données jusque présents soit non disponibles, soit peu exploitables. Par exemple, nous pourrions visualiser et analyser les rapports entre caractéristiques des lésions traumatiques et localisations anatomiques, ainsi que leur retentissement fonctionnel.

Une définition du big data – pas de médecine personnalisée sans médecine sociale

A ce jour, le big data a tous les attributs de ce qui pourrait devenir une « bulle », comme l'ont été les télécoms il y a une vingtaine d'années. Il ne dispose pas de définition consensuelle,[213] se nourrit presque toujours des mêmes exemples emblématiques, qui, à bien y regarder, ne correspondent jamais aux critères du big data, ou à peine. Sont en général mis en avant les exemples tirés des réseaux sociaux, type Facebook ou Twitter, ou encore l'utilisation de la publicité ou de l'annonce ciblée, par Google entre autres. Les données recoupées ne sont finalement pas si nombreuses, sans parler de leur qualité, relativement peu hétérogènes, et le big se limitent rapidement à un traitement de gros volumes, si possible rapidement. Le big data en santé est essentiellement au stade de promesse, et se définit – par défaut, en creux – surtout par la médecine personnalisée, de précision ou prédictive, ce qui ramène la plupart du temps à une médecine du génome.[59,61,257] Grossièrement, il s'agit d'étendre la bibliothèque des facteurs de risque connus, en séquençant systématiquement l'intégralité du génome.

Le dernier rapport McKinsey d'avril 2015 (document confidentiel) recense les usages et les attentes du big data en biomédecine. Le document se veut le plus exhaustif possible et le plus représentatif. Le choix des exemples reflète certainement le plus haut degré de qualité disponible actuellement. Or, les exemples présentés n'ont rien d'impressionnants ou de convaincants. Ainsi, telle entreprise affiche le succès de son système prédictif et l'illustre par une réduction de 13% des risques cardiovasculaires à 5 ans – et non des événements indésirables ou de la mortalité – à mettre en relation avec une prescription et une observance 6 fois plus importante concernant les statines. Or, parmi les facteurs de risque cardiovasculaire les moins susceptibles d'être causaux, se trouve l'hypercholestérolémie.[2,258] En particulier, l'hypercholestérolémie chez une personne sans antécédent personnel de pathologie cardiovasculaire, et hors hyperlipidémie familiale, n'est pas une indication à traitement. Le meilleur succès du système est donc de permettre le traitement d'un facteur de risque qui a toutes les chances de ne pas jouer de rôle dans la pathogenèse. Un autre exemple est celui d'un système prédictif d'utilisation du système de soins qui, pour être optimisé, s'est reposé sur le test d'une dizaine de techniques étiquetées prédictives, pour la plupart –sauf une, un réseau de neurones – linéaires (des régressions), et l'utilisation de presque 70 variables. Le meilleur modèle retenu est un modèle basé sur 69 variables, intégrées dans un modèle de régression logistique. L'aire sous la courbe correspondant à ses performances était de 0.74, ce qui est loin d'être idéal en termes de vrais positifs et vrais négatifs. Quant à la production scientifique actuelle, dans le domaine biomédical et concernant le big data, il existe un peu

moins de 900 articles indexés dans Pubmed (recherche sur le terme « big data »). La majorité ont été publiés dans les deux dernières années, et la très vaste majorité sont des articles de mise au point ou de position – pas de recherche.

Que l'accent du big data en biomédecine soit porté sur la médecine du génome, sur une possible précision issue de corrélations entre des observations et un gène ou un mutant donné, ce que certains, dont la voix est minoritaire,[60,259] mettent en doute, est à la fois gênant et un contre-sens. En effet, tout l'intérêt potentiel du big data réside non pas tant dans l'exhaustivité des données, ni encore dans d'énormes quantités d'observations, mais avant tout dans le recoupement d'informations qui d'ordinaire ne sont pas rapprochées. C'est finalement une opportunité rare de redéfinir une partie de ce que peut être l'interdisciplinarité, et d'élargir, sans parler de tester, les dimensions usuelles de la biomédecine. C'est l'occasion de valider, modifier, quantifier les intrications entre individus, groupes, société, systèmes de soins ou d'éducation. C'est encore la possibilité de dessiner le paysage de nos relations interpersonnelles,[77,260,261] notre rapport aux autres et à la santé, d'approcher la structuration dynamique des sociétés, comme aurait pu en rêver Bourdieu, qui n'en avait alors pas les moyens techniques.[256] Nous avions eu l'occasion, au travers d'une publication, de proposer ce type d'approche et de représentation de l'espace psychologique en population général, basés sur l'intégration d'informations multiples, médicales, symptomatologiques, biologiques et sociales.[77]

De fait, le big data en santé ne devrait pas dissocier la médecine personnalisée de la médecine sociale, ce qui est aussi vrai que de dire qu'il n'existe pas d'individu sans société et inversement. Le big data est le moyen de pratiquer l'interdisciplinarité, basée sur l'observation, et reposant sur l'utilisation de techniques adaptées à ses spécificités principales que sont la haute dimensionnalité, l'hétérogénéité et souvent, l'aspect distribué des observations.

Bibliographie

1. Lefèvre T. Formes et fonctions, en théories et en pratiques. [Internet] [Thèse d'université]. Brest : Université de Bretagne Occidentale (UBO); 2014 [consulté le 4 juin 2015]. Disponible sur : https://tel.archives-ouvertes.fr/tel-01078992/document

2. Lefèvre T, Stindel E, Ansart S, Roux C. Mathematics in medicine: beyond iatromathematics. Lancet 2014;383(9916):513.

3. Lefèvre T, Lepresle A, Chariot P. Detangling complex relationships in forensic data: principles and use of causal networks and their application to clinical forensic science. Int J Legal Med 2015; doi: 10.1007/s00414-015-1164-8. [Epub ahead of print]

4. Berkman LF, Kawachi I, Maria Glymour M, eds. Social Epidemiology. New York: Oxford University Press; 2000. 428 p.

5. Krieger N. Commentary: Society, biology and the logic of social epidemiology. Int J Epidemiol 2001;30(1):44‑6.

6. Kasl SV, Jones BA. Social epidemiology: towards a better understanding of the field. Int J Epidemiol 2002;31(6):1094‑7.

7. Chariot P, Debout M. Traité de médecine légale et de droit de la santé. Paris: Vuibert; 2010. 717 p.

8. Beauthier JP, Hédouin V. Traité de Médecine Légale. 2e édition. Bruxelles : De Boeck; 2011. 1056 p.

9. Évaluation du nouveau schéma d'organisation de la médecine légale - IGAS - Inspection générale des affaires sociales [Internet]. 2011 [consulté le 4 juin 2015]. Disponible sur: http://www.igas.gouv.fr/spip.php?article399

10. Mucchielli L. L'invention de la violence. Des peurs, des chiffres et des faits. Paris: Fayard; 2011. 344 p.

11. Fassin D. L'ombre du monde : Une anthropologie de la condition carcérale. Paris: Seuil; 2015. 601 p.

12. Chan CH, Tiwari A, Fong DYT, Ho PC. Post-traumatic stress disorder among Chinese women survivors of intimate partner violence: a review of the literature. Int J Nurs Stud 2010;47(7):918‑25.

13. Lagdon S, Armour C, Stringer M. Adult experience of mental health outcomes as a result of intimate partner violence victimisation: a systematic review. Eur J Psychotraumatology 2014;5.

14. Trevillion K, Oram S, Feder G, Howard LM. Experiences of domestic violence and mental disorders: a systematic review and meta-analysis. PloS One 2012;7(12):e51740.

15. Maniglio R. The impact of child sexual abuse on health: a systematic review of reviews. Clin Psychol Rev 2009;29(7):647‑57.

16. Miller AB, Esposito-Smythers C, Weismoore JT, Renshaw KD. The relation between child maltreatment and adolescent suicidal behavior: a systematic review and critical examination of the literature. Clin Child Fam Psychol Rev 2013;16(2):146-72.

17. Nanni V, Uher R, Danese A. Childhood maltreatment predicts unfavorable course of illness and treatment outcome in depression: a meta-analysis. Am J Psychiatry 2012;169(2):141‑51.

18. Homma Y, Wang N, Saewyc E, Kishor N. The relationship between sexual abuse and risky sexual behavior among adolescent boys: a meta-analysis. J Adolesc Health Off Publ Soc Adolesc Med 2012;51(1):18‑24.

19. Chalmers AF. Qu'est-ce que la science? Paris: Le Livre de Poche; 1990. 286 p.

20. Cohen HW. P values: use and misuse in medical literature. Am J Hypertens 2011;24(1):18‑23.

21. Falissard B. Statistics in brief: when to use and when not to use a threshold p value. Clin Orthop 2012;470(1):315‑6.

22. Greenland S, Poole C. Living with p values: resurrecting a Bayesian perspective on frequentist statistics. Epidemiol Camb Mass 2013;24(1):62‑8.

23. Neyman J, Pearson ES. On the Problem of the Most Efficient Tests of Statistical Hypotheses. Philos Trans R Soc Math Phys Eng Sci 1933;231(694-706):289‑337.

24. Smulders YM. A two-step manuscript submission process can reduce publication bias. J Clin Epidemiol 2013;66(9):946‑7.

25. Niveau de preuve et gradation des recommandations de bonne pratique - État des lieux - HAS. Haute Autorité de Santé [Internet]. [consulté le 4 juin 2015]. Disponible sur: http://www.has-sante.fr/portail/jcms/c_1600564/fr/niveau-de-preuve-et-gradation-des-recommandations-de-bonne-pratique-etat-des-lieux

26. Ioannidis JPA. Why most published research findings are false. PLoS Med 2005;2(8):e124.

27. Ioannidis JA, Khoury MJ. Assessing value in biomedical research: The pqrst of appraisal and reward. JAMA 2014;312(5):483-4.

28. Ioannidis JPA. Research accomplishments that are too good to be true. Intensive Care Med 2014;40(1):99‑101.

29. Chalmers I, Bracken MB, Djulbegovic B, Garattini S, Grant J, Gülmezoglu AM, et al. How to increase value and reduce waste when research priorities are set. Lancet 2014;383(9912):156-65.

30. Ioannidis JPA, Greenland S, Hlatky MA, Khoury MJ, Macleod MR, Moher D, et al. Increasing value and reducing waste in research design, conduct, and analysis. Lancet 2014;383(9912):166-75.

31. Ludwig J, Duncan GJ, Gennetian LA, Katz LF, Kessler RC, Kling JR, et al. Neighborhood Effects on the Long-Term Well-Being of Low-Income Adults. Science 2012;337(6101):1505‑10.

32. Brook R, Ware J, Rogers W, Keeler E, Davies A, Barry C. The effect of coinsurance on the health of adults. Results from the RAND Health Insurance Experiment. N Engl J Med 1983;309:1426‑34.

33. Godlee F. Evidence based medicine: flawed system but still the best we've got. BMJ 2014;348(2):g440‑g440.

34. Spence D. Evidence based medicine is broken. BMJ 2014;348(1):g22‑g22.

35. Smith R, Rennie D. Evidence based medicine--an oral history. BMJ 2014;348(34):g371‑g371.

36. Feyerabend P. Contre la méthode. Esquisse d'une théorie anarchiste de la connaissance. Paris: Seuil; 1988. 349 p.

37. Noether E, Kosmann-Schwarzbach Y, Meersseman L. Les théorèmes de Noether: invariance et lois de conservation au XXe siècle : avec une traduction de l'article original "Invariante Variationsprobleme. Palaiseau : Éd. de l'École polytechnique; 2004. 173 p.

38. Typologie [Internet]. Wikipédia. 2015 [consulté le 4 juin 2015]. Disponible sur : http://fr.wikipedia.org/w/index.php?title=Typologie&oldid=111700340.

39. Trésor de la langue française informatisé. Typologie [Internet]. [consulté le 4 juin 2015]. Disponible sur : http://atilf.atilf.fr/dendien/scripts/tlfiv5/advanced.exe?8;s=3609764700.

40. Linné C Système de la nature: classe première du règne animal, contenant les quadrupèdes vivipares et les cétacées. Whitefish: Kessinger publishing; réed 2009. 330 p.

41. Grémy JP, Moan JL. Analyse de la démarche de construction de typologies dans les sciences sociales. Inform Sci Hum [Internet]. 1977 [consulté le 4 juin 2014]. Disponible sur : http://halshs.archives-ouvertes.fr/halshs-00650400.

42. Parizot I. Soigner les exclus. Paris: Presses Universitaires de France; 2003. 256 p.

43. Coenen-Huther J. Classifications, typologies et rapport aux valeurs. Rev Eur Sci Soc Eur J Soc Sci 2007;(XLV-138):27‑40.

44. Durkheim É. Les règles de la méthode sociologique. Paris: Éd. Flammarion; 2010. 333 p.

45. Durkheim É. Le suicide : É tude de sociologie. Presses Universitaires de France; 2007. 463 p.

46. Oquendo MA, Baca-García E, Mann JJ, Giner J. Issues for DSM-V: suicidal behavior as a separate diagnosis on a separate axis. Am J Psychiatry 2008;165(11):1383‑4.

47. Dwivedi YY, ed. The Neurobiological Basis of Suicide. Boca Raton (FL): CRC Press; 2012. 482 p.

48. Kupfer D, Frank E, Phillips M. Major depressive disorder: new clinical, neurobiological, and treatment perspectives. Lancet 2012;379:1045‑55.

49. Stoltenberg SF, Christ CC, Highland KB. Serotonin system gene polymorphisms are associated with impulsivity in a context dependent manner. Prog Neuropsychopharmacol Biol Psychiatry 2012;39(1):182‑91.

50. Zhurov V, Stead JDH, Merali Z, Palkovits M, Faludi G, Schild-Poulter C, et al. Molecular Pathway Reconstruction and Analysis of Disturbed Gene Expression in Depressed Individuals Who Died by Suicide. PLoS ONE 2012;7(10):e47581.

51. Hvistendahl M. Making Sense of a Senseless Act. Science 2012;338(6110):1025‑7.

52. Halbwachs M. Les causes du suicide. Paris: Presses universitaires de France; 2002. 432 p.

53. Halbwachs M. La mémoire collective. Paris: Albin Michel; 1997. 304 p.

54. Halbwachs. Morphologie sociale. Paris: Armand Colin; 1999. 190 p.

55. Hanson RK, Morton-Bourgon K. L'exactitude des évaluations du risque de récidive chez les délinquants sexuels une méta-analyse [Internet]. 2007 [consulté le 4 juin 2015]. Disponible sur: http://epe.lac-bac.gc.ca/100/200/301/psepc-sppcc/accuracy_of_recidivism-f/PS3-1-2007-1F.pdf.

56. Hureau J, Olié J. Évaluation de la dangerosité psychiatrique et criminologique. [Internet]. 2012 [consulté le 4 juin 2015].Disponible sur: http://www.academie-medecine.fr/publication100100060/.

57. Senon J, Voyer M, Paillard C, Jaafari N. Dangerosité criminologique : données contextuelles, enjeux cliniques et expertaux. L'Information Psychiatr 2009;85:719‑25.

58. Rietveld CA, Medland SE, Derringer J, Yang J, Esko T, Martin NW, et al. GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment. Science 2013;340(6139):1467‑71.

59. The Lancet. Moving toward precision medicine. Lancet 2011;378(9804):1678.

60. Coote JH, Joyner MJ. Is precision medicine the route to a healthy world? Lancet 2015;385(9978):1617.

61. Chaussabel D, Pulendran B. A vision and a prescription for big data-enabled medicine. Nat Immunol 2015;16(5):435‑9.

62. Latour B. Nous n'avons jamais été modernes : essai d'anthropologie symétrique. Paris: La Découverte; 1997. 205 p.

63. Shapin S, Schaffer S. Leviathan and the air-pump Hobbes, Boyle, and the experimental life. Princeton: Princeton University Press; 2011. 448 p.

64. Latour B, Woolgar S, Biezunski M. La vie de laboratoire : La production des faits scientifiques. Paris: Editions La Découverte; 2005. 299 p.

65. Latour B, Biezunski M. La science en action : Introduction à la sociologie des sciences. Nouvelle éd. Paris: Editions La Découverte; 2005. 664 p.

66. Noyer J-M, Carmes M. L'irrésistible montée de l'algorithmique : méthodes et concepts en SHS [Internet]. 2013 [consulté le 4 juin 2015]. Disponible sur: http://archivesic.ccsd.cnrs.fr/sic_00917623.

67. Lucrèce. De la nature. Paris: Editions Flammarion; 1997. 552 p.

68. Héraclite. Fragments. Paris: Flammarion; 2002. 374 p.

69. Piazza L, Lummen TTA, Quiñonez E, Murooka Y, Reed BW, Barwick B, et al. Simultaneous observation of the quantization and the interference pattern of a plasmonic near-field. Nat Commun 2015;6:6407.

70. Shannon CE. A Mathematical Theory of Communication. Bell Syst Tech J 1948;27:379‑423.

71. Tucci RR. Shannon Information Theory Without Shedding Tears. arXiv:12082737 [Internet]. 2012 [consulté le 4 juin 2015]; Disponible sur: http://arxiv.org/abs/1208.2737.

72. Linder R. Les plans d'expériences: un outil indispensable à l'expérimentateur. Paris: Presses de l'École nationale des ponts et chaussées; 2005. 320 p.

73. Regier DA, Narrow WE, Kuhl EA, Kupfer DJ. The conceptual development of DSM-V. Am J Psychiatry 2009;166(6):645‑50.

74. Adam D. Mental health: On the spectrum. Nature 2013;496(7446):416‑8.

75. Ketter TA, Wang PW, Becker OV, Nowakowska C, Yang Y. Psychotic bipolar disorders: dimensionally similar to or categorically different from schizophrenia? J Psychiatr Res 2004;38(1):47‑61.

76. Keshavan MS, Morris DW, Sweeney JA, Pearlson G, Thaker G, Seidman LJ, et al. A dimensional approach to the psychosis spectrum between bipolar disorder and schizophrenia: The Schizo-Bipolar Scale. Schizophr Res 2011;133(1-3):250‑4.

77. Lefèvre T, Lepresle A, Chariot P. An alternative to current psychiatric classifications: a psychological landscape hypothesis based on an integrative, dynamical and multidimensional approach. Philos Ethics Humanit Med 2014;9:12.

78. Thom R. Stabilité structurelle et morphogénèse; essai d'une théorie générale des modèles. Reading, Mass.: W.A. Benjamin; 1972. 362 p.

79. Burroni E. La Topologie des espaces métriques. Paris: Ellipses Marketing; 2005. 210 p.

80. Berger M, Gostiaux B. Géométrie différentielle: variétés, courbes et surfaces. Paris: Presses universitaires de France; 1992. 520 p.

81. Zeeman EC. Catastrophe theory: selected papers, 1972-1977. Reading: Addison-Wesley Pub. Co., Advanced Book Program; 1977. 685 p.

82. Sussmann HJ, Zahler RS. A critique of applied catastrophe theory in the behavioral sciences. Behav Sci 1978;23(4):383‑9.

83. Laurent M. Systèmes biologiques à dynamique non linéaire propriétés, analyse et modélisation. Paris: Ellipses; 2013. 360 p.

84. Bagley RJ, Glass L. Counting and Classifying Attractors in High Dimensional Dynamical Systems. J Theor Biol 1996;183(3):269‑84.

85. Ivancevic VG, Ivancevic TT. Geometrical Dynamics of Complex Systems. A Unified Modelling Approach to Physics, Control, Biomechanics, Neurodynamics and Psycho-Socio-Economical Dynamics. New York : Springer; 2006. 824 p.

86. Luke DA, Stamatakis KA. Systems science methods in public health: dynamics, networks, and agents. Annu Rev Public Health 2012;33:357‑76.

87. Takens F. Detecting strange attractors in turbulence. In: Rand D, Young LS, eds. Dynamical Systems and Turbulence, Warwick 1980. New York : Springer-Verlag; 1981. p. 366‑81.

88. Pecora LM, Moniz L, Nichols J, Carroll TL. A unified approach to attractor reconstruction. Chaos Interdiscip J Nonlinear Sci 2007;17(1):013110.

89. Marmot MG, Smith GD, Stansfeld S, Patel C, North F, Head J, et al. Health inequalities among British civil servants: the Whitehall II study. Lancet 1991;337(8754):1387‑93.

90. Goldberg M, Chevalier A, Imbernon E, Coing F, Pons H. The epidemiological information system of the French national electricity and gas company: the SI-EPI project. Med Lav 1996;87(1):16‑28.

91. Jayasinghe S. Conceptualising population health: from mechanistic thinking to complexity science. Emerg Themes Epidemiol 2011;8(1):2.

92. Bishop CM. Pattern recognition and machine learning. New York: Springer; 2006. 738 p.

93. Jain AK. Data clustering: 50 years beyond K-means. Pattern Recognit Lett 2010;31(8):651‑66.

94. Theodoridis S, Koutroumbas K. Pattern Recognition. 4th Revised edition. San Diego : Academic Press Inc; 2008. 984 p.

95. Kaufman L, Rousseeuw PJ. Finding groups in data : an introduction to cluster analysis. Hoboken, N.J.: Wiley; 2005. 368 p.

96. Tufféry S. Data mining et statistique décisionnelle. Paris: Éditions Technip; 2010. 705 p.

97. Lee JA, Verleysen M. Nonlinear dimensionality reduction. London: Springer; 2007. 309 p.

98. Izenman AJ. Modern multivariate statistical techniques : regression, classification, and manifold learning. New York : Springer; 2008. 733 p.

99. Tenenbaum JB, de Silva V, Langford JC. A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science 2000;290(5500):2319‑23.

100. Camastra F. Data Dimensionality Estimation Methods: A Survey. Pattern Recognit 2003;36:2945‑54.

101. Whitehead M, Dahlgren G. What can be done about inequalities in health? Lancet 1991;338(8774):1059‑63.

102. OMS | Commission des déterminants sociaux de la santé - rapport final [Internet]. [consulté le 3 juin 2015]. Disponible sur: http://www.who.int/social_determinants/thecommission/finalreport/fr/.

103. Chauvin P, Lebas J. Inégalités et disparités sociales de santé. In : Bourdillon F, Brücker G, Tabuteau D, eds. Traité de santé publique. 2ème édition revue et augmentée. Paris: Flammarion Médecine Sciences; 2007. p. 331‑41.

104. Evans RG, Barer ML, Marmor TR. Why are some people healthy and others not? : the determinants of health of populations. New York: A. de Gruyter; 1994. 402 p.

105. Karanikolos M, Mladovsky P, Cylus J, Thomson S, Basu S, Stuckler D, et al. Financial crisis, austerity, and health in Europe. Lancet 2013;381(9874):1323‑31.

106. Eckersley R, Dixon J, Douglas RM. The social origins of health and well-being. New York : Cambridge University Press; 2001. 368 p.

107. Cooper H. Investigating socio-economic explanations for gender and ethnic inequalities in health. Soc Sci Med 2002;54(5):693‑706.

108. Leclerc A, Fassin D, Grandjean H, Kaminski M, Lang T. Les inégalités sociales de santé. Paris: Editions la Découverte; 2010. 448 p.

109. Chauvin P. Précarisation sociale et état de santé : le renouvellement d'un paradigme épidémiologique. In: Lebas J, Chauvin P. Précarité et santé. Paris: Flammarion Médecine Sciences; 1998. p. 59‑74.

110. Lynch J, Kaplan G. Socioeconomic position. In: Berkman LF, Kawachi I Social epidemiology. New York: OUP; 2000. p. 13‑35.

111. Tellnes G, ed. Urbanisation and health : new challenges in health promotion and prevention. Oslo : Oslo Academic Press; 2005. 368 p.

112. Pearce N, Davey Smith G. Is social capital the key to inequalities in health? Am J Public Health 2003;93(1):122‑9.

113. Berney L. Lifecourse influences on health in old age. Understanding health inequalities. Open University Press. Buckingham: Graham H; 2000. p. 79‑95.

114. Saint Claire L. Rival truths: common sense and social psychological explanations in health an illness. New York : Psychology Press; 2003. 288 p.

115. Strecher V. The health belief model and health behavior. In: Gochman DS, ed. Handbook of health behaviour research, Personal and social determinants. New York: Springer-Verlag US; 1997. p. 71‑91.

116. Joubert M, Chauvin P, Facy F, Ringa V, eds. Précarisation, risques et santé. Paris: Editions Inserm; 2001.

117. Berkman LF, Melchior M, Chastang J-F, Niedhammer I, Leclerc A, Goldberg M. Social integration and mortality: a prospective study of French employees of Electricity of France-Gas of France: the GAZEL Cohort. Am J Epidemiol 2004;159(2):167‑74.

118. Chauvin P, Parizot I, Revet S. Santé et expériences de soins: De l'individu à l'environnement social. Inserm; 2005. 292 p.

119. Diez Roux AV. Investigating neighborhood and area effects on health. Am J Public Health 2001;91(11):1783‑9.

120. Frumkin H, Frank LD, Jackson R. Urban Sprawl and Public Health: Designing, Planning, and Building for Healthy Communities. Chicago: Island Press; 2004. 366 p.

121. Navarro V. The Political and Social Contexts of Health. Amityville: Baywood Publishing Company, Inc.; 2004. 250 p.

122. Dean H. Welfare rights and social policy. Harlow: Prentice Hall; 2002. 272 p.

123. Lombrail P. Accès aux soins. In: Leclerc A, Fassin D, Grandjean H et al Les inégalités sociales de santé. Paris: Inserm/La Découverte; 2000. p 403‑18.

124. Andersen RM. Revisiting the behavioral model and access to medical care: does it matter? J Health Soc Behav 1995;36(1):1‑10.

125. Babitsch B, Gohl D, von Lengerke T. Re-revisiting Andersen's Behavioral Model of Health Services Use: a systematic review of studies from 1998-2011. Psychosoc Med 2012;9:Doc11.

126. Chauvenet A. Médecines au choix, médecine de classes. Paris: Presses universitaires de France; 1978. 255 p.

127. Herzlich C, Moscovici S. Santé et maladie analyse d'une représentation sociale. Paris: École des hautes études en sciences sociales; 1969. 210 p.

128. Ehrenberg A. La fatigue d'être soi. Dépression et société. Paris: O. Jacob; 1999. 318 p.

129. Mackenbach JP. The persistence of health inequalities in modern welfare states: The explanation of a paradox. Soc Sci Med 2012;75(4):761‑9.

130. Montaut A. Santé et recours aux soins des femmes et des hommes, Premiers résultats de l'enquête Handicap-Santé 2008. Etudes et Résultats, DREES. 2010;717.

131. BEH n°2-3-4 Numéro thématique - Santé et recours aux soins des migrants en France [Internet]. 2012 [consulté le 4 juin 2015]. Disponible sur: http://www.invs.sante.fr/Publications-et-outils/BEH-Bulletin-epidemiologique-hebdomadaire/Archives/2012/BEH-n-2-3-4-2012.

132. Jusot F, Berchet C. État de santé et recours aux soins des immigrés en France: une revue de la littérature. In: BEH n°2-3-4 Numéro thématique - Santé et recours aux soins des migrants en France [Internet]. 2012. p 17-21. [consulté 4 mai 2015]. Disponible sur: http://www.invs.sante.fr/Publications-et-outils/BEH-Bulletin-epidemiologique-hebdomadaire/Archives/2012/BEH-n-2-3-4-2012.

133. Rondet C, Lapostolle A, Soler M, Grillo F, Parizot I, Chauvin P. Are immigrants and nationals born to immigrants at higher risk for delayed or no lifetime breast and cervical cancer screening? The results from a population-based survey in Paris metropolitan area in 2010. PloS One. 2014;9(1):e87046.

134. Boisguérin B. État de santé et recours aux soins des bénéficiaires de la CMU, Un impact qui se consolide entre 2000 et 2003. Etudes et Résultats, DREES. 2004;294.

135. Morin T. Classification des dépressifs selon leur type de recours aux soins. Etudes et Résultats, DREES. 2007;577.

136. Gouyon M. Une typologie des recours urgents ou non programmés à la médecine de ville. Dossiers solidarité et santé, les professions de santé et leurs pratiques. DREES; 2006.

137. Gouyon M. L'enquête sur le recours au spécialiste en médecine de ville en 2007. DREES, document de travail; 2010.

138. Chapireau F. Les recours aux soins spécialisés en santé mentale. Etudes et Résultats, DREES. 2006;533.

139. Chevreul K, Durand-Zaleski I, Bahrami S, Hernández-Quevedo C, Mladovsky P. France: Health system review. Health Syst Transit 2010;12(6):1‑219.

140. Rodwin VG. The health care system under French national health insurance: lessons for health reform in the United States. Am J Public Health 2003;93(1):31‑7.

141. HCSP. Evaluation du plan psychiatrie et santé mentale 2005-2008 [Internet]. 2011 [consulté 4 juin 2015]. Disponible sur: http://www.hcsp.fr/docspdf/avisrapports/hcspr20111006_evalplapsysantementale.pdf.

142. Allonier C, Dourgnon P, Rochereau T. Enquête santé et protection sociale 2006, un panel pour l'analyse des politiques de santé, la santé publique et la recherche en économie de la santé. Institut de recherche et de documentation en économie de la santé IRDES. Paris; 2008.

143. Médecins du Monde. Access to healthcare for vulnerable populations. update of legislation in 10 European countries. [Internet]. 2013. [consulté 4 juin 2015]. Disponible sur: http://www.medecinsdumonde.org/Access-to-healthcare-in-Europe-in-times-of-crisis-and-rising-xenophobia.

144. La Convention internationale des droits de l'enfant | Unicef France [Internet]. 2014 [consulté le 4 juin 2015]. Disponible sur: http://www.unicef.fr/contenu/info-humanitaire-unicef/la-convention-internationale-des-droits-de-lenfant.

145. Insee - Population - Fiches thématiques - Population immigrée - Immigrés - Insee Références - Édition 2012 [Internet]. [consulté le 4 juin 2015]. Disponible sur: http://www.insee.fr/fr/themes/document.asp?reg_id=0&id=3713.

146. Debré I. Les mineurs isolés étrangers en France [Internet]. 2010 [consulté 4 juin 2015]. Disponible sur: https://documentation.outre-mer.gouv.fr/Record.htm?idlist=1&record=19103612124919218949.

147. Réseau européen des migrations - Politiques, pratiques et données statistiques sur les mineurs isolés étrangers en 2014 - InfoMIE.net [Internet]. 2015 [consulté le 4 juin 2015]. Disponible sur: http://infomie.net/spip.php?article2107.

148. Aspinwall LG, Brown TR, Tabery J. The Double-Edged Sword: Does Biomechanism Increase or Decrease Judges' Sentencing of Psychopaths? Science 2012;337(6096):846‑9.

149. Gill P. When DNA goes on trial. Nature 2009;460(7251):34‑5.

150. Gilbert N. Science in court: DNA's identity crisis. Nat News 2010;464(7287):347‑8.

151. Cressey D. UK forensic science slammed by inquiry. Nature [Internet]. 25 juill 2013 [consulté le 4 juin 2015]; Disponible sur: http://www.nature.com/news/uk-forensic-science-slammed-by-inquiry-1.13444?WT.ec_id=NEWS-20130730.

152. Collectif. Interdisons les tests d'âge osseux sur les jeunes immigrés. Le Monde.fr [Internet]. 17 janv 2015 [consulté le 4 juin 2015]; Disponible sur: www.lemonde.fr/idees/article/2015/01/17/interdisons-les-tests-d-age-osseux-sur-les-jeunes-immigres_4558355_3232.html.

153. Baumard M. Immigration : les députés maintiennent les tests osseux. Le Monde.fr [Internet]. 13 mai 2015 [consulté le 4 juin 2015]; Disponible sur: www.lemonde.fr/societe/article/2015/05/13/immigration-les-deputes-maintiennent-les-tests-osseux_4632905_3224.html.

154. Jolivet A. Migrations, santé et soins en Guyane [Thèse d'université]. Paris: Université Pierre et Marie Curie; 2014 [consulté le 4 juin 2015]. Disponible sur: http://www.theses.fr/2014PA066124.

155. Termes clés de la migration [Internet]. Organisation internationale pour les migrations. [consulté le 4 juin 2015]. Disponible sur: http://www.iom.int/fr/termes-cles-de-la-migration.

156. Zimmerman C, Kiss L, Hossain M. Migration and Health: A Framework for 21st Century Policy-Making. PLoS Med 2011;8(5):e1001034.

157. Kristiansen M, Mygind A, Krasnik A. Health effects of migration. Dan Med Bull 2007;54(1):46‑7.

158. Darmon N, Khlat M. An overview of the health status of migrants in France, in relation to their dietary practices. Public Health Nutr 2001;4(2):163‑72.

159. MIGREUROP R. Atlas des migrants en Europe: Géographie critique des politiques migratoires. Paris: Armand Colin; 2012. 144 p.

160. Steel Z, Silove D, Brooks R, Momartin S, Alzuhairi B, Susljik I. Impact of immigration detention and temporary protection on the mental health of refugees. Br J Psychiatry J Ment Sci 2006;188:58‑64.

161. Keller AS, Rosenfeld B, Trinh-Shevrin C, Meserve C, Sachs E, Leviss JA, et al. Mental health of detained asylum seekers. Lancet 2003;362(9397):1721‑3.

162. Silove D, Steel Z, Mollica R. Detention of asylum seekers: assault on health, human rights, and social development. Lancet 2001;357(9266):1436‑7.

163. Chariot P, Caussinus H. Age estimation in undocumented migrant adolescents: medical response to judicial authorities. Presse Médicale 2015;44(1):99‑100.

164. Focardi M, Pinchi V, De Luca F, Norelli G-A. Age estimation for forensic purposes in Italy: ethical issues. Int J Legal Med 2014;128(3):515‑22.

165. Rudolf E. Comments to Focardi et al., Age estimation for forensic purposes in Italy: ethical issues. Int J Legal Med 2014; doi: 10.1007/s00414-014-1043-8 [Epub ahead of print].

166. Pruvost M-O, Boraud C, Chariot P. Skeletal age determination in adolescents involved in judicial procedures: from evidence-based principles to medical practice. J Med Ethics 2010;36(2):71‑4.

167. Schmeling A, Olze A, Reisinger W, Geserick G. Age estimation of living people undergoing criminal proceedings. Lancet 2001;358(9276):89‑90.

168. Schmeling A, Reisinger W, Geserick G, Olze A. Age estimation of unaccompanied minors. Part I. General considerations. Forensic Sci Int 2006;159 Suppl 1:S61‑4.

169. Scientific Programme | Congress of the International Academy of Legal Medicine [Internet]. [consulté le 5 mai 2015]. Disponible sur: http://www.ialmdubai.ae/scientific-programme/.

170. AGFAD. Study Group on Forensic Age Diagnostics of the German Association of Forensic Medicine [Internet]. [consulté le 4 juin 2015]. Disponible sur: http://agfad.uni-muenster.de/english/empfehlungen.htm.

171. Rudolf E, Kramer J, Gebauer A, Bednar A, Recsey Z, Zehetmayr J, et al. Standardized medical age assessment of refugees with questionable minority claim-a summary of 591 case studies. Int J Legal Med 2015;129(3):595‑602.

172. Schmeling A, Grundmann C, Fuhrmann A, Kaatsch H-J, Knell B, Ramsthaler F, et al. Criteria for age estimation in living individuals. Int J Legal Med 2008;122(6):457‑60.

173. Olze A, Reisinger W, Geserick G, Schmeling A. Age estimation of unaccompanied minors. Part II. Dental aspects. Forensic Sci Int 2006;159 Suppl 1:S65‑7.

174. Maber M, Liversidge HM, Hector MP. Accuracy of age estimation of radiographic methods using developing teeth. Forensic Sci Int 2006;159 Suppl 1:S68‑73.

175. Kiran CS, Reddy RS, Ramesh T, Madhavi NS, Ramya K. Radiographic evaluation of dental age using Demirjian's eight-teeth method and its comparison with Indian formulas in South Indian population. J Forensic Dent Sci 2015;7(1):44‑8.

176. Demirjian A, Goldstein H, Tanner JM. A new system of dental age assessment. Hum Biol 1973;45(2):211‐27.

177. Schmidt S, Nitz I, Ribbecke S, Schulz R, Pfeiffer H, Schmeling A. Skeletal age determination of the hand: a comparison of methods. Int J Legal Med 2013;127(3):691‐8.

178. Schmidt S, Koch B, Schulz R, Reisinger W, Schmeling A. Studies in use of the Greulich-Pyle skeletal age method to assess criminal liability. Leg Med Tokyo Jpn 2008;10(4):190‐5.

179. Mansourvar M, Ismail MA, Raj RG, Kareem SA, Aik S, Gunalan R, et al. The applicability of Greulich and Pyle atlas to assess skeletal age for four ethnic groups. J Forensic Leg Med 2014;22:26‐9.

180. Wittschieber D, Schulz R, Vieth V, Küppers M, Bajanowski T, Ramsthaler F, et al. Influence of the examiner's qualification and sources of error during stage determination of the medial clavicular epiphysis by means of computed tomography. Int J Legal Med 2014;128(1):183‐91.

181. Kellinghaus M, Schulz R, Vieth V, Schmidt S, Schmeling A. Forensic age estimation in living subjects based on the ossification status of the medial clavicular epiphysis as revealed by thin-slice multidetector computed tomography. Int J Legal Med 2010;124(2):149‐54.

182. Schulz R, Mühler M, Reisinger W, Schmidt S, Schmeling A. Radiographic staging of ossification of the medial clavicular epiphysis. Int J Legal Med 2008;122(1):55‐8.

183. Bassed RB, Briggs C, Drummer OH. Age estimation using CT imaging of the third molar tooth, the medial clavicular epiphysis, and the spheno-occipital synchondrosis: a multifactorial approach. Forensic Sci Int 2011;212(1-3):273.e1‐5.

184. Hulst R van. A Statistically Significant Future for Bayes' Rule. Science 2013;341(6144):343‐343.

185. Efron B. Bayes' Theorem in the 21st Century. Science 2013;340(6137):1177‐8.

186. Schmeling A, Olze A, Reisinger W, König M, Geserick G. Statistical analysis and verification of forensic age estimation of living persons in the Institute of Legal Medicine of the Berlin University Hospital Charité. Leg Med Tokyo Jpn 2003;5 Suppl 1:S367‑71.

187. Ontell FK, Ivanovic M, Ablin DS, Barlow TW. Bone age in children of diverse ethnicity. AJR Am J Roentgenol 1996;167(6):1395‑8.

188. Loder RT, Estle DT, Morrison K, Eggleston D, Fish DN, Greenfield ML, et al. Applicability of the Greulich and Pyle skeletal age standards to black and white children of today. Am J Dis Child 1960 1993;147(12):1329‑33.

189. Olze A, Taniguchi M, Schmeling A, Zhu B-L, Yamada Y, Maeda H, et al. Comparative study on the chronology of third molar mineralization in a Japanese and a German population. Leg Med Tokyo Jpn 2003;5 Suppl 1:S256‑60.

190. Clarot F, Le Dosseur P, Vaz E, Proust B. Skeletal maturation and ethnicity. Leg Med 2004;6(2):141‑2.

191. Meijerman L, Maat GJR, Schulz R, Schmeling A. Variables affecting the probability of complete fusion of the medial clavicular epiphysis. Int J Legal Med 2007;121(6):463‑8.

192. Etiemble A. Les mineurs isolés étrangers en France : évaluation quantitative de la population accueillie à l'Aide sociale à l'Enfance, les termes de l'accueil et de la prise en charge [Internet]. Rennes: Quest'us; 2002. 272 p. Disponible sur: http://infomie.net/spip.php?article12.

193. Thom R. Prédire n'est pas expliquer. Paris: Flammarion; 2009. 171 p.

194. Zabet D, Rérolle C, Pucheux J, Telmon N, Saint-Martin P. Can the Greulich and Pyle method be used on French contemporary individuals? Int J Legal Med 2015;129(1):171‑7.

195. Saint-Martin P, Rérolle C, Dedouit F, Bouilleau L, Rousseau H, Rougé D, et al. Age estimation by magnetic resonance imaging of the distal tibial epiphysis and the calcaneum. Int J Legal Med 2013;127(5):1023‑30.

196. Schulz R, Schiborr M, Pfeiffer H, Schmidt S, Schmeling A. Forensic age estimation in living subjects based on ultrasound examination of the ossification of the olecranon. J Forensic Leg Med 2014;22:68‑72.

197. Descartes R. Discours de la méthode. Flammarion; 2000. 189 p.

198. Bernard C. Introduction à l'étude de la médecine expérimentale. Flammarion; 2008. 381 p.

199. Canguilhem G. Le normal et le pathologique. 11e éd. Presses Universitaires de France; 2009. 240 p.

200. Morin E. On complexity. Cresskill: Hampton Press; 2008.

201. Bertalanffy L von. Théorie générale des systèmes. Paris: Dunod; 2012. 328 p.

202. Deisboeck TS, Kresh JY. Complex systems science in biomedicine. New York: Springer; 2006. 864 p.

203. Galea S, Riddle M, Kaplan GA. Causal thinking and complex system approaches in epidemiology. Int J Epidemiol 2009;39(1):97‑106.

204. Lorenz EN. Deterministic Nonperiodic Flow. J Atmospheric Sci 1963;20(2):130‑41.

205. West BJ. Fractal Physiology and the Fractional Calculus: A Perspective. Front Physiol 2010(14);1:12.

206. Lee J, Lee D. An improved cluster labeling method for support vector clustering. IEEE Trans Pattern Anal Mach Intell 2005;27(3):461‑4.

207. Yilmaz O, Achenie LEK, Srivastava R. Systematic tuning of parameters in support vector clustering. Math Biosci 2007;205(2):252‑70.

208. Fahim AM, Saake G, Salem AM, Torkey FA, Ramadan MA. K-Means for Spherical Clusters with Large Variance in Sizes. WASET 2008;45:177-182.

209. Armstrong JJ, Zhu M, Hirdes JP, Stolee P. K-Means Cluster Analysis of Rehabilitation Service Users in the Home Health Care System of Ontario: Examining the Heterogeneity of a Complex Geriatric Population. Arch Phys Med Rehabil 2012;93(12):2198-205.

210. Simpson TI, Armstrong JD, Jarman AP. Merged consensus clustering to assess and improve class discovery with microarray data. BMC Bioinformatics 2010;11(1):590.

211. Monti S, Tamayo P, Mesirov J, Golub T. Consensus clustering – A resampling-based method for class discovery and visualization of gene expression microarray data. Machine learning, functional genomics special issue 2003;91‑118.

212. Laguës M, Lesne A. Invariances d'échelle: des changements d'états à la turbulence. Paris: Belin; 2008. 367 p.

213. Ward JS, Barker A. Undefined By Data: A Survey of Big Data Definitions. ArXiv13095821 Cs [Internet]. 2013 [consulté le 5 juin 2015]; Disponible sur: http://arxiv.org/abs/1309.5821.

214. Wang C, Chen M-H, Schifano E, Wu J, Yan J. A Survey of Statistical Methods and Computing for Big Data. ArXiv150207989 Math Stat [Internet]. 2015 [consulté le 5 juin 2015]; Disponible sur: http://arxiv.org/abs/1502.07989.

215. Al-Jarrah OY, Yoo PD, Muhaidat S, Karagiannidis GK, Taha K. Efficient Machine Learning for Big Data: A Review. ArXiv150305296 Cs [Internet]. 2015 [consulté le 5 juin 2015]; Disponible sur: http://arxiv.org/abs/1503.05296.

216. Wolfe PJ. Making sense of big data. Proc Natl Acad Sci USA 2013;110(45):18031‑2.

217. Chiolero A. Big data in epidemiology: too big to fail? Epidemiology 2013;24(6):938‑9.

218. Laney D. 3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety [Internet]. 2001 [consulté le 5 juin 2015] Disponible sur: http://blogs.gartner.com/doug-

laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

219. Bellman R. Dynamic programming. Mineola: Dover Publications; 2003. 384 p.

220. Roweis ST, Saul LK. Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science 2000;290(5500):2323‑6.

221. Zhang Z, Zha H. Principal Manifolds and Nonlinear Dimensionality Reduction via Tangent Space Alignment. SIAM J Sci Comput 2004;26(1):313‑38.

222. Gorban AN. Principal manifolds for data visualization and dimension reduction. Berlin: Springer; 2008. 364 p.

223. Nadler B, Lafon S, Coifman RR, Kevrekidis IG. Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. Appl Comput Harmon Anal 2006;21(1):113‑27.

224. Hinton GE, Salakhutdinov RR. Reducing the Dimensionality of Data with Neural Networks. Science 2006;313(5786):504‑7.

225. Cox TF, Cox MAA. Multidimensional scaling. Boca Raton: Chapman & Hall/CRC; 2001. 328 p.

226. Pearson K. On lines and planes of closest fit to systems of points in space. Philos Mag. 1901;2(6):559‑72.

227. Saporta G. Probabilités, analyse des données et statistique. 3e édition révisée. Paris: Éditions Technip; 2011. 622 p.

228. Donoho DL. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. Proc Natl Acad Sci USA 2003;100(10):5591‑6.

229. Maaten LJP van der, Postma EO, Herik HJ van den. Dimensionality Reduction: A Comparative Review. [Rapport technique] Maastricht: Maastricht university; 2008. 22 p.

230. Weng S, Zhang C, Lin Z, Zhang X. Mining the structural knowledge of high-dimensional medical data using isomap. Med Biol Eng Comput 2005;43(3):410‑2.

231. Shi J, Luo Z. Nonlinear dimensionality reduction of gene expression data for visualization and clustering analysis of cancer tissue samples. Comput Biol Med 2010;40(8):723‑32.

232. Gorban AN, Zinovyev A. Principal manifolds and graphs in practice: from molecular biology to dynamical systems. Int J Neur Syst 2010;20(03):219-32.

233. Préteceille E. La division sociale de l'espace francilien [Internet]. Paris : Observatoire Sociologique du Changement - Sciences Po & CNRS; 2003 [consulté le 6 mai 2015]. Disponible sur: https://halshs.archives-ouvertes.fr/halshs-00130291/document.

234. Chauvin P, Parizot I. Vulnérabilités sociales, santé et recours aux soins dans les quartiers défavorisés franciliens. Editions de la DIV. Paris; 2007. 150 p.

235. Chauvin P, Parizot I. Les inégalités sociales et territoriales de santé dans l'agglomération parisienne : une analyse de la cohorte SIRS. Editions de la DIV. Paris; 2009. 105 p.

236. Mackenbach JP, Kunst AE, Cavelaars AE, Groenhof F, Geurts JJ. Socioeconomic inequalities in morbidity and mortality in western Europe. The EU Working Group on Socioeconomic Inequalities in Health. Lancet 1997;349(9066):1655‑9.

237. Marmot MG, Wilkinson RG. Social determinants of health. New York: Oxford University Press; 2006. 376 p.

238. Smith PC. Universal health coverage and user charges. Health Econ Policy Law 2013;8(4):529‑35.

239. Gusmano MK, Weisz D, Rodwin VG, Lang J, Qian M, Bocquier A, et al. Disparities in access to health care in three French regions. Health Policy 2014;114(1):31‑40.

240. Watanabe R, Hashimoto H. Horizontal inequity in healthcare access under the universal coverage in Japan; 1986-2007. Soc Sci Med 2012;75(8):1372‑8.

241. Glazier RH, Agha MM, Moineddin R, Sibley LM. Universal health insurance and equity in primary care and specialist office visits: a population-based study. Ann Fam Med 2009;7(5):396‑405.

242. Wang T-F, Shi L, Nie X, Zhu J. Race/ethnicity, insurance, income and access to care: the influence of health status. Int J Equity Health. 2013;12:29.

243. Kawachi I. Social Ties and Mental Health. J Urban Health Bull N Y Acad Med 2001;78(3):458‑67.

244. Bourdieu P. Questions de sociologie. Paris: Editions de Minuit; 2002. 288 p.

245. Riou J, Lefèvre T, Parizot I, Lhuissier A, Chauvin P. Is there still a French eating model? A taxonomy of eating behaviors in adults living in the Paris metropolitan area in 2010. PloS One 2015;10(3):e0119161.

246. Pearl J. Probabilistic reasoning in intelligent systems : networks of plausible inference. San Mateo: Morgan Kaufmann Publishers; 1991. 552 p.

247. Pearl J. Causality : models, reasoning, and inference. Cambridge: Cambridge University Press; 2010. 400 p.

248. Hopfield JJ. Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci USA 1982;79(8):2554‑8.

249. Erb RJ. The backpropagation neural network--a Bayesian classifier. Introduction and applicability to pharmacokinetics. Clin Pharmacokinet 1995;29(2):69‑79.

250. Hernández-Rabadán DL, Ramos-Quintana F, Guerrero Juk J. Integrating SOMs and a Bayesian classifier for segmenting diseased plants in uncontrolled environments. ScientificWorldJournal 2014;2014:214674.

251. Pernkopf F, Wohlmayr M. Stochastic margin-based structure learning of Bayesian network classifiers. Pattern Recognit 2013;46(2):464‑71.

252. Belabbas M-A, Wolfe PJ. Spectral methods in machine learning and new strategies for very large datasets. Proc Natl Acad Sci USA 2009;106(2):369‑74.

253. Jones N. Computer science: The learning machines. Nature 2014;505(7482):146‑8.

254. Pe'er D. Bayesian Network Analysis of Signaling Networks: A Primer. Sci STKE 2005;2005(281):pl4‑pl4.

255. Lee SY. Basic and advanced structural equation models for medical and behavioural sciences. Hoboken : Wiley; 2012. 367 p.

256. Bourdieu P. La distinction. Les Editions de Minuit; 1979. 672 p.

257. Goldberger JJ. Personalized Medicine vs Guideline-Based Medicine. JAMA 2013;309(24):2559.

258. Ioannidis JA. More than a billion people taking statins?: Potential implications of the new cardiovascular guidelines. JAMA 2013;311(5):463-4.

259. Nebert DW, Zhang G. Personalized Medicine: Temper Expectations. Science 2012;337(6097):910‑910.

260. Baedke J. The epigenetic landscape in the course of time: Conrad Hal Waddington's methodological impact on the life sciences. Stud Hist Philos Sci Part C Stud Hist Philos Biol Biomed Sci 2013;44(4, Part B):756‑73.

261. Wang J, Zhang K, Xu L, Wang E. Quantifying the Waddington landscape and biological paths for development and differentiation. Proc Natl Acad Sci USA 2011;108(20):8257‑62.

Annexes

Annexe 1. Questionnaire SIRS - questions spécifiques à l'étude

Questions liées aux recours aux soins

D9. A quand remonte votre dernière visite chez un dentiste pour un simple contrôle, c'est-à-dire pour voir si tout va bien, sans avoir mal ni de problèmes particuliers ?
*!! **Consigne :** un simple détartrage est considéré comme un contrôle*

- 1 an ou moins ☐ → *Passez à* **D10**
- entre 1 et 2 ans ☐
- entre 2 et 3 ans ☐
- 3 ans ou plus ☐
- Vous n'avez jamais eu ce type de consultation ☐

} *Passez à* **D9B**

" Nous allons parler à présent du recours à des soins de santé "

F4. Que vous l'ayez ou non déclaré comme médecin traitant, avez-vous un médecin régulier qui vous connaît déjà et que vous allez consulter en priorité si vous êtes malade (médecin de famille ou médecin de quartier par exemple) ?

- Oui ☐ → *Passez à* **F5** *puis* **F6**
- Non ☐ → *Passez à* **F6**

F10. Au cours des 12 derniers mois, avez-vous eu un bilan de santé dans le cadre de la Sécurité Sociale ?

- Oui ☐
- Non ☐
- *(Ne sait pas)* ☐

F11. Au cours des 12 derniers mois, combien de fois avez-vous été pour **vous-même** aux Urgences d'un hôpital ou d'une clinique ?
*!! **Consigne :** compter le nombre total de visites même si elles concernaient un même problème de santé*

☐☐ *fois au cours des 12 derniers mois*

F12. Au cours des 12 derniers mois, combien de fois avez-vous vu **pour vous-même** et **en urgence** un médecin à votre domicile qu'il s'agisse de SOS médecins, du SAMU, des pompiers, etc. ?

☐☐ *fois au cours des 12 derniers mois*

Si oui : **F16A.** Quels médecins spécialistes avez-vous consultés **au cours des 12 derniers mois** ? Etait-ce..	Oui	Non	**F16B.** ***Si oui,*** combien de fois au cours des 12 derniers mois **(pour vous-même)**... *!! Enquêteur : présentez carton F16B*							
			En ville dans un cabinet privé	En ville dans un dispensaire	A l'hôpital en secteur privé	A l'hôpital en secteur public	A l'hôpital mais ne sait pas dans quel secteur	Dans une clinique privée	A domicile	Lieu inconnu ou autre
un gynécologue	☐	☐								
un cardiologue	☐	☐								
un dermatologue	☐	☐								
un ORL	☐	☐								
un rhumatologue	☐	☐								
un diabétologue	☐	☐								
un psychiatre	☐	☐								
un autre spécialiste (*Précisez* →) _____	☐	☐								
un autre spécialiste (*Précisez* →) _____	☐	☐								
un autre spécialiste (*Précisez* →) _____	☐	☐								
un autre spécialiste (*Précisez* →) _____	☐	☐								

F17. Toujours au cours des 12 derniers mois et pour **vous-même**, en dehors d'une consultation à proprement parler, avez-vous demandé un avis médical à un proche ou un ami qui est médecin ?

• Non ..... ☐ • 1 ou 2 fois ..... ☐ • entre 3 et 10 fois..... ☐ • plus de 10 fois ..... ☐

F18. Enfin, au cours des 12 derniers mois, avez-vous consulté pour vous-même…

… un ostéopathe ou un acupuncteur ?

• Oui..... ☐ • Non..... ☐ • *(Déjà cité en F16 ou F14)* ..... ☐

… d'autres médecines parallèles ou alternatives ?

• Oui..... ☐ • Non..... ☐ • *(Déjà cité en F16 ou F14)* ..... ☐

Questions liées aux variables explicatives

A- FICHE MENAGE

Pouvez-vous énumérer toutes les personnes qui résident habituellement dans votre logement, en précisant leur prénom et quelques caractéristiques

	1	2	3	4	5	6	7	8	9
A1. Prénom									
A2. Sexe Homme / Femme	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐
A3. Lien avec la personne de référence (=chef de ménage) **1.** personne de référence (PR) **2.** conjoint de PR **3.** enfant de PR ou du conjoint **4.** père / mère de PR ou du conjoint **5.** autre parent de PR ou du conjoint **6.** salarié logé ou domestique **7.** ami, non apparenté	☐	☐	☐	☐	☐	☐	☐	☐	☐
A4. Age en années révolues *(si moins d'un an codez 0)*	☐☐	☐☐	☐☐	☐☐	☐☐	☐☐	☐☐	☐☐	☐☐
A5. Niveau d'études **1.** jamais scolarisé **2.** maternelle ou primaire **3.** secondaire **4.** supérieur	☐	☐	☐	☐	☐	☐	☐	☐	☐
A6. Occupation actuelle **1.** exerce un emploi **2.** apprenti / stage **3.** élève / étudiant **4.** chômeur **5.** retraité **6.** au foyer **7.** congé parental temps plein **8.** autre : *précisez sur les pointillés*	☐ ---------- ---------- ☐☐	☐ ---------- ---------- ☐☐	☐ ---------- ---------- ☐☐	☐ ---------- ---------- ☐☐	☐ ---------- ---------- ☐☐	☐ ---------- ---------- ☐☐	☐ ---------- ---------- ☐☐	☐ ---------- ---------- ☐☐	☐ ---------- ---------- ☐☐
A7. pour les enfants de moins de 18 ans **1.** tout le temps dans le ménage **2.** en garde alternée	☐	☐	☐	☐	☐	☐	☐	☐	☐

B1. Sexe de la personne sélectionnée ***(= sexe indiqué en A2)***

• Homme ..... ☐ • Femme ..... ☐

D2. Comment est votre état de santé physique ?

• Très bon..... ☐ • Bon ..... ☐ • Moyen ..... ☐ • Mauvais ..... ☐ • Très mauvais .. ☐

D4. Êtes-vous limité depuis au moins 6 mois à cause d'un problème de santé, dans les activités que les gens font habituellement ?
*!! **Consigne :** problème de santé physique ou psychologique*
*!! **Consigne :** pour les femmes enceintes, ne considérer que les limitations dues aux grossesses pathologiques*

• Oui, fortement limité..... ☐ • Oui, limité, mais pas fortement ... ☐ • Non, pas limité du tout..... ☐

D14. Avez-vous une maladie ou un problème de santé qui soit chronique ou de caractère durable ?
*!! **Consigne :** Une maladie chronique est une maladie qui a duré ou peut durer pendant une période de **6 mois ou plus***

• Oui ..... ☐ • Non ..... ☐ • *(Ne sait pas)* ... ☐

D24. Avez-vous ou avez-vous eu récemment, dans votre entourage proche, des personnes atteintes d'une maladie grave ?

• Oui ..... ☐
• Non ..... ☐

" Concernant votre couverture maladie... "

E1. Actuellement, avez-vous une couverture maladie ?

• Oui, la Sécurité sociale standard ..... ☐
• Oui, la Sécurité sociale **de base** par le biais de la CMU (Couverture maladie universelle) ..... ☐
• Oui, l'aide médicale d'Etat (AME) ..... ☐
} *Passez à* **E2A**

• Non, aucune couverture maladie ..... ☐
• *(Ne sait pas)* ..... ☐
} *Passez à* **F1**

E2A. Bénéficiez-vous d'une couverture maladie complémentaire ?

• Oui, par la CMU (Couverture Maladie Universelle) ..... ☐
• Oui, par une mutuelle ou une assurance privée..... ☐
• Non, aucune ..... ☐
• *(Ne sait pas)* ..... ☐

E2B. Êtes-vous pris en charge à 100% **pour raison médicale** par la **Sécurité sociale** ?

• Oui ..... ☐ → *Passez à* **E3**
• Non ..... ☐
• *(Ne sait pas)* ..... ☐
} *Passez à* **F1**

F1. En général, vous consultez un médecin **généraliste**… *(les modalités 2 et 3 peuvent éventuellement être cochées toutes les deux)*
!! Consigne : *en dehors du suivi gynéco éventuellement fait par un généraliste*

1. Uniquement lorsque vous ne pouvez plus faire autrement ..... ☐
2. Dès que vous ne vous sentez pas bien ..... ☐
3. Régulièrement pour un suivi ou pour voir si tout va bien ..... ☐

F2. Avez-vous dans votre entourage une personne qui est médecin ou qui travaille dans le domaine de la santé et qui pourrait vous conseiller ou vous orienter en cas de besoin ?

• Oui..... ☐ → *Passez à* **F3** • Non..... ☐ → *Passez à* **F4**

O4A. En cas de besoin, est-ce que vous pourriez compter sur quelqu'un, qu'il s'agisse de membres de votre ménage, de membres de votre famille, d'amis, de collègues ou de voisins, pour…				**O4B.** ***Si oui,*** Pourriez-vous compter pour cela sur des…							
				Membres du ménage		Autres membres de la famille (hors ménage)		Amis, collègues		Voisins	
	Oui	Non		Oui	Non	Oui	Non	Oui	Non	Oui	Non
• Vous aider dans la vie quotidienne, vous donner un coup de main (*)… ?	☐	☐	*Si oui* →	☐	☐	☐	☐	☐	☐	☐	☐
• Vous aider financièrement ou matériellement (nourriture, vêtements, etc.) ?	☐	☐	*Si oui* →	☐	☐	☐	☐	☐	☐	☐	☐
• Vous apporter un soutien moral ou affectif ? .	☐	☐	*Si oui* →	☐	☐	☐	☐	☐	☐	☐	☐

* ***!! Consigne :*** *vous aider à vous rendre quelque part, pour quelques menus travaux chez vous, garder vos enfants, etc.*

O11. D'une façon générale, vous diriez que vous vous sentez très seul, plutôt seul, plutôt entouré ou très entouré ?

• Très seul ..... ☐ • Plutôt entouré ..... ☐
• Plutôt seul ..... ☐ • Très entouré..... ☐

Annexes 2. Communications affichées et orales

Société française de médecine légale (séance du 9 mars 2015)

Lefèvre T, Chauvin P, Chariot P. Big data et médecine légale : est-il pertinent d'intégrer des sources d'information multiples ? Exemple de l'estimation de l'âge. Société Française de Médecine Légale, séance du 9 mars 2015. Communication orale.

23e congrès international de médecine légale (Dubaï, 2015)

Lefèvre T, Chauvin P, Chariot P. Age estimation in living persons: is the integration of multiple information sources relevant? The contribution of novel multivariate methods to forensic sciences. 23rd Congress of the International Academy of Legal Medicine, Dubai, Qatar, 19-21 janvier 2015. Communication orale.

Ecole doctorale Pierre Louis (St Malo, 2014)

Lefèvre T, Rondet C, Parizot I, Chauvin P. Application de techniques de *clustering* à des données de santé : les 4 profils d'utilisation du système de soins en Ile de France. Séminaire de l'école doctorale Pierre Louis, St Malo, 20-22 octobre 2014. Communication affichée, présentée oralement.

7e journée des doctorants et post-doctorants de l'IFR 65 (Paris, 2013)

Lefèvre T, Chauvin P. Techniques non linéaires de réduction de la dimension. Principes et revue brève. 7ème journée des doctorants et post-doctorants de l'IFR65, Paris, 19 juin 2013. Communication orale.

Ecole doctorale Pierre Louis (St Malo, 2012)

Lefèvre T, Chauvin P. Pour le respect de l'environnement et de la géométrie : techniques non linéaires de réduction de la dimension – mieux que l'ACP ? Séminaire de l'école doctorale Pierre Louis, St Malo, 8-10 octobre 2012. Communication affichée, présentée oralement.

Annexes 3. Articles publiés dans le cadre de la thèse

Annexe 3.1 Cadre général d'utilisation des techniques de *clustering*

Available online at
ScienceDirect
www.sciencedirect.com
Elsevier Masson France
EM|consulte
www.em-consulte.com
Revue d'Épidémiologie et de Santé Publique 63 (2015) 9–19
Revue d'Épidémiologie et de Santé Publique
Epidemiology and Public Health

Original article

A general framework for a reliable multivariate analysis and pattern recognition in high-dimensional epidemiological data, based on cluster robustness: A tutorial to enrich the epidemiologists' toolkit

Un cadre général pour la recherche de groupes homogènes à partir de données épidémiologiques de haute dimension, basé sur la robustesse des groupes : un tutoriel pour l'épidémiologiste

T. Lefèvre [a,b,c,d,e,*], P. Chauvin [a,b]

[a] *Inserm, UMR_S 1136, department of social epidemiology, Pierre-Louis institute of epidemiology and public health, 27, rue de Chaligny, 75012 Paris, France*
[b] *UMR_S 1136, UPMC université Paris 06, Sorbonne universités, 75646 Paris, France*
[c] *Inserm, UMR_S 1101, laboratory of medical information processing, 29609 Brest, France*
[d] *UMR_S 1101, université de Bretagne occidentale, 29609 Brest, France*
[e] *UMR_S 1101, institut Mines-Telecom, Telecom Bretagne, 29609 Brest, France*

Received 28 April 2014; accepted 1 December 2014
Available online 17 January 2015

Abstract

Background. – In an epidemiologist's toolbox, three main types of statistical tools can be found: means and proportions comparisons, linear or logistic regression models and Cox-type regression models. All these techniques have their own multivariate formulations, so that biases can be accounted for. Nonetheless, there is an entire set of natively massive multivariate techniques, which are based on weaker assumptions than classical statistical techniques are, and which seem to be underestimated or remain unknown to most epidemiologists. These techniques are used for pattern recognition or clustering – that is, for retrieving homogeneous groups in data without any a priori about these groups. They are widely used in connex domains such as genetics or biomolecular studies.

Methods. – Most clustering techniques require tuning specific parameters so that groups can be identified in data. A critical parameter to set is the number of groups the technique needs to discover. Different approaches to find the optimal number of groups are available, such as the silhouette approach and the robustness approach. This article presents the key aspects of clustering techniques (how proximity between observations is defined and how to find the number of groups), two archetypal techniques (namely the *k*-means and PAM algorithms) and how they relate to more classical statistical approaches.

Results. – Through a theoretical, simple example and a real data application, we provide a complete framework within which classical epidemiological concerns can be reconsidered. We show how to (i) identify whether distinct groups exist in data, (ii) identify the optimal number of groups in data, (iii) label each observation according to its own group and (iv) analyze the groups identified according to separate and explicative data. In addition, how to achieve consistent results while removing sensitivity to initial conditions is explained.

Conclusions. – Clustering techniques, in conjunction with methods for parameter tuning, provide the epidemiologist with substantial additional tools. They differ from the usual approaches based on hypothesis-testing because no assumptions are made on the data and these clustering techniques are natively multivariate.

Keywords: Cluster; Epidemiologic methods; Epidemiology; Hypothesis test; Multivariate analysis

Résumé

Position du problème. – Les épidémiologistes disposent essentiellement de trois grandes sortes d'outils pour traiter leurs données : les tests de comparaisons de moyenne et de proportions, les modèles de régression linéaire ou logistique et les modèles de survie type modèles de Cox. Tous

* Corresponding author. Inserm, UMR_S 1136, department of social epidemiology, Pierre-Louis institute of epidemiology and public health, 27, rue de Chaligny, 75012 Paris, France.
E-mail address: thomas.lefevre@inserm.fr (T. Lefèvre), pierre.chauvin@inserm.fr (P. Chauvin).

http://dx.doi.org/10.1016/j.respe.2014.12.017
0398-7620/

ces outils possèdent leur formulation multivariée, ce qui permet de contrôler un minimum les biais. Il existe cependant tout un ensemble de techniques nativement multivariées reposant sur des hypothèses moins fortes que les techniques statistiques classiques, et qui semblent demeurer sous-estimées ou mal connues. Ces techniques, dites de *clustering* ou de classification, sont utilisées pour l'identification de groupes homogènes à partir de données, et ce sans a priori sur ces groupes. Elles sont largement utilisées dans des domaines connexes à l'épidémiologie, comme la génétique.

Méthodes. – La majorité des techniques de *clustering* nécessitent l'ajustement de paramètres qui leur sont spécifiques. Un paramètre particulièrement critique est le nombre de groupes à découvrir dans les données. Différentes approches existent qui permettent de déterminer le nombre optimal de groupes à découvrir, comme l'approche par la silhouette ou par la robustesse. Les auteurs présentent ici les aspects principaux liés aux techniques de *clustering* (de quelle façon l'on définit la proximité entre deux observations, comment déterminer le nombre de groupes à découvrir), deux techniques archétypiques (les algorithmes des *k* moyennes et PAM) et comment les articuler aux méthodes statistiques plus classiques.

Résultats. – Nous proposons un cadre général de traitement des données à l'aide des techniques de *clustering* au travers d'un exemple théorique simple puis d'une application sur données réelles. Nous montrons comment (i) déterminer s'il existe des groupes distincts dans les données, (ii) déterminer le nombre optimal de groupes, (iii) labelliser chaque observation selon le groupe auquel elle appartient, (iv) analyser les groupes selon des données séparées, explicatives. Enfin, nous expliquons comment obtenir des groupes consistants en s'affranchissant des problèmes de sensibilité aux conditions initiales.

Conclusions. – L'utilisation conjointe de techniques de *clustering* et de méthodes d'ajustement des paramètres de ces techniques permet d'enrichir les outils classiques de l'épidémiologiste. Ces techniques sont nativement multivariées et diffèrent des approches statistiques basées sur les tests d'hypothèses en ce sens qu'elles ne nécessitent aucun a priori sur les données à étudier.

Mots clés : Analyse multivariée ; Cluster ; Épidémiologie ; Méthodes épidémiologiques ; Test d'hypothèse

1. Background

Epidemiologists, and more generally speaking researchers working in clinical medical fields, make extensive use of statistical tools, which have basically become the technical core of their activity. The classical testing approach is intended for experimental or quasi-experimental settings, and consists in more or less sophisticated hypothesis tests. To approach causality, multiple and relevant covariates have to be added to statistical models to control for biases and spread the overall variability over different factors. The trend for massive multivariate statistical testing peaked as genetics and variant studies or even genome-wide studies appeared, as well as the use of large data sets extracted from health information systems or cohort surveys, for which there can be no unique or even clear hypothesis, but which still have to be analyzed, reduced and made comprehensive.

Artificial intelligence as well as sensing [1], computer vision and imaging [2–5], have for several years contributed to the development of pattern recognition in data sets, also known as clustering techniques. For the last few years, such techniques have been increasingly used to process data in the biomolecular and genetics fields [6–11], in ecology [12] and biochemistry [13], but more rarely in medicine [14] and public health [15–17]. Another remarkable example is provided by the international ENCODE project, which used pattern recognition techniques massively in the systematic search for the functionality of ''junk'' DNA [18].

Pattern recognition asks a general question: are there, in a given data set, any shared specificities, which make people belong to the same homogeneous group? This key question addresses a wide category of problems (e.g., classification and stratification of populations). Most of the current literature covers new techniques, or variants of existing ones, and their applications to genetics. More recently, a number of authors have provided general statistical, mathematical grounds for clustering [19,20]. Nonetheless, to the best of our knowledge, no general framework has been suggested, allowing epidemiologists to easily make use of these powerful tools. We intend here to provide a few easy-to-use solutions, with a clear explanation, so that pattern recognition can enrich the epidemiologist's statistical toolbox.

Our purpose is not to provide an exhaustive review – several useful books have already accomplished this task (e.g., [21–23]) – but a possible, reliable method to get started with pattern recognition. We first present key aspects of classical techniques in pattern recognition, then we discuss data preparation and how to estimate optimal settings in a general fashion, and finally how to return to determinant analysis. The general diagram is provided in Fig. 1. A complete simple example is given for illustrative purposes.

Note that the literature discusses two different uses of pattern recognition techniques. One consists in ''machine learning'' [24]: a specific technique is trained on previously labeled data sets, so that it can be used as an automated classifier on fresh, newly acquired data sets. An example of their possible application is disease diagnosis, e.g., in pathological anatomy. In this case, there is a gold standard (e.g., the skills and knowledge of the pathologist) that is used to train and correct the technique. This application is not covered in this paper. The other use of pattern recognition techniques is presented herein: when no gold standard is available, when there is no ''external truth'' about the data set to explore, these techniques can be used to discover homogenous groups in data (if any exist).

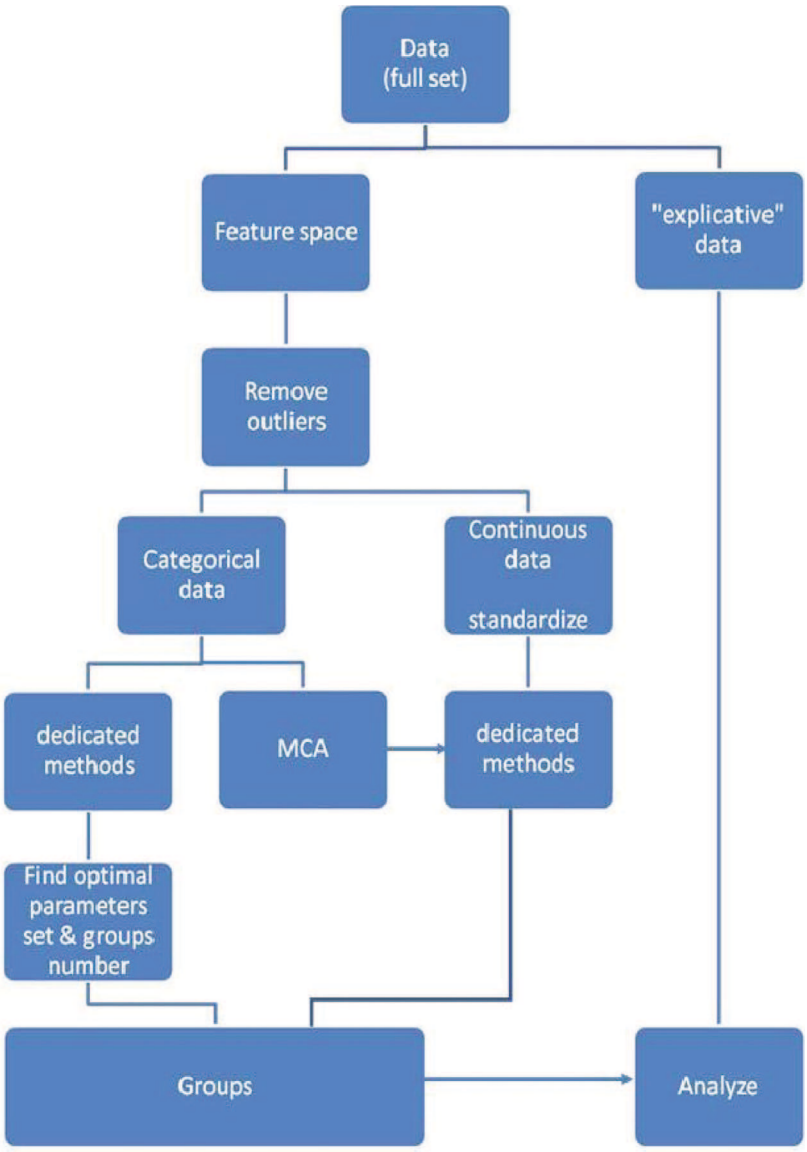


Fig. 1. A general framework for the multivariate search for patterns and their analysis. Initial data set may be split into two subsets, one consisting of the variables of interest (feature space) and the other "explicative" variables. The latter is used in classical analysis for association measure retrieval. MCA: multiple correspondence analysis.

2. Methods

2.1. General considerations

2.1.1. Assumptions

With the oldest dating from the late 1960s within a constantly evolving field, many different techniques share the common goal of separating people into groups, ideally homogeneous groups, on the basis of a given set of information. We will not expound on the technical mathematical details, easily found elsewhere if needed [21–23], but will emphasize some practical keys and issues, and the main specificities, limitations and advantages of a few carefully selected methods. A more detailed and complementary review of clustering can be found in [25]. All the techniques presented in this paper are based on only two assumptions: such groups do exist in data and these groups can be separated into a "flat" space. By "flat" we mean that no specific intrinsic geometry of the data set is taken into account: any point of the data sets can be reached in the same way, with isotropy.

2.1.2. Algorithm parameters

In the absence of fully automatic methods (which would seem dubious to many people, given the great diversity in clustering problems), different kinds of parameters have to be tuned depending on which method is chosen, answering the two following questions.

2.1.2.1. How many groups need to be found?. Some techniques are conceived so that the number of groups (clusters) in data is "automatically" determined, as provided by neural network-based techniques (e.g., self-organizing maps [SOM] or self-organizing tree algorithms [SOTA]) [26,27]. In

most techniques, this is not the case and the number of groups to retrieve in data has to be specified as a parameter.

2.1.2.2. How close are two individuals?. There is no magic in data mining: the best that can be done is to constrain the fewest possible data and make the fewest possible assumptions. Since this is a question of proximity, clustering techniques require the use of different distances or, sometimes, of (dis)similarity measures. The most frequently used distances are the Euclidean and the Manhattan distances (also called L2 and L1 distances). The Euclidean distance corresponds to the classical, everyday and "real life" distance, and involves square values according to the Pythagoras theorem. The Manhattan distance is named after how distance is computed in a grid-patterned city such as Manhattan, by adding segments of streets, rather than distances as the crow flies, from one point to another. Manhattan distances usually lead to crisper results than Euclidean distances, but the two are normally not inconsistent with one another. The use of one rather than the other essentially stems from the kind of data to explore. If data resemble data traffic, with strong geographical constraints such as roads, the Manhattan distance seems to fit reality better. If data resemble birds' migration, the Euclidean distance may be more appropriate. More generally speaking, continuous variables are likely to be treated with the Euclidean distance. Indeed, the most delicate problem concerning the way to assess proximity between two persons or observations rises when it comes to categorical data, or even worse, mixed data (categorical and continuous data). In these cases, using a (dis)similarity-based method can be an elegant solution. Several (dis)similarity measures for categorical data have been suggested and reviewed in the literature [28,29].

Here, for a relatively simple and straightforward introduction, we will present and use geometry-based or similarity-based techniques (i.e., based on the use of a distance or similarity measure) only. One should be aware that other techniques are available, which do not exactly rely on similarity assumptions. We discuss two of them: self-organizing maps (SOM, a version of neural networks) [23,26] and support vector clustering (SVC, a version based on SVM, support vector machines) [30,31]. These techniques are useful and powerful and can circumvent some of the limitations of the algorithms presented, but they also have drawbacks that prevent us from presenting them here. For example, in SVC, the complexity of the SVM usually combines with the complexity of network algorithms: when applying SVM techniques to a classification problem, SVC requires the choice of a "kernel" adapted to the type of data to process, which needs fine parameter tuning.

2.1.3. Sensitivity to initial conditions

Given a run of the method on our data set and given a set of parameters, will we obtain the same (qualitative) results if we run it again? Are the groups stable, reproducible from one run to another? A single run of a given technique can lead to a marginal result, i.e., a result that is not representative of the data's underlying structure. Techniques may be more or less sensitive to initial conditions, but two facts are worth remembering here:

- it is rare, perhaps even impossible, to analyze data sets that are strange enough for their analysis to deliver very peculiar results that cannot be trusted;
- similarly, it seems unreasonable to rely on a single test in classical studies, and it remains irrelevant to conclude after a single run: in other words, sensitivity analyses have to be conducted just as in the usual methods.

2.1.4. Shapes of the groups identified

In classical hypothesis-testing methods, a very typical assumption is the Gaussian characteristic of the samples studied. Here, groups may have different shapes depending on which method is used. The most common is the "shell", a compact shape, an ellipsoid shape in the feature space [25,32,33]. This is particularly true for first-generation methods and their descendants, such as *k*-means with Euclidean distance. Generally, these methods ensure relatively robust results and run efficiently.

2.1.5. Categorical, continuous or mixed data

Dedicated algorithms exist for categorical or mixed data, such as the PAM algorithm, which are specific enough to justify a particular treatment (since their geometry differs from a continuous data space) [22]. Most available methods apply to continuous data, and it is always possible to use a multiple correspondence analysis (MCA) [34] to transform categorical data into continuous data, and then apply a continuous data algorithm.

2.1.6. Outliers

In most cases, outliers should be removed before multivariate analysis (see Fig. 1) whatever their relative weight against the rest of the data, even if some techniques are less or even not sensitive to outliers. We also recommend treating outliers before any further consideration.

2.2. Algorithm example 1: the k-means algorithm

The *k*-means algorithm is one of the most well-known and oldest clustering algorithms [21,23,35]. Many variants have been suggested based on its principle, mainly to overcome its main flaws. It is designed to operate on continuous data and requires three types of parameter: the number *k* of clusters to be retrieved, the distance to be used (usually, Euclidean), and the *k* initial observations to initialize it. The principle of *k*-means is based on physical assumptions that natural groups are compact and well-structured enough to present the lowest global inertia. Then this algorithm consists of minimizing the inertia of each *k* group, by consecutively adding or releasing elements from one cluster to another, until the moment when moving an element does not change anything significantly in the *k* inertia. If *S* denotes the set of the *k* clusters S_k to be found, the algorithm consists in retrieving *S* such as:

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x_j \in S_i} \left\| x_j - \mu_i \right\|^2 \quad (1)$$

where μ_i is the barycenter of the cluster S_i.

K-means provide shell-shaped clusters by construction. It is sensitive to initial conditions, which means that the choice of the *k* initial elements may be of some importance. It is also sensitive to outliers.

2.3. Algorithm example 2: the PAM algorithm

The PAM (an acronym for partitioning around medoids) algorithm [22,23] can handle either categorical or continuous variables, or even mixed variables. It is reputed to be less sensitive to outliers and initial conditions than *k*-means, which it extends. It can take either a dissimilarity matrix or a classical observation matrix as arguments, and Euclidean or Manhattan distances. When the *k*-means algorithm is based on "virtual" data points (since it computes barycenters that are not likely to be "real" individuals) and on minimizing a quadratic error (the *k* inertia), the PAM algorithm uses "medoids" ("real" data points) and tends to minimize a sum of dissimilarities. Indeed, a medoid is a point of a cluster whose average dissimilarity to all the objects in the same cluster is minimal; in other words, it is the most centrally located point in the cluster. Then the PAM algorithm runs in the same way as the *k*-means algorithm: it selects *k* medoids and tries iteratively all points against them. Points are attributed to the nearest medoid in terms of dissimilarity. The algorithm stops when no significant changes are to be found in data point attribution to the *k* clusters.

2.4. Finding the "best" fitting results: testing for group validity

Several indices or measures have been suggested to assess group validity, what is usually referred to as internal validity. Two classical indices for internal validity are the Dunn index and the silhouette, even if several others have been suggested.

The Dunn index compares the size of the clusters with the distances between clusters. The greater the distance relative to the size of a given cluster, the larger the index is, suggesting better clustering. The index is denoted V(S) and is computed as:

$$Dunn = V(S) = \min\left\{\min_{j\neq i}\left[\frac{d_s(S_i, S_j)}{\max\Delta(S_k)}\right]\right\} \quad (2)$$

where d_s stands for the distance between cluster S_i and S_j, and $\Delta(S_i)$ for the size of the cluster S_i.

The silhouette may be one of the oldest indices [22], but it remains quite reliable whatever technique is used to cluster data. The silhouette width of a cluster is the sum of each observation silhouette that contributes to this cluster. The observation silhouette ranges between −1 and 1. A silhouette of 1 means a correct cluster attribution, while −1 indicates an erroneous cluster attribution, and 0 stands for an observation that could either have been attributed to its present cluster or another one. The silhouette of a given data point *x* is given by the formula:

$$S(x) = \frac{b(x) - a(x)}{\max_x[b(x), a(x)]} \quad (3)$$

where *a(x)* is the average distance between *x* and the other data points of the same cluster, and *b(x)* is the minimum of the average distances between *x* and data points from the other clusters.

Empirically, the same authors who proposed the silhouette width approach also provided cut-off values for cluster intrinsic validity: less than 0.25 = no substantial structure was found; 0.26–0.50 = a weak, potentially artificial structure, to be validated by other techniques; 0.51–0.70 = a reasonable structure; 0.71–1.00 = strong confidence in the structure found.

2.5. Finding the "best" fitting results: group stability

Another key issue is to assess whether the clusters obtained with a given set of parameters and a given technique are stable. In other words, when re-running the same algorithm on the same data, with the same set of parameters, how sure can we be of obtaining similar results? Here again, different measures of stability can be found in the literature [36–38]. A useful approach consists in evaluating the stability of groups while iterating the same method with the same parameter set, on randomly chosen and reshuffled subsamples of the data. This index is called cluster robustness, and is presented in [39,40]. Its main parameters are the proportion of observations from the original data set to be used and the number of iterations needed.

For example, for between two and six clusters, the PAM algorithm can be challenged on 100 subsamples made of 80% of the original data set, reshuffled for each iteration. Robustness will then be computed as the propensity for two observations to be assigned to the same cluster throughout the 100 iterations. If all the couples of observations are assigned to the same clusters at each iteration, then the result is likely to be robust and reliable. Such resampling-based methods are also used to address the sensitivity to initial conditions.

2.6. Further group analysis: testing for differences, looking for association measurements

Given a clustering result, data are enriched with one more categorical variable, which is the cluster identifier to which every patient or individual belongs. Obviously, at that point, any classical tests can then be used for group comparisons. When a further objective is to search for the factors associated with belonging to a specific cluster rather than another one, the usual logistic or multinomial regression models can then be used as usual, with every observation now labeled with a cluster identifier as an outcome (see [41] for a recent example).

3. Results

We present here a simple, theoretical example and then a real data application to illustrate our general framework.

3.1. A simple, theoretical example

We created a data set of 200 patients, originating from four different groups in terms of height and weight. For the purpose of this example, let us postulate that we have a group of young

frail children, one of obese adolescents, one of relatively small female adults and a group of relatively large male adults. We assume that heights and weights within a group follow a Gaussian distribution, centered on their respective mean. Data have been generated with the R software version 2.14.2 (64 bits), running on Windows 7, and the clusterCons package. The R code we used for data generation as well as for the following analysis is provided in the supplementary data. Our training objective is to distinguish between four qualitative groups to be discovered in these data using clustering algorithms.

We used four different techniques, namely PAM, *k*-means algorithms, hierarchical clustering (hclust) [23,42] and Divisive Analysis (DIANA) algorithms [22]. We used the cluster robustness approach to guide our search for optimal parameters. Since four groups were generated, we expected the optimal number of clusters to be retrieved to be four. We tested several algorithms with assumed numbers of clusters ranging from two to six. We ran the comparison algorithm on the complete data, with 100 repetitions and an 80% reshuffle proportion. We also provided results on the so-called internal validity of the clusters, through the silhouette criterion, for the four techniques mentioned above, plus two additional nonlinear techniques: the SOM algorithm and its tree variation, the SOTA algorithm. We give graphical results of data labeling according to the number of assumed clusters, as well as quantitative results on the clusters' robustness and silhouettes, according to the number of assumed clusters.

The graphical results are displayed in Figs. 2 and 3. First of all, the silhouette approach gave us valuable and clear information on the optimal clustering of the data for both algorithms used. There was no doubt that (i) four was the best choice and (ii) it led to strong, confident structures with stable silhouette values of about 0.75.

The silhouette and robustness results are given in Fig. 4. The robustness approach, based on extensive resampling, shuffling and iterative runs of the same algorithm, taught us that four clusters were the most likely to be the best choice as well, except for the *k*-means algorithm, for which five clusters seemed to be the best option. Indeed, mean robustness is higher for four clusters, with the narrowest range of values.

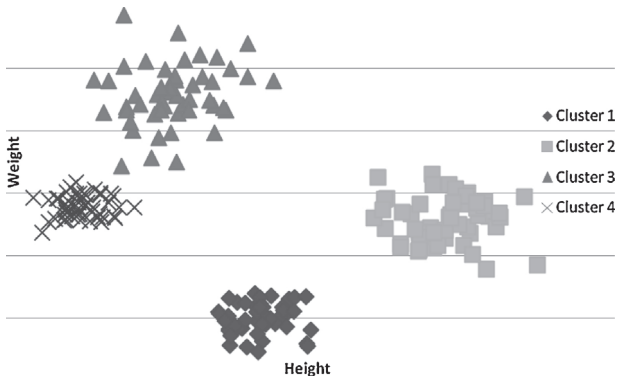


Fig. 2. Distribution of a study population by weight and height: graphical representation of a simple, theoretical example, with its four "real" clusters as they have been built.

3.2. A real data application: searching for specific profiles in healthcare resource utilization data

The data presented here were extracted from the French SIRS cohort study, with the latest data collected in 2010 in a representative sample of 3000 French-speaking adults in the Paris metropolitan area (Paris and its suburbs, a region with a population of 6.5 million). The SIRS methodology and further detailed characteristics have been previously described elsewhere, for example in [43].

For a more comprehensible view of how the French healthcare system is solicited – or not – by its users, one may want to identify specific and homogeneous profiles of healthcare resource utilization. The SIRS cohort study provides us with numerous variables that account for various aspects of the healthcare system utilization. Here, we chose 17 of them, which we grouped with respect to specific aspects of the French healthcare system [44]: primary care (six variables), the indirect access to a specialist (IAS, one variable), paramedical or alternative care (two variables), places for healthcare (six variables) and emergency care (including emergency units as well as unplanned care, two variables). We then applied our method to these variables, identically as for the previous simple example. Groups were primarily sought using the PAM algorithm. The detailed results are given in Fig. 5 and Table 1.

Starting with 17 characteristics of healthcare resource utilization, we ended with four clearly separated and meaningful utilization profiles, as provided by our group robustness analysis. Type 1 intensively uses all types of healthcare services, presents particularly high rates of consultation with general practitioners (GPs), IAS, home visits and consultations in emergency units. Type 2 presents the lowest level of utilization of healthcare services, whereas type 3 mainly consults GPs or uses emergency healthcare services and type 4 is the most likely to frequently consult specialists in ambulatory settings and to make use of nonconventional care. Noticeably, these four groups are statistically well differentiated, according to most variables.

In the second step, factors associated with these profiles, such as sociodemographic factors, could be analyzed, using multivariate models, for example. Also, healthcare system variables and other factors of interest can be jointly analyzed with specific clustering techniques called co-clustering techniques [10], which will not be addressed here.

4. Discussion

We presented here a number of powerful techniques for naturally massive multivariate analysis based on minimalist (mainly geometrical) assumptions. Pattern recognition or clustering techniques provide us with a vast choice of techniques, even if many of them are variations of the same one. It is still a relatively young and very active field of research with considerable experience in various domains. It broadens the usual scope of hypothesis-testing, without excluding it, and requires new habits, particularly rethinking the way we conceive of homogenous groups, since the results provided

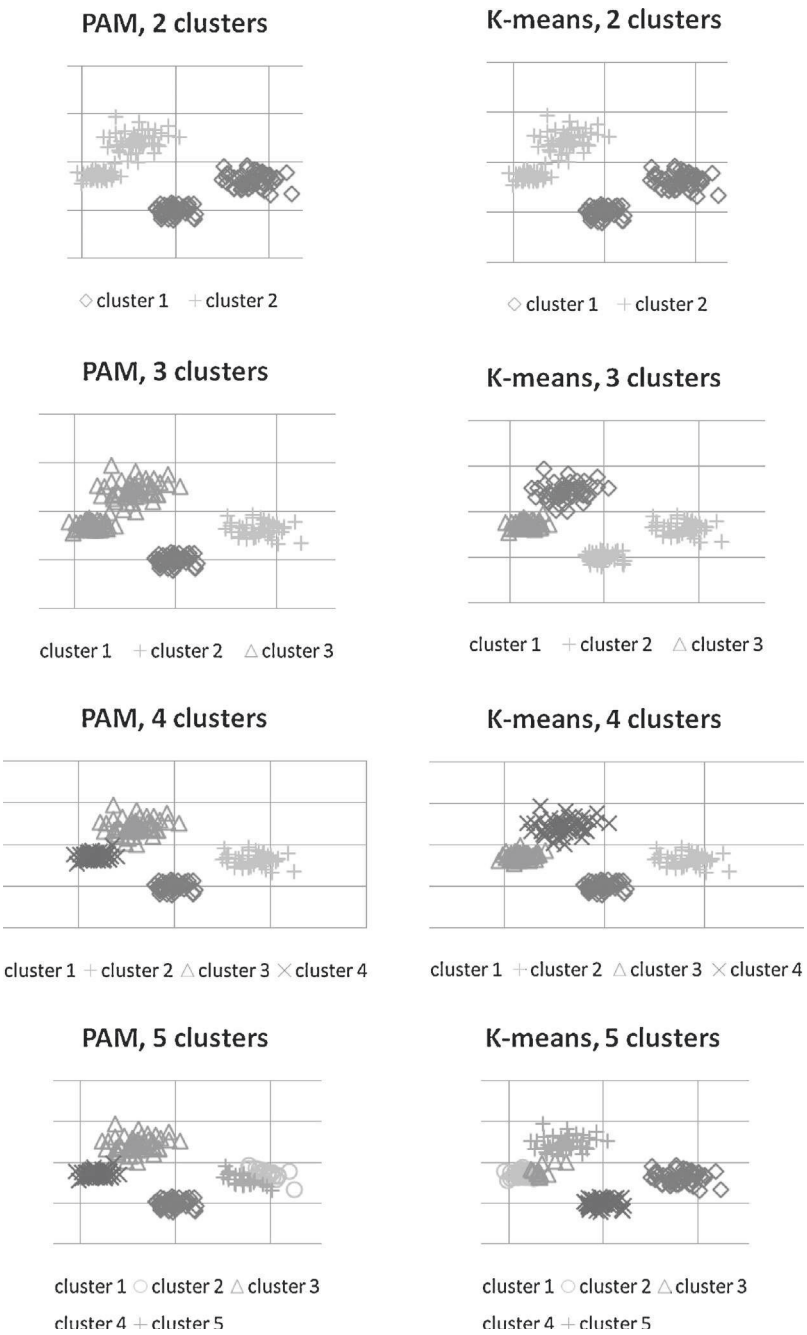


Fig. 3. Graphical representation of the results of two clustering algorithms applied to the theoretical example, searching for two to five clusters, respectively.

by these techniques are not attached to a significance cut-off, separating the ''true'' from the ''false'' as is usually done.

Several indices and approaches can help the researcher find a path through data analyses and provide evidence for reliable results. We have suggested the use of the silhouette width technique and the robustness technique, as well as a general framework for data exploration, and showed how such algorithms and indices operate through a simple example.

In this simple, theoretical example, since we knew the correct classification of people among the four possible groups, we could have computed the ratio of people misclassified using this or that technique, according to the number of clusters assumed. In the real world, as in our real data example, we are often faced with the absence of any ''gold standard'' or ''oracle'' information about the potential ''true'' nature of the clusters to be found. In this simple, theoretical example as well as in the real data application, one could argue that for the PAM or DIANA algorithms, two clusters may have appeared to be the choice to retain. Would it have been a mistake? Not necessarily: a look at the graphical display of the data shows that separating data into two groups was possibly a good choice (if not the only one) since it did not misclassify people. Actually, it looks as if the data had been examined from a higher scale: distinguishing two clearly separated groups does not prevent these two groups

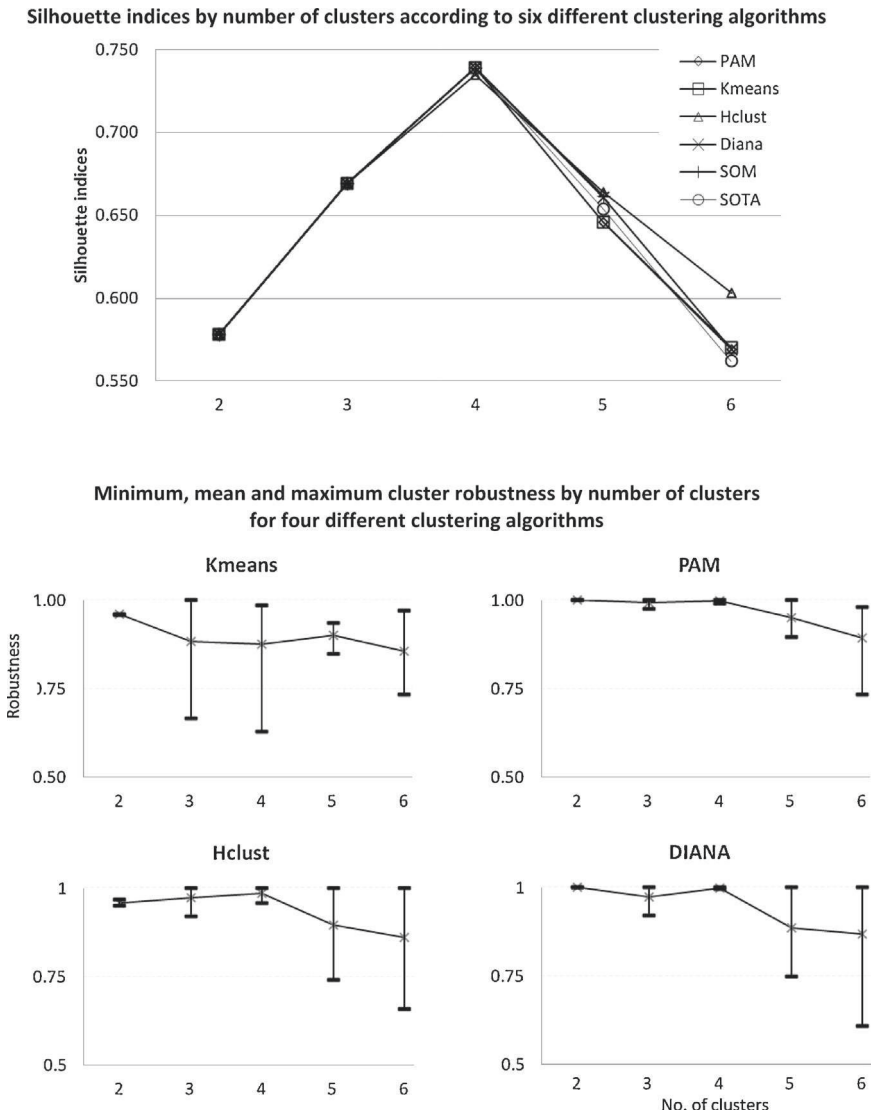


Fig. 4. Silhouette and robustness for different clustering algorithms applied to the theoretical example. PAM: partitioning around medoids; Hclust: hierarchical clustering; DIANA: divisive analysis; SOM: self-organizing map; SOTA: self-organizing tree algorithm.

from being further broken down into two other subgroups, at a lower scale. Robustness values showed us that two, three or four might have been a more or less acceptable choice. It is worth noting that robustness values decreased after $k = 4$.

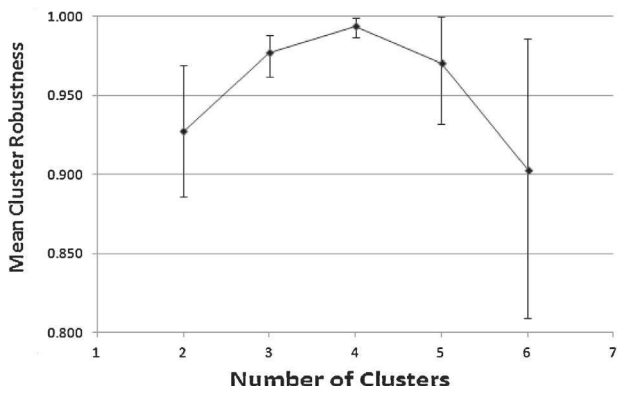


Fig. 5. Group robustness for the PAM algorithm applied to the SIRS cohort data.

The danger of searching for very precise or high numbers of clusters deserves emphasis here. Indeed, in the (numerous) situations where data are diffuse and overlapping, the best number of clusters may be much higher and, ultimately, it may even reach the total number of individuals (each person becoming his or her own personal group). Conversely, looking for a small number of patterns when data are not high-dimensional (namely two, maybe three) may lead to erroneous statements or, at the least, to poorly informative results. Moreover, one must remember that if two is the expected number of groups, then the usual statistical techniques may be used with equal success as those proposed here.

More generally speaking, pattern recognition – and especially the few techniques presented herein – suffers from several shortcomings. The first-generation algorithms (such as *k*-means, hierarchical clustering, PAM and their variants) all operate natively on a flat space, which means that they do not take into consideration that the actual intrinsic dimensionality of the data sets is almost surely not the dimensionality given by

Table 1
Proportions of the different types of resource utilization according to the four types of healthcare utilization, Paris metropolitan area, 2010.

	Type 1 (30.0%)	Type 2 (21.0%)	Type 3 (25.7%)	Type 4 (23.3%)	All types	*P*
Primary care						
Date of the last dentist consultation						< 0.001
< 2 years	84.3	68.1	72.9	87	78.2	
2–3 years	4	9.8	9.2	4.3	6.8	
> 3 years	7.8	16.8	13.1	6.2	10.9	
Never	3.9	5.3	4.8	2.5	4.1	
Having a referring GP	96	71.5	93.2	90.6	88.1	< 0.001
Frequency of consultation with a GP[a]						< 0.001
0	2.4	60.4	–	6.7	16.9	
1	9	39.6	–	23.7	17.7	
2	17	–	31.8	23.9	18.1	
3–5	38.6	–	47.9	34.6	30.5	
6+	33	–	20.3	11.1	16.8	
Frequency of consultation with a DAS[a]						< 0.001
0	49.3	87.7	93.6	–	57.8	
1	22.3	12.2	6.4	40.2	20.2	
2	11.4	0.1	–	25.1	9.1	
3–5	9.4	–	–	16.3	6.5	
6+	7.6	–	–	18.3	6.4	
Has had a medical check-up in a dedicated center[a]	7	5.1	5.7	3.6	5.4	0.127
Frequency of requesting medical advice from relatives[a]						0.0057
0	78.1	85.3	84.4	78.3	81.4	
1–2	14.2	10.7	12.3	14.7	13	
3–10	6.1	3.8	3.4	6.3	4.9	
11+	1.6	0.2	–	0.7	0.6	
Paramedical and alternative care[a]						
Has consulted an acupuncturist/osteopath	22.7	6.4	9.4	17.1	17.1	< 0.001
Has consulted for nonconventional/ alternative healthcare	7.2	2.1	3.3	6.9	4.9	< 0.001
Indirect access specialists[a]						
Frequency of consultation with an IAS						< 0.001
0	–	81.1	78.4	68	54.8	
1	1.3	12.2	21.6	28.6	15.3	
2	29.3	1.8	–	3.3	9.4	
3–5	40.8	4.9	–	0.1	12.6	
6+	28.7	–	–	–	0.8	
Places for healthcare[a]						
Public hospital or clinic-GP	12.3	3.1	14.1	7.6	9.4	< 0.001
Public hospital or clinic-specialist	42.5	5.5	7.2	19.9	19.6	< 0.001
Private hospital or clinic-GP	14.2	1.8	10.5	6.3	8.4	< 0.001
Private hospital or clinic-specialist	21.1	1.5	3.5	9.3	9.3	< 0.001
Ambulatory settings-GP	91	34.9	90.9	88.6	76.8	< 0.001
Ambulatory settings-specialist	83	22.3	16.3	88.8	53.4	< 0.001
Emergency care[a]						
Frequency of home visits						0.002
0	87.1	96.5	92.8	91.7	91.8	
1	9.2	3.5	5.4	6.6	6.3	
2	2.5	–	1.2	1.6	1.4	
3–5	1.2	–	0.6	–	0.5	
6+	–	–	–	–	–	
Frequency of consultations in emergency units						< 0.001
0	77.5	89.8	80.7	83	82.6	
1	15.9	9.7	15.7	13	13.7	
2	4	0.1	2.2	2.1	2.3	
3–5	2.2	–	1.2	1.6	1.3	
6+	0.4	–	0.2	0.3	0.2	

GP: general practitioner; DAS: direct access specialist; IAS: indirect access specialist. All results are expressed in percentages.
[a] All reference periods were the last 12 months.

its number of variables (i.e., there is redundancy between variables). It implies that some distances between data points are not correct, because those points are not real neighbors.

They also operate in high-dimensional spaces, which are prone to being difficult, if not impossible, to handle correctly when dimensionality increases: this is known as the "curse of dimensionality": the higher the dimensionality, the less two data points must be distinguished from one another, implying that they cannot accurately separate two points belonging to distinct clusters.

Most of the techniques presented herein are linear methods (with the exception of the SOM and SVC techniques mentioned). This implies that if there are indeed two groups to be discovered, but that they are not separated by a line or a plane (respectively, a hyperplane), then the algorithms will fail to retrieve the correct structures. Lastly, we have already mentioned that these first-generation algorithms can only produce compact, shape-constrained clusters.

There are several ways to address these shortcomings – such as nonlinear dimensionality reduction to circumvent the curse of dimensionality, nonlinear clustering or kernel-based techniques (such as SOM or SVC) – and, whatever their limitations are, we believe that these methods are worth epidemiologists, confronted with large data sets and/or the need for cluster research, getting to know.

5. Available software and packages

Either free or commercial suites, the main statistical packages available provide at least partial solutions for pattern recognition – clustering: SAS (proc fastclust), SPSS (SPSS TwoStep Cluster Component), STATA (linkage, *k*-means), R, and Matlab. In some cases, proprietary solutions can even be found (e.g., SPSS).

For R, several packages can be used, such as cluster [45], clusterCons [40], and clv [46], clValid [47]. The R software runs on Linux or Windows machines, is freely downloadable, and all packages are fully documented and often come with code examples.

Dissimilarity matrices can be computed using a dedicated algorithm, called DAISY [22], also available in the R Cluster package.

Disclosure of interest

The authors declare that they have no conflicts of interest concerning this article.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.respe.2014.12.017.

References

[1] Xu M, Wei C. Remotely sensed image classification by complex network eigenvalue and connected degree. Comput Math Methods Med 2012;2012:632–703.
[2] Wang Y, Shi H, Ma S. A new approach to the detection of lesions in mammography using fuzzy clustering. J Int Med Res 2011;39:2256–63.
[3] Connolly AC, Guntupalli JS, Gors J, Hanke M, Halchenko YO, Wu Y-C, et al. The representation of biological classes in the human brain. J Neurosci 2012;32:2608–18.
[4] Khotanlou H, Afrasiabi M. Segmentation of Multiple Sclerosis Lesions in Brain MR Images Using Spatially Constrained Possibilistic Fuzzy C-Means Classification. J Med Signals Sens 2011;1:149–55.
[5] Alexandrov T, Kobarg JH. Efficient spatial segmentation of large imaging mass spectrometry datasets with spatially aware clustering. Bioinformatics 2011;27:i230–8.
[6] Prabhakara S, Acharya R. Unsupervised two-way clustering of metagenomic sequences. J Biomed Biotechnol 2012;2012:153647.
[7] Priyadarshini G, Sarmah R, Chakraborty B, Bhattacharyya DK, Kalita JK. An effective graph-based clustering technique to identify coherent patterns from gene expression data. Int J Bioinform Res Appl 2012;8:18–37.
[8] Nascimento M, Sáfadi T, Fonseca E, Silva F, Nascimento ACC. Bayesian model-based clustering of temporal gene expression using autoregressive panel data approach. Bioinformatics 2012;28(15):2004–7.
[9] Maji P, Das C. Relevant and significant supervised gene clusters for microarray cancer classification. IEEE Trans Nanobioscience 2012;11:161–8.
[10] Hwang T, Atluri G, Xie M, Dey S, Hong C, Kumar V, et al. Co-clustering phenome-genome for phenotype classification and disease gene discovery. Nucleic Acids Res 2012;40(19):e146.
[11] Dasgupta A, Sun YV, König IR, Bailey-Wilson JE, Malley JD. Brief review of regression-based and machine learning methods in genetic epidemiology: the Genetic Analysis Workshop 17 experience. Genet Epidemiol 2011;35(Suppl 1):S5–11.
[12] Martín-López B, Iniesta-Arandia I, García-Llorente M, Palomo I, Casado-Arzuaga I, Amo DGD, et al. Uncovering Ecosystem Service Bundles through Social Preferences. PLoS One 2012;7:e38970.
[13] Ramoji A, Neugebauer U, Bocklitz T, Foerster M, Kiehntopf M, Bauer M, et al. Toward a spectroscopic hemogram: Raman spectroscopic differentiation of the two most abundant leukocytes from peripheral blood. Anal Chem 2012;84:5335–42.
[14] Sutherland ER, Goleva E, King TS, Lehman E, Stevens AD, Jackson LP, et al. Cluster analysis of obesity and asthma phenotypes. PLoS One 2012;7:e36631.
[15] Conry MC, Morgan K, Curry P, McGee H, Harrington J, Ward M, et al. The clustering of health behaviours in Ireland and their relationship with mental health, self-rated health and quality of life. BMC Public Health 2011;11:692.
[16] Muntaner C, Chung H, Benach J, Ng E. Hierarchical cluster analysis of labour market regulations and population health: a taxonomy of low- and middle-income countries. BMC Public Health 2012;12:286.
[17] Armstrong JJ, Zhu M, Hirdes JP, Stolee P. K-Means Cluster Analysis of Rehabilitation Service Users in the Home Health Care System of Ontario: Examining the Heterogeneity of a Complex Geriatric Population. Arch Phys Med Rehab 2012;93(12):2198–205.
[18] Consortium TEP. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489:57–74.
[19] Dougherty ER, Brun M. A probabilistic theory of clustering. Pattern Recognit 2004;37:917–25.
[20] Brun M, Sima C, Hua J, Lowey J, Carroll B, Suh E, et al. Model-based evaluation of clustering validation measures. Pattern Recognit 2007;40:807–24.
[21] Theodoridis S, Koutroumbas K. Pattern Recognition. 4th Revised edition, Academic Press Inc; 2008.
[22] Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis. Hoboken, N.J: Wiley; 2005.
[23] Izenman AJ. Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning. Illustrated edition, Springer-Verlag New York Inc; 2008.
[24] Marsland S. Machine Learning: An Algorithmic Perspective. 1st ed. Boca Raton: Chapman and Hall/CRC; 2009.
[25] Jain AK. Data clustering: 50 years beyond K-means. Pattern Recognit Lett 2010;31:651–66.

[26] Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci U S A 1999;96:2907–12.
[27] Hinton GE, Salakhutdinov RR. Reducing the Dimensionality of Data with Neural Networks. Science 2006;313:504–7.
[28] Boriah S, Chandola V, Kumar V. Similarity Measures for Categorical Data: A Comparative Evaluation. In: Proceedings of SIAM Data Mining Conference; 2008.
[29] Gower JC. A General Coefficient of Similarity and Some of Its Properties. Biometrics 1971;27:857.
[30] Lee J, Lee D. Dynamic Characterization of Cluster Structures for Robust and Inductive Support Vector Clustering. IEEE Trans Pattern Anal Mach Intell 2006;28:1869–74.
[31] Ben-Hur A, Horn D, Siegelmann HT, Vapnik V. Support Vector Clustering. J Mach Learn Res 2001;2:125–37.
[32] Mu-Chun Su, Chien-Hsing Chou. A modified version of the K-means algorithm with a distance based on cluster symmetry. IEEE Trans Pattern Anal Mach Intell 2001;23:674–80.
[33] Fahim AM, Saake G, Salem AM, Torkey FA, Ramadan MA. K-Means for Spherical Clusters with Large Variance in Sizes. Int J Comput Sci 2009;4(3):145.
[34] Tufféry S. Data mining et statistique décisionnelle. Paris: Éd. Technip; 2010.
[35] Lloyd S. Least squares quantization in PCM. IEEE Trans Inf Theory 1982;28:129–37.
[36] Dalton L, Ballarin V, Brun M. Clustering algorithms: on learning, validation, performance, and applications to genomics. Curr Genomics 2009;10:430–45.
[37] Susmita, Datta S. clValid: An R Package for Cluster Validation. J Stat Softw 2008;25.
[38] Dunn J. Well separated clusters and optimal fuzzy-partitions. J Cybernetics 1974;4:95–104.
[39] Monti S, Tamayo P, Mesirov J, Golub T. Consensus clustering – A resampling-based method for class discovery and visualization of gene expression microarray data. Mach Learn J 2003;52(1–2):91–118.
[40] Simpson TI, Armstrong JD, Jarman AP. Merged consensus clustering to assess and improve class discovery with microarray data. BMC Bioinformatics 2010;11:590.
[41] Rebholz CE, Rueegg CS, Michel G, Ammann RA, Von der Weid NX, Kuehni CE, et al. Clustering of health behaviours in adult survivors of childhood cancer and the general population. Br J Cancer 2012;107(2):234–42.
[42] Jain AK, Murty MN, Flynn PJ. Data clustering: a review. ACM Comput Surv 1999;31:264–323.
[43] Martin-Fernandez J, Grillo F, Parizot I, Caillavet F, Chauvin P. Prevalence and socioeconomic and geographical inequalities of household food insecurity in the Paris region, France, 2010. BMC Public Health 2013;13:486.
[44] Chevreul K, Durand-Zaleski I, Bahrami S, Hernández-Quevedo C, Mladovsky P. France: Health system review. Health Syst Transit 2010;12:1–219.
[45] Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K. Cluster: Cluster Analysis Basics and Extensions. R package version 1.14.4; 2013.
[46] Nieweglowski L. clv: Cluster Validation Techniques; 2009.
[47] Brock G, Pihur V, Datta S, Datta S. clValid: Validation of Clustering Results. R package; 2011.

Annexe 3.2 Les 4 profils de recours aux soins en Ile-de France

PLOS ONE

RESEARCH ARTICLE

Applying Multivariate Clustering Techniques to Health Data: The 4 Types of Healthcare Utilization in the Paris Metropolitan Area

Thomas Lefèvre[1,2]*, Claire Rondet[1,3], Isabelle Parizot[4], Pierre Chauvin[1,2]

1. Inserm, UMRS 1136, Pierre Louis Institute of Epidemiology and Public Health, Department of Social Epidemiology, Paris, France, **2.** Sorbonne Universités, UPMC Univ Paris 06, UMRS 1136, Paris, France, **3.** Sorbonne Universités, UPMC Univ Paris 06, Faculty of Medicine Pierre and Marie Curie, Department of general practice, Paris, France, **4.** CNRS, UMR 8997, Centre Maurice Halbwachs, Research group on social inequalities, Paris, France

*thomas.lefevre@inserm.fr

OPEN ACCESS

Citation: Lefèvre T, Rondet C, Parizot I, Chauvin P (2014) Applying Multivariate Clustering Techniques to Health Data: The 4 Types of Healthcare Utilization in the Paris Metropolitan Area. PLoS ONE 9(12): e115064. doi:10.1371/journal.pone.0115064

Editor: Kimon Divaris, UNC School of Dentistry, University of North Carolina-Chapel Hill, United States of America

Received: May 15, 2014

Accepted: November 11, 2014

Published: December 15, 2014

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. All relevant data have been deposited to Dryad (DOI:10.5061/dryad.9v79s).

Funding: Funding for this study was provided by the Institute for Public Health Research (IReSP) and the Interministerial Committee for Urban Affairs. 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.

Abstract

Background: Cost containment policies and the need to satisfy patients' health needs and care expectations provide major challenges to healthcare systems. Identification of homogeneous groups in terms of healthcare utilisation could lead to a better understanding of how to adjust healthcare provision to society and patient needs.

Methods: This study used data from the third wave of the SIRS cohort study, a representative, population-based, socio-epidemiological study set up in 2005 in the Paris metropolitan area, France. The data were analysed using a cross-sectional design. In 2010, 3000 individuals were interviewed in their homes. Non-conventional multivariate clustering techniques were used to determine homogeneous user groups in data. Multinomial models assessed a wide range of potential associations between user characteristics and their pattern of healthcare utilisation.

Results: We identified four distinct patterns of healthcare use. Patterns of consumption and the socio-demographic characteristics of users differed qualitatively and quantitatively between these four profiles. Extensive and intensive use by older, wealthier and unhealthier people contrasted with narrow and parsimonious use by younger, socially deprived people and immigrants. Rare, intermittent use by young healthy men contrasted with regular targeted use by healthy and wealthy women.

Conclusion: The use of an original technique of massive multivariate analysis allowed us to characterise different types of healthcare users, both in terms of

resource utilisation and socio-demographic variables. This method would merit replication in different populations and healthcare systems.

Introduction

In the European context of cost-containment policies and the post-2008 economic and financial crisis [1], cost optimisation and, in some countries, cost reduction of public expenditure has become unavoidable and the healthcare system is no exception. For this reason, the healthcare system may need to be adapted to cost-containment goals while at the same time meeting patients' needs and expectations as closely as possible. This requires, among other issues, accurate characterisation of healthcare resource utilisation by the user population, as well as identification of determinants of use.

Many studies have previously addressed the use of the healthcare system (either individual services, or globally) by the general population or by specific population subgroups. For example, several studies have examined healthcare system utilisation from a systemic point of view or from a decision-making approach [2–4], or by subgroups of the population, such as cancer survivors [5], migrants [6, 7], or the underserved and low-income people [8–10]. In addition, determinants of utilisation of specific healthcare services have been investigated, including mental healthcare services [11], emergency care units [12], primary care resources [13], dental care [14] and specialist consultations [15]. Associations between health insurance and healthcare research have also been regularly documented [16].

It has been suggested that healthcare systems themselves could not be analysed through a classical reductionist approach but should be considered as complex systems [17] which require analysis with non-conventional techniques. In particular, it could be interesting to identify distinct groups of patients which would exhibit different homogeneous patterns of resource utilisation. If such groups can be identified, then factors associated with each utilisation profile can be examined using conventional approaches [18–20].

Identifying such utilisation patterns requires the use of particular multivariate techniques, which are capable of taking into account a vast amount and variety of variables simultaneously, documented from the largest population possible. These techniques, particularly clustering techniques, have been applied and validated in a wide range of areas of medicine, including genetics [21–23], imaging [24–26], clinical medicine [27, 28] and public health [29].

In this study, we aimed to identify and characterise distinct profiles of users of the French healthcare system in an urban environment, through analysis of data from a representative, population-based study in the Paris metropolitan area, using clustering techniques.

Methods

This work is based on the SIRS cohort study that received legal authorization from two French national authorities for non-biomedical research: the Comité consultatif sur le traitement de l'information en matière de recherche dans le domaine de la santé (CCTIRS) and the Commission nationale de l'informatique et des libertés (CNIL) [30]. The participants provide their verbal informed consent. Written consent was not necessary because this survey did not fall into the category of biomedical research (as defined by French law).

This study represents a cross-sectional analysis of data collected in the SIRS cohort study in 2010 among a representative sample of 3,000 French-speaking adults in the Paris metropolitan area (Paris and its suburbs, a region with a population of 6.5 million).

The SIRS cohort

The SIRS cohort was constituted in 2005 using a 3-level random sampling method. In a first step, 50 census blocks (with about 2000 inhabitants each) were randomly selected using a stratification based on socioeconomic status and whether they qualified or not for ''underprivileged urban area'' according to the central government list. In the next step, 60 households were randomly chosen from a complete list of households within each selected census block. In the final step, one adult was randomly selected from each household by the birthday method. The refusal rate among the newly contacted people was 29%. The methodology of the SIRS study and detailed characteristics of the study population have been described previously elsewhere, for example in [31].

Characterisation of healthcare utilisation

A comprehensive, detailed profile of the French healthcare system is provided in reference [32]. Interviewees were asked in detail about their own use of healthcare services during the twelve months preceding the interview. All responses were codes as categorical variables and all reference periods were the last twelve months. Resource use was grouped into categories as detailed below. Unless otherwise specified, all consultation frequencies fell into one of five categories (none, only once, only twice, 3–5 times, or $\geq$6 times).

Primary care: French people seeking healthcare may consult a general practitioner (GP), whether as an end in itself or as an entry point to specialists (the French system has adopted this gate-keeping model since 2004 [32]). Patients need to respect this procedure so that they can be reimbursed. Exceptions are made for four kinds of specialists who can be consulted directly, namely gynaecologists (who are mainly community-based in France), ophthalmologists, paediatricians and psychiatrists; these four specialities will henceforward be referred to as direct access specialists (DAS). We used six variables to characterise primary care utilisation, namely date of the latest dental consultation (4 categories: less than 2 years, between 2 and 3 years, more than 3 years, never),

having declared a referring GP (yes/no), frequency of GP consultation, frequency of DAS consultation, undergoing a medical check-up in a dedicated Social security centre (yes/no), frequency of requests for medical advice from friends and relatives (4 categories: none, 1 or 2, 3 to 10, more than 10 times).

Indirect access to a specialist (IAS): IAS concerns all other specialists except DAS. The patient may access to them only when referred by their GP (or from their own initiative but at full cost). A single variable was documented, the frequency of IAS consultations.

Paramedical or alternative care: two variables were considered: having consulted an acupuncturist or an osteopath (yes/no) and having consulted for non-conventional or alternative healthcare (yes/no). Traditional Chinese medicine fell into the latter category.

Site of healthcare consumption: in France, healthcare can be delivered in three principal settings: public hospitals or clinics, private hospitals or clinics, and community settings. Since the place of consultation was systematically documented for each medical consultation over the previous twelve months, six distinct variables were considered: having consulted (at least once) a GP in a public hospital or clinic, a private hospital or clinic, or in a community setting, and having consulted (at least once) a specialist in a public hospital or clinic, a private hospital or clinic, or in a community setting.

Emergency care: two variables documented healthcare utilisation in emergency situations, depending on the place where care was delivered; these were the frequency of home visits for emergency reasons and the frequency of consultations in an emergency unit.

Population characteristics and factors associated with healthcare utilisation

Five dimensions were explored as possibly associated with healthcare utilisation, all of them made up of three items (except for the socioeconomic status, with four items).

Demographics: age (5 categories: 18–29, 30–44, 45–59, 60–74, and 75 years old or more), gender, origin (distinguishing, as previously reported [33, 34], between French people born to two French parents, French people born to at least one foreign parent, and foreign immigrants).

Socioeconomic status: education level (none or primary/secondary/tertiary), employment status (employed, unemployed, inactive or retired), monthly household income per consumption unit (in quintiles, and computed as the total household income divided by the number of consumption units [adult: 1; child $\geq$14 years: 0.5; child $<$14 years: 0.3]), according to the usual OECD-modified scale recommended by Eurostat, and health insurance status (full coverage by the statutory health insurance – SHI, SHI plus a voluntary health insurance - VHI, full coverage by a special insurance for the poor, partial coverage by the SHI only, and no insurance at all).

Stance regarding health and medicine: general attitude toward medical consultation (people were asked if they generally consult a doctor as a last resort, or as soon as they are not feeling well), having a relative or a friend suffering from a severe condition, and having medical professionals among relatives.

Social integration: feeling of isolation (very isolated, rather isolated, rather supported, very supported), level of social support (low, medium, high) and frequency of social contacts (quartiles), both as described in [34].

Perceived health: as measured by the Minimum European Health Module [35, 36] that assembles the global perceived health status (good, average, bad), the global activity limitation indicator (presence of a long-standing activity limitation in the previous six months), and the presence of a chronic or long standing health problems over the twelve months.

These dimensions are similar to those identified by Anderson in the late 60's [37, 38]. In his Behavioral Model of Health Service Use (BMHSU), this author distinguished between three classes of factors: predisposing factors (such as age, gender, education, occupation, social relationships, attitudes and knowledge related to health services and professionals), enabling factors (such as income or health insurance), and need factors (such as perceived health status or functional disability). According to this model, Andersen suggested that the respective roles of these factors may provide clues for measuring equity in service use. For example, if the main drivers of health service use are need factors, access can be considered equitable. Conversely, if the main drivers are constituted of social factors, beliefs and enabling factors, access can be considered as not equitable.

Clustering methods

The use of clustering techniques available for health scientists has been described previously [39–41]. Clustering techniques require the determination of some data-specific parameters, such as the number of groups to be retrieved. In order to identify the different types of healthcare system utilisation, we used the partitioning around medoïds (PAM) algorithm with the Euclidean distance as a reference analysis and applied it to the healthcare system utilisation variables [40]. A resampling-based scheme and cluster-robustness approach [42] was used to determine the key parameters of the algorithm and in particular the number of clusters. Other clustering methods or sets of parameters were further used for sensitivity analyses. The PAM algorithm was run with alternative distance or similarity measure (Manhattan distance and Gower measure [40]). A fuzzy-logic version of PAM, the FANNY algorithm [40], was also applied to the data, with several values for the fuzziness parameter. All analysis was conducted on R 2.13.1 (R Foundation for Statistical Computing, 2012), with the clusterCons package.

Statistical analyses

We accounted for the three-level sampling design of the SIRS cohort by using the *survey* command options of the STATA IC 10 software (STATA Corp, 2007) for

descriptive statistics and multinomial models. Classical ratio tests (chi-square or exact Fisher) were used to compare population characteristics according to their type of care utilisation.

We used multinomial regression models to investigate significant associations between variables of each of the five dimensions studied, and the types of healthcare utilisation as categorical outcomes. Adjusted odd ratios (OR) are reported with a p value for linear trend. Statistical significance was assessed at a bilateral p value<0.05.

Results

Cluster identification

The optimal number of individual clusters that would account for the data was four. This was the value at which mean cluster robustness was maximal and the range of robustness values narrowest (Fig. 1). A cluster robustness of unity indicates that no member of a given cluster was likely to be assigned to another cluster when the algorithm was reiterated and also directly reflects the stability of the cluster. In our analysis, mean robustness for the four-cluster model was high >0.99, indicating that very few individuals could not be assigned unequivocally to a given cluster.

Sensitivity analyses

No qualitative differences in cluster distribution or robustness were identified by reiterating the algorithm with alternative distance or similarity measures. Qualitatitively identical findings were obtained in all models (results not shown).

The four types of healthcare utilisation

The four clusters identified in our data were associated with four distinct types of healthcare utilisation, accounting for 30.0% (Type 1), 21.0% (Type 2), 25.7% (Type 3) and 23.3% (Type 4) of the study population. Table 1 shows the contribution of each variable of healthcare utilisation to each Type.

Type 1 represents the largest users of primary care. These individuals used all available resources, including GPs, social security centres for check-ups and medical advice from relatives, and used these resources extensively. For example, 71.6% had consulted their GP three times or more in the past twelve months and also consulted DAS extensively (50.7% with at least one visit). They were also the most frequent users of IAS, with 100% having consulted at least one IAS in the last 12 months and 69.5% consulted more than two times. They were also the largest users of paramedical or alternative care. Type 1 individuals consulted in all settings (principally in public hospitals for specialists and in community care for GPs) and were the largest users of the private sector. Finally, they were also the principal users of emergency care, both in the home (12.9%) and in emergency units (22.5%).

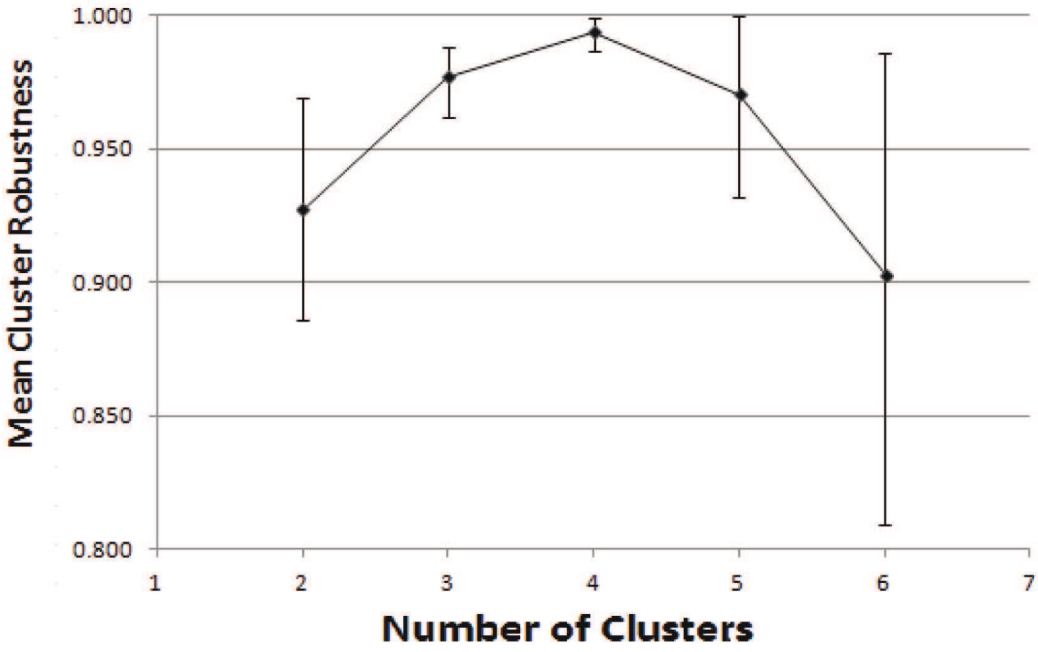


Fig. 1. Cluster robustness according to the number of assumed clusters in the data. Mean cluster robustness, together with minimum and maximum values, is presented as a function of the number of searched clusters.

doi:10.1371/journal.pone.0115064.g001

Type 2 was the mirror image of Type 1. Together with Type 3, individuals in Type 2 were the least frequent users of primary care. In Type 2, 28.5% of individuals had no referring GP and only 12.3% had consulted a DAS (furthermore, only once in most cases). Although 18.9% had consulted an IAS in the previous year, but only 4.9% consulted more than twice. These individuals rarely used paramedical or alternative resources. Whatever the setting, Type 2 users had the lowest rate of healthcare utilisation, and this was especially true for private hospitals and clinics. Type 2 individuals rarely required emergency care in the home (3.5%) or in emergency units (9.8%) and, when they did, they usually (9.7%) consulted only once and hardly ever more than twice (0.1%).

Healthcare resource utilisation by Type 3 users was closer to that observed in Type 2 than that in Types 1 or 4. Type 3 individuals were characterised by extensive recourse to GPs, with 20.3% of users have consulted more than six times in the last twelve months. In contrast, they rarely consulted DAS (6.4%) or IAS (only 21.6% of them had consulted an IAS and never more than once). They seldom used paramedical or alternative care (9.4% and 3.3% respectively). When consulting GPs, they were more likely to consult in community care, and had little use for the public sector. Nonetheless, Type 3 users constituted the second most frequent user of emergency resources, especially in emergency units (19.3% consulted at least once in emergency units).

Type 4 shared similarities with Type 1, in that it was constituted by people who were heavy users of the healthcare system. Type 4 did not present a particularly

Table 1. Resource utilisation by the four types of healthcare utilisation in the Paris metropolitan area, 2010.

	Type 1	Type 2	Type 3	Type 4	All types	p
	(30.0%)	(21.0%)	(25.7%)	(23.3%)	N=3006	
Primary care						
Date of the last dentist consultation						<0.001
<2 yr	84.3	68.1	72.9	87.0	78.2	
2–3 yr	4.0	9.8	9.2	4.3	6.8	
>3 yr	7.8	16.8	13.1	6.2	10.9	
never	3.9	5.3	4.8	2.5	4.1	
Having a referring GP						<0.001
	96.0	71.5	93.2	90.6	88.1	
Frequency of consultation with a GP						<0.001
0	2.4	60.4	-	6.7	16.9	
1	9.0	39.6	-	23.7	17.7	
2	17.0	-	31.8	23.9	18.1	
3–5	38.6	-	47.9	34.6	30.5	
6+	33.0	-	20.3	11.1	16.8	
Frequency of consultation with a DAS						<0.001
0	49.3	87.7	93.6	-	57.8	
1	22.3	12.2	6.4	40.2	20.2	
2	11.4	0.1	-	25.1	9.1	
3–5	9.4	-	-	16.3	6.5	
6+	7.6	-	-	18.3	6.4	
Has had a medical check-up in a dedicated centre						0.127
	7.0	5.1	5.7	3.6	5.4	
Frequency of requests for medical advice from relatives						0.0057
0	78.1	85.3	84.4	78.3	81.4	
1–2	14.2	10.7	12.3	14.7	13.0	
3–10	6.1	3.8	3.4	6.3	4.9	
11+	1.6	0.2	-	0.7	0.6	
Paramedical and alternative care						
Has consulted an acupuncturist/osteopath						<0.001
	22.7	6.4	9.4	17.1	17.1	
Has consulted for non-conventional/alternative healthcare						<0.001
	7.2	2.1	3.3	6.9	4.9	
Indirect access specialists						
Frequency of consultation with an IAS						<0.001
0	-	81.1	78.4	68.0	54.8	
1	1.3	12.2	21.6	28.6	15.3	
2	29.3	1.8	-	3.3	9.4	

Table 1. Cont.

	Type 1 (30.0%)	Type 2 (21.0%)	Type 3 (25.7%)	Type 4 (23.3%)	All types N=3006	p
3–5	40.8	-	-	0.1	12.6	
6+	28.7	-	-	-	0.8	
Site of healthcare consumption						
Public hospital or clinic-GP	12.3	3.1	14.1	7.6	9.4	<0.001
Public hospital or clinic-specialist	42.5	5.5	7.2	19.9	19.6	<0.001
Private hospital or clinic-GP	14.2	1.8	10.5	6.3	8.4	<0.001
Private hospital or clinic-specialist	21.1	1.5	3.5	9.3	9.3	<0.001
Ambulatory settings-GP	91.0	34.9	90.9	88.6	76.8	<0.001
Ambulatory settings-specialist	83.0	22.3	16.3	88.8	53.4	<0.001
Emergency care						
Frequency of home visits						0.002
0	-	81.1	78.4	68.0	54.8	
1	1.3	12.2	21.6	28.6	15.3	
2	29.3	1.8	-	3.3	9.4	
3–5			40.8	-	-	0.1
6+	28.7	-	-	-	0.8	
Frequency of consultations in emergency units						<0.001
0	77.5	89.8	80.7	83.0	82.6	
1	15.9	9.7	15.7	13.0	13.7	
2	4.0	0.1	2.2	2.1	2.3	
3–5	2.2	-	1.2	1.6	1.3	
6+	0.4	-	0.2	0.3	0.2	

GP: General Practitioner. DAS: Direct Access Specialist. IAS: Indirect Access Specialist. All results are expressed in percentages.

doi:10.1371/journal.pone.0115064.t001

high rate of GP consultations (being the third-most frequent user) but had the highest use of DAS (100%). Type 4 individuals were also the second most frequent users who referred to relatives for medical advice, and who consulted for paramedical or alternative care. Type 4 was also the second highest user of IAS, although the frequency of consultation was relatively low (21.6% had consulted once and none consulted more than once). Individuals in Type 4 used all settings when consulting specialists (public and private hospitals, as well as community care). For emergency care, Type 4 presented the lowest level of use compared to the other Types, with 17.0% having consulted in emergency units, 13.0% only once, and 1.6% between three and five times (second user in that case).

Factors associated with healthcare utilization

Univariate associations between independent variables and each of the four profiles of healthcare utilisation are presented in Table 2. Only three variables were not associated with significant differences between profiles, namely having

Table 2. Population characteristics of the four types of healthcare utilisation in the Paris metropolitan area, 2010.

	Type 1	Type 2	Type 3	Type 4	All types	p
	(30.0%)	(21.0%)	(25.7%)	(23.3%)	N=3006	
Model 1: Demographics						
Age (yr.)						<0.001
18–29	11.6	32.1	21.2	23.7	21.8	
30–44	22.8	32.0	35.0	39.5	31.9	
45–59	25.5	23.4	20.2	23.7	23.3	
60–74	24.2	9.3	14.1	9.6	14.7	
75+	15.9	3.3	9.5	3.5	8.3	
Gender						<0.001
Male	42.6	64.3	59.6	21.0	47.0	
Female	57.4	35.7	40.4	79.0	53.0	
Origin						0.006
French, born to French parents						
	72.2	63.3	61.1	69.1	66.6	
French, born to at least one foreign parent						
	18.3	22.0	23.2	20.1	20.8	
Foreigner	9.5	14.7	15.7	10.8	12.6	
Model 2: Socioeconomic status						
Education level						<0.001
Tertiary	59.3	54.5	48.7	63.1	56.5	
Secondary	32.5	38.0	42.3	32.2	36.2	
Primary or none	8.2	7.5	9.0	4.7	7.3	
Employment status						<0.001
Employed	47.9	61.6	55.5	63.1	56.7	
Unemployed	5.7	8.6	9.3	7.1	7.6	
Inactive	46.4	29.8	35.2	29.8	35.7	
Income (quintiles)						
1st	16.5	22.8	24.6	18.9	20.6	<0.001
2nd	14.2	22.4	23.0	18.5	19.3	
3rd	22.3	21.8	19.7	23.2	21.7	
4th	20.0	18.2	15.6	17.8	18.0	
5th	27.0	14.8	17.1	21.6	20.4	
Health insurance status						<0.001
SHI+VHI	94.1	78.5	85.3	89.6	87.1	
special insurance for the poor	1.7	4.1	2.1	3.0	2.6	
SHI only	3.6	16.6	12.4	6.8	9.7	
None	0.6	0.8	0.2	0.6	0.6	
Model 3: Stance regarding health and medicine						
General attitude toward medical consultation						<0.001
As a last resort	33.0	68.7	38.7	43.2	45.4	
As soon as not feeling well						

Table 2. Cont.

	Type 1	Type 2	Type 3	Type 4	All types	p
	(30.0%)	(21.0%)	(25.7%)	(23.3%)	N=3006	
	67.0	31.3	61.3	56.8	54.6	
Having a relative or a friend suffering from a severe condition						0.002
No	51.5	42.7	41.8	53.0	47.3	
Yes	48.5	57.3	58.2	47	52.3	
Having medical professionals among relatives						0.215
No	55.7	59.3	61.8	55.8	58.1	
Yes	44.3	40.7	38.2	44.2	44.2	
Model 4: Social integration						
Isolation feeling						0.251
Very supported	29.7	35.2	32.0	34.5	32.7	
Rather supported	55.6	54.1	53.7	52.5	54.1	
Rather isolated	12.5	9.6	12.1	12.3	11.6	
Very isolated	2.2	0.1	0.2	0.7	1.6	
Level of social support						<0.001
High	81.9	90.7	87.4	91.3	87.6	
Medium	13.8	6.5	9.1	6.2	9.1	
Low	4.3	2.8	3.5	2.5	3.3	
Frequency of social contacts (quartiles)						0.054
1st	25.0	23.3	22.5	18.7	22.5	
2nd	24.2	24.9	27.1	21.1	24.4	
3rd	26.3	21.6	22.0	30.4	25.1	
4th	24.5	30.2	28.4	29.7	28.0	
Model 5: Health status						
Perceived health status						<0.001
Good	62.3	92.2	77.4	82.4	78.0	
Average	28.5	6.8	20.0	15.2	18.0	
Bad	9.2	1.0	2.6	2.4	4.0	
Chronic or long standing health problem						<0.001
No	40.2	86.4	64.4	72.2	64.8	
Yes	59.8	13.6	35.6	27.8	35.2	
Long standing activity limitation						<0.001
No	63.1	94.4	81.1	86.5	80.6	
Yes	36.9	5.6	18.9	13.5	19.4	

All results are expressed in percentages.

doi:10.1371/journal.pone.0115064.t002

Table 3. Characteristics associated with the type of healthcare utilisation: multinomial logistic regression model (with Type 4 as reference), Paris metropolitan area, 2010.

	Type 1	Type 2	Type 3	p*
	OR [CI 95%]	OR [CI 95%]	OR [CI 95%]	
Model 1: Demographics				
Age (yr.)				<0.001
18–29	Ref	Ref	Ref	
30–44	1.18 [0.72–1.93]	0.60 [0.37–0.96]	0.99 [0.67–1.45]	
45–59	2.24 [1.35–3.73]	0.77 [0.48–1.31]	1.01 [0.67–1.55]	
60–74	5.44 [3.28–9.03]	0.80 [0.48–1.31]	1.87 [1.17–2.97]	
75+	11.14 [5.67–21.89]	1.00 [0.47–2.12]	4.45 [2.49–7.94]	
Gender				<0.001
Male	Ref	Ref	Ref	
Female	0.32 [0.22–0.46]	0.14 [0.11–0.20]	0.17 [0.11–0.24]	
Origin				0.0375
French, born to French parents				
	Ref	Ref	Ref	
French, born to at least one foreign parent				
	1.10 [0.79–1.54]	1.32 [0.89–1.95]	1.54 [1.11–2.17]	
Foreigner	1.19 [0.72–1.99]	1.67 [1.06–2.61]	1.95 [1.30–2.93]	
Model 2: Socioeconomic status				
Educational level				0.0119
Tertiary	Ref	Ref	Ref	
Secondary	1.23 [0.89–1.69]	1.25 [0.91–1.71]	1.61 [1.24–2.10]	
Primary or none	1.94 [1.09–3.48]	1.67 [0.92–3.01]	2.26 [1.39–3.69]	
Employment status				<0.001
Employed	Ref	Ref	Ref	
Unemployed	1.40 [0.65–3.00]	0.87 [0.43–1.76]	1.20 [0.61–3.36]	
Inactive	2.08 [1.59–2.71]	1.01 [0.77–1.34]	1.25 [0.98–1.61]	
Income per consumption unit (quintiles)				<0.001
1st	Ref	Ref	Ref	
2nd	0.96 [0.61–1.51]	1.11 [0.70–1.76]	1.05 [0.68–1.61]	
3rd	1.39 [0.87–2.20]	0.92 [0.53–1.60]	0.82 [0.51–1.29]	
4th	1.63 [0.96–2.77]	1.11 [0.62–1.99]	0.91 [0.57–1.47]	
5th	1.80 [1.20–2.70]	0.79 [0.51–1.21]	0.90 [0.56–1.43]	
Health insurance status				0.0058
SHI+VHI	Ref	Ref	Ref	
special insurance for the poor				
	0.55 [0.24–1.22]	1.45 [0.69–3.05]	0.59 [0.26–1.34]	
SHI only	0.63 [0.29–1.37]	2.72 [1.39–5.35]	1.78 [0.89–3.56]	
None	1.01 [0.19–5.41]	1.60 [0.41–6.32]	0.28 [0.04–2.03]	
Model 3: Stance regarding health and medicine				

Table 3. Cont.

	Type 1	Type 2	Type 3	p*
	OR [CI 95%]	OR [CI 95%]	OR [CI 95%]	
General attitude toward medical consultation				<0.001
As a last resort	Ref	Ref	Ref	
As soon as not feeling well				
	1.55 [1.19–2.01]	0.33 [0.23–0.47]	1.15 [0.84–1.58]	
Having a relative or a friend suffering from a severe condition				0.0058
No	Ref	Ref	Ref	
Yes	0.97 [0.71–1.32]	0.62 [0.44–0.85]	0.67 [0.48–0.90]	
Having medical professionals among relatives				0.3301
No	0.96 [0.77–1.20]	1.19 [0.90–1.59]	1.20 [0.94–1.55]	
Yes	Ref	Ref	Ref	
Model 4: Social integration				
Isolation feeling				0.1295
Very supported	Ref	Ref	Ref	
Rather supported	1.11 [0.81–1.52]	0.94 [0.69–1.29]	1.01 [0.72–1.41]	
Rather isolated	0.88 [0.60–1.31]	0.68 [0.43–1.09]	0.90 [0.66–1.22]	
Very isolated	2.84 [1.08–7.50]	1.45 [0.46–4.58]	3.06 [1.06–8.84]	
Level of social support				0.0384
High	Ref	Ref	Ref	
Medium	2.32 [1.40–3.84]	1.03 [0.62–1.72]	1.43 [0.90–2.28]	
Low	1.72 [0.84–3.56]	1.11 [0.59–2.11]	1.28 [0.66–2.49]	
Frequency of social contacts (quartiles)				0.0370
1st	1.42 [0.96–2.13]	1.28 [0.87–1.90]	1.21 [0.80–1.82]	
2nd	1.29 [0.90–1.87]	1.20 [0.87–1.64]	1.32 [0.92–1.89]	
3rd	1.03 [0.70–1.53]	0.71 [0.51–1.00]	0.75 [0.46–1.22]	
4th	Ref	Ref	Ref	
Model 5: Health status				
Perceived health status				<0.001
Good	Ref	Ref	Ref	
Average	1.40 [0.97–2.03]	0.53 [0.33–0.84]	1.18 [0.84–1.67]	
Bad	1.65 [0.71–3.84]	0.75 [0.24–2.36]	0.78 [0.33–1.87]	
Chronic or long standing health problem				<0.001
No	Ref	Ref	Ref	
Yes	2.91 [2.27–3.71]	0.48 [0.34–0.67]	1.34 [1.02–1.76]	
Long standing activity limitation				<0.001
No	Ref	Ref	Ref	
Yes	2.24 [1.44–3.48]	0.56 [0.36–0.88]	1.34 [0.90–2.00]	

*p value for overall trend. Type 4 is the reference type for the estimation of all the Odd-Ratios.

doi:10.1371/journal.pone.0115064.t003

medical professionals among relatives, feelings of isolation and frequency of social contacts.

The five multivariate multinomial models are successively presented in Table 3, with Type 4 being considered as the reference type. As in the univariate analysis, many significant associations were observed. For instance, the probability of belonging to Type 1 increased with age, and women were most likely to belong to Type 4. Foreigners was more likely to belong to Type 3 (OR=1.95, 95% CI= [1.30–2.93]), as did individuals with a low education level (primary school or none; OR=2.26, 95% CI= [1.39–3.69]). Inactive or wealthy people were most likely to belong to Type 1 (OR=2.08, 95% CI= [1.59–2.71], and OR=1.80, 95% CI= [1.20–2.70], respectively). Individuals with a SHI only had the highest probability of belonging to Type 2. Referring to their GP for the slightest health issue was an attitude associated with Type 1 (OR=1.55, 95% CI= [1.19–2.01]), while the opposite attitude was associated with Type 2. Among the three variables related to social integration, only the frequency of social contacts tended to be associated with the type of healthcare utilisation; although the association was not significant, the point estimate indicated that frequent social contacts were more characteristic of Type 4 people. In terms of health status, reporting a chronic condition was significantly associated with Types 1 and 3 (OR=2.91, 95% CI= [2.27–3.71], and OR=1.34, 95% CI= [1.02–1.76], respectively), while reporting a good health status tended to be more frequent in Type 2 (OR=0.53, CI= [0.33–0.84]).

Discussion

In this study, we took advantage of a database which was representative of the general population of French-speaking adults in the Paris metropolitan area. Because data were recorded from face-to-face interviews, independently from medical registers or medical consumption records, our sample has the advantage of taking into account non-users of healthcare. Social and subjective variables are particularly richly documented in the SIRS cohort, which was originally designed to study social inequalities in health and access to healthcare. We used an original and methodologically robust approach to identify homogenous and consistent types of healthcare system users.

We identified four different types of healthcare user through this approach. The findings of the cluster analysis exhibited strong robustness in terms of sensitivity to parameter tuning and of group stability. One type of user (Type 1) typically consisted of elderly individuals of French origin, who were wealthy but unhealthy, inactive and socially isolated, and who benefited from a good health insurance and took advantage of all kinds of healthcare services, which they used extensively. Type 4 was typically constituted by young, working women of French origin, with a high educational level, who tended to be wealthy and healthy, socially integrated and supported and fully insured. These users were the most likely to frequently consult specialists in the community and to make use of non-conventional care. A

third type (Type 2) was constituted by young men, frequently foreigners, who tended to be unemployed and rather poor, healthy but with a mediocre access to health insurance, and who had the lowest utilisation of healthcare services. The last type (Type 3) was constituted of a population of diverse ages, often foreigners, with a poor educational level and low incomes. These users were typically inactive, with a mediocre health insurance, rather socially isolated and unhealthy, and principally used GP services or emergency healthcare.

Our study has certain limitations. Firstly, we dealt with declarative data, without any linkage to medical records or objective measures, so we were unable to estimate possible reporting and recall biases. It is also specific to the French healthcare system and we can thus make no direct comparison with or extrapolate to healthcare systems of other countries, each of which has specific regulation policies (especially in terms of gate-keeping, extent of public and supplementary health insurance and out-of-pocket payments) and provision of healthcare services. For example, if individuals with Type 4 profile present a higher use of specialists, it is partly both because most of them are women and consult a gynaecologist every year. In France, gynaecologists represent general ''women's health'' doctors, responsible for all aspects of gynaecological follow-up (including contraception and cervical smears) and who are directly accessible without going through a GP. The place of the gynaecologist in the French healthcare system is atypical and not found in many other countries. Also, despite the existence of a universal basic health insurance, income and insurance status may still influence access to healthcare, which could account for the Type 3 profile. Apart from individuals with the lowest income levels and those suffering from costly chronic diseases, approximately 30% of health expenses are supported by patients (co-payment). They may be reimbursed by a voluntary (supplementary) insurance; but sometimes only partially, according to their contract. In many situations, people also have to pay upfront and are then reimbursed by the basic public health insurance. The gate-keeping system can be bypassed, although this incurs a higher cost or lower reimbursement for patients. In addition, access to IAS and prescription of paraclinical tests such as imaging or laboratory analyses can be prescribed without consulting a GP during an emergency unit consultation instead (with the possibility to get them at the same time rather than in a second step after the GP consultation).

Moreover, even in France, the Paris metropolitan area region is not representative of the whole country, being very urbanised, more wealthy on average and with a higher density of medical provision than the rest of the country, but also with much more social inequalities and spatial segregation than other French regions [43].

Technically speaking, the robustness and stability of the four clusters could result, at least in part, from too many constraints in the analysis methods. In other words, the identified clusters may grossly reflect reality, but lack accuracy. If so, this is likely to be linked to the intrinsic geometric assumptions underlying the clustering methods we used. The overall shape of the groups identified cannot represent complicated configurations, such as reticulated patterns, and the

clusters generated by the model are spheroid with little scope for interpenetration. Moreover, we may have encountered a lack of statistical power for some underrepresented categories such as the utilisation of emergency resources.

From the health inequality point of view, our results help clarify some of the differences in behaviours with respect to healthcare system use and to opportunities for healthcare. While it is usually assumed that social inequalities in access to healthcare mainly stem from economic inequalities [44–46], it would have been expected that the introduction of universal health insurance in France in 1999 should have removed such barriers [47]. However, it is clear that this has not happened systematically [16, 43, 48]. In fact, the social causes and processes underlying inequalities in access to healthcare are complex, going beyond purely economical or materialistic factors, and involving different psychosocial and behavioural factors as well [49, 50]. Our study may help unravel some of this complexity and diversity. Indeed, we observed several associations between the type of healthcare resource utilisation and social factors previously found to be associated with healthcare indicators, such as objective or perceived health status [45, 51] or, more broadly, health expectations and perceived needs [52], social capital [53] or social integration [54]. These determinants coexist, but contribute to different extents to the four types of utilisation. For example, health status, measured by chronic diseases and functional limitations, was indeed associated with greater use of the healthcare system, together with higher educational level and higher income. Stance regarding healthcare was also found to influence extent of use of the healthcare system and established differences regarding gender and healthcare use were observed. Among all the variables evaluated, only having medical professionals among relatives was not significantly associated with profiles; this may be explained by this question being too general, as it could be interpreted in a wide variety of different ways with respect to closeness, confidence, availability and the professional skills involved. With respect to social integration, the frequency of social contacts appeared to be poorly discriminant. In this context, we believe that the ''crude'' frequency of social contacts, without any further details on their frequency, quality, context or content, is less meaningful than direct interrogation of global and subjective feelings of isolation. For the latter, significant differences were found for people reporting a ''very isolated'' status in Types 1 and 3 (Table 3). Taken together, these results, which are consistent with the literature, provide some face validity of our typology.

As mentioned above, this typology can be interpreted according to Andersen's BMHSU, in terms of equity in access to healthcare services. When looking at the variables that discriminated the four types of user best ($OR \geq 2$ or ≤ 0.5), we observed that the most important predisposing factors for the pattern of healthcare utilisation were age, gender, origins, educational level, a feeling of social isolation and the general attitude toward medical consultation (Types 2 and 4 only), whereas the most prominent enabling factors were the feeling of social isolation and health insurance (Types 2 and 4 only).

Individuals corresponding to the Type 1 profile need to have access to services due to a higher prevalence of chronic conditions and functional limitations, and

they also use these services, both because they can afford to (they have adequate financial resources and more time to access services) and because they present habits and perceptions that make them prone to use them. Our study provides no information on whether their access to healthcare meet all their needs, but their pattern of healthcare use does not seem to be explained in terms of predisposing factors such as gender or educational level. Individuals corresponding to the Type 2 profile, who are in majority young males, have a low level of healthcare utilisation which reflects their low perceived needs. At the same time, these individuals are those with the highest proportion of basic insurance status only. This is an obstacle to access to healthcare which is not very reassuring regarding equity in access to health services, particularly since young people may underestimate their real health needs. Individuals corresponding to the Type 3 profile present low rates of use and multiple negative predisposing factors. For example, they are more likely to be foreigners, with a lower educational level and low financial resources. On the other hand, they are likely to suffer from chronic conditions and functional limitations. It is for this profile that the healthcare system is the most likely to be inequitable. Indeed, these individuals predominantly use services from GPs and emergency units, and far less often from specialists. Individuals corresponding to the Type 4 profile have few expressed needs but use services, especially from GPs and DAS, intensively, preferring community care. These individuals have the adequate resources to use services, both economically (income, health insurance) and culturally (female gender, higher educational level, higher social support). In conclusion, of the four profiles described, one (Type 3) and possibly two (Type 2) profiles are in a situation where the French healthcare (and insurance) system is the most likely to be inequitable.

Finally, we demonstrated that the method used was able to reveal stable and meaningful structures in our data without resorting to the usual reductionism of classical studies on healthcare utilisation. For this reason, we think that a similar multivariate clustering method would merit replication in other datasets derived from other contexts, such as non-urban populations or countries with other healthcare systems, in order to confirm or refine our findings.

Author Contributions

Conceived and designed the experiments: TL IP PC. Performed the experiments: TL IP PC. Analyzed the data: TL CR IP PC. Contributed reagents/materials/analysis tools: TL CR IP PC. Contributed to the writing of the manuscript: TL CR IP PC.

References

1. **Karanikolos M, Mladovsky P, Cylus J, Thomson S, Basu S, et al.** (2013) Financial crisis, austerity, and health in Europe. The Lancet 381: 1323–1331. doi:10.1016/S0140-6736(13)60102-6.

2. **Geitona M, Zavras D, Kyriopoulos J** (2007) Determinants of healthcare utilization in Greece: implications for decision-making. Eur J Gen Pr 13: 144–150. doi:10.1080/13814780701541340.

3. **Parkhurst J** (2008) Understanding determinants of health service use from a systems perspective. J Health Serv Res Policy 13: 122–123. doi:10.1258/jhsrp.2008.008019.

4. **Balabanova D, McKee M, Pomerleau J, Rose R, Haerpfer C** (2004) Health service utilization in the former soviet union: evidence from eight countries. Health Serv Res 39: 1927–1950. doi:10.1111/j.1475-6773.2004.00326.x.

5. **Treanor C, Donnelly M** (2012) An international review of the patterns and determinants of health service utilisation by adult cancer survivors. BMC Health Serv Res 12: 316. doi:10.1186/1472-6963-12-316.

6. **Dias SF, Severo M, Barros H** (2008) Determinants of healthcare utilization by immigrants in Portugal. BMC Health Serv Res 8: 207. doi:10.1186/1472-6963-8-207.

7. **Cabieses B, Tunstall H, Pickett KE, Gideon J** (2012) Understanding differences in access and use of healthcare between international immigrants to Chile and the Chilean-born: a repeated cross-sectional population-based study in Chile. Int J Equity Heal 11: 68. doi:10.1186/1475-9276-11-68.

8. **Stewart M, Reutter L, Makwarimba E, Rootman I, Williamson D, et al.** (2005) Determinants of health-service use by low-income people. Can J Nurs Res Rev Can Rech En Sci Infirm 37: 104–131.

9. **Droomers M, Westert GP** (2004) Do lower socioeconomic groups use more health services, because they suffer from more illnesses? Eur J Public Health 14: 311–313.

10. **Habicht J, Kunst AE** (2005) Social inequalities in healthcare services utilisation after eight years of healthcare reforms: a cross-sectional study of Estonia, 1999. Soc Sci Med 1982 60: 777–787. doi:10.1016/j.socscimed.2004.06.026.

11. **Fleury M-J, Grenier G, Bamvita J-M, Perreault M, Kestens Y, et al.** (2012) Comprehensive determinants of health service utilisation for mental health reasons in a Canadian catchment area. Int J Equity Heal 11: 20. doi:10.1186/1475-9276-11-20.

12. **Carret MLV, Fassa AG, Kawachi I** (2007) Demand for emergency health service: factors associated with inappropriate use. BMC Health Serv Res 7: 131. doi:10.1186/1472-6963-7-131.

13. **Busato A, Künzi B** (2008) Primary care physician supply and other key determinants of healthcare utilisation: the case of Switzerland. BMC Health Serv Res 8: 8. doi:10.1186/1472-6963-8-8.

14. **Finlayson TL, Gansky SA, Shain SG, Weintraub JA** (2010) Dental utilization among Hispanic adults in agricultural worker families in California's Central Valley. J Public Health Dent 70: 292–299. doi:10.1111/j.1752-7325.2010.00184.x.

15. **Harrington DW, Wilson K, Rosenberg M, Bell S** (2013) Access granted! barriers endure: determinants of difficulties accessing specialist care when required in Ontario, Canada. BMC Health Serv Res 13: 146. doi:10.1186/1472-6963-13-146.

16. **Glazier RH, Agha MM, Moineddin R, Sibley LM** (2009) Universal health insurance and equity in primary care and specialist office visits: a population-based study. Ann Fam Med 7: 396–405. doi:10.1370/afm.994.

17. **Lipsitz LA** (2012) Understanding healthcare as a complex system: The foundation for unintended consequences. JAMA 308: 243–244. doi:10.1001/jama.2012.7551.

18. **Armstrong JJ, Zhu M, Hirdes JP, Stolee P** (2012) K-Means Cluster Analysis of Rehabilitation Service Users in the Home Healthcare System of Ontario: Examining the Heterogeneity of a Complex Geriatric Population. Arch Phys Med Rehabil. Available: http://www.ncbi.nlm.nih.gov/pubmed/22705468. Accessed 2012 Jul 2.

19. **Westert GP, Satariano WA, Schellevis FG, van den Bos GA** (2001) Patterns of comorbidity and the use of health services in the Dutch population. Eur J Public Health 11: 365–372.

20. **Newcomer SR, Steiner JF, Bayliss EA** (2011) Identifying subgroups of complex patients with cluster analysis. Am J Manag Care 17: e324–332.

21. **Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, et al.** (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci U S A 96: 2907–2912.

22. **Dalton L, Ballarin V, Brun M** (2009) Clustering algorithms: on learning, validation, performance, and applications to genomics. Curr Genomics 10: 430–445. doi:10.2174/138920209789177601.

23. **Hwang T, Atluri G, Xie M, Dey S, Hong C, et al.** (2012) Co-clustering phenome-genome for phenotype classification and disease gene discovery. Nucleic Acids Res. Available: http://www.nar.oxfordjournals.org/cgi/doi/10.1093/nar/gks615. Accessed 2012 Aug 24.

24. **Alexandrov T, Kobarg JH** (2011) Efficient spatial segmentation of large imaging mass spectrometry datasets with spatially aware clustering. Bioinforma Oxf Engl 27: i230–238. doi:10.1093/bioinformatics/btr246.

25. **Guevara P, Poupon C, Rivière D, Cointepas Y, Descoteaux M, et al.** (2011) Robust clustering of massive tractography datasets. NeuroImage 54: 1975–1993. doi:10.1016/j.neuroimage.2010.10.028.

26. **Wang Y, Shi H, Ma S** (2011) A new approach to the detection of lesions in mammography using fuzzy clustering. J Int Med Res 39: 2256–2263.

27. **Sutherland ER, Goleva E, King TS, Lehman E, Stevens AD, et al.** (2012) Cluster analysis of obesity and asthma phenotypes. PloS One 7: e36631. doi:10.1371/journal.pone.0036631.

28. **Lochner C, Hemmings SMJ, Kinnear CJ, Niehaus DJH, Nel DG, et al.** (2005) Cluster analysis of obsessive-compulsive spectrum disorders in patients with obsessive-compulsive disorder: clinical and genetic correlates. Compr Psychiatry 46: 14–19. doi:10.1016/j.comppsych.2004.07.020.

29. **Conry MC, Morgan K, Curry P, McGee H, Harrington J, et al.** (2011) The clustering of health behaviours in Ireland and their relationship with mental health, self-rated health and quality of life. BMC Public Health 11: 692. doi:10.1186/1471-2458-11-692.

30. **Claudot F, Alla F, Fresson J, Calvez T, Coudane H, et al.** (2009) Ethics and observational studies in medical research: various rules in a common framework. Int J Epidemiol 38: 1104–1108. doi:10.1093/ije/dyp164.

31. **Martin-Fernandez J, Grillo F, Parizot I, Caillavet F, Chauvin P** (2013) Prevalence and socioeconomic and geographical inequalities of household food insecurity in the Paris region, France, 2010. BMC Public Health 13: 486. doi:10.1186/1471-2458-13-486.

32. **Chevreul K, Durand-Zaleski I, Bahrami S, Hernández-Quevedo C, Mladovsky P** (2010) France: Health system review. Health Syst Transit 12: 1–219.

33. **Vallée J, Cadot E, Grillo F, Parizot I, Chauvin P** (2010) The combined effects of activity space and neighbourhood of residence on participation in preventive health-care activities: The case of cervical screening in the Paris metropolitan area (France). Health Place 16: 838–852. doi:10.1016/j.healthplace.2010.04.009.

34. **Rondet C, Soler M, Ringa V, Parizot I, Chauvin P** (2013) The role of a lack of social integration in never having undergone breast cancer screening: Results from a population-based, representative survey in the Paris metropolitan area in 2010. Prev Med. doi:10.1016/j.ypmed.2013.06.016.

35. **The EHEMU/EHLEIS team** (2010) The Minimum European Health Module. Background documents. EHEMU Technical report.

36. **Cox B, Oyen HV, Cambois E, Jagger C, Roy S le, et al.** (2009) The reliability of the Minimum European Health Module. Int J Public Health 54: 55–60. doi:10.1007/s00038-009-7104-y.

37. **Andersen R** (1995) Revisiting the behavioral model and access to medical care: does it matter? J Health Soc Behav 36: 1–10.

38. **Babitsch B, Gohl D, von Lengerke T** (2012) Re-revisiting Andersen's Behavioral Model of Health Services Use: a systematic review of studies from 1998–2011. Psychosoc Med 9: Doc11.

39. **Theodoridis S, Koutroumbas K** (2008) Pattern Recognition. 4th Revised edition. Academic Press Inc. 984 p.

40. **Kaufman L, Rousseeuw PJ** (2005) Finding groups in data: an introduction to cluster analysis. Hoboken, N.J.: Wiley.

41. **Jain AK** (2010) Data clustering: 50 years beyond K-means. Pattern Recognit Lett 31: 651–666. doi:10.1016/j.patrec.2009.09.011.

42. **Monti S, Tamayo P, Mesirov J, Golub T** (2003) Consensus clustering – A resampling-based method for class discovery and visualization of gene expression microarray data. MACHINE LEARNING, FUNCTIONAL GENOMICS SPECIAL ISSUE. 91–118.

43. **Gusmano MK, Weisz D, Rodwin VG, Lang J, Qian M, et al.** (2013) Disparities in access to healthcare in three French regions. Heal Policy Amst Neth. doi:10.1016/j.healthpol.2013.07.011.

44. **Mackenbach JP, Kunst AE, Cavelaars AE, Groenhof F, Geurts JJ** (1997) Socioeconomic inequalities in morbidity and mortality in western Europe. The EU Working Group on Socioeconomic Inequalities in Health. Lancet 349: 1655–1659.

45. **Mackenbach JP** (2012) The persistence of health inequalities in modern welfare states: The explanation of a paradox. Soc Sci Med 75: 761–769. doi:10.1016/j.socscimed.2012.02.031.

46. **Marmot MG, Wilkinson RG** (2006) Social determinants of health. Oxford; New York: Oxford University Press.

47. **Smith PC** (2013) Universal health coverage and user charges. Heal Econ Policy Law 8: 529–535. doi:10.1017/S1744133113000285.

48. **Watanabe R, Hashimoto H** (2012) Horizontal inequity in healthcare access under the universal coverage in Japan; 1986–2007. Soc Sci Med 1982 75: 1372–1378. doi:10.1016/j.socscimed.2012.06.006.

49. **Evans RG, Barer ML, Marmor TR** (1994) Why are some people healthy and others not?: the determinants of health of populations. New York: A. de Gruyter.

50. **Berkman LF, Kawachi I** (2000) Social Epidemiology. Oxford University Press. 414 p.

51. **Wang T-F, Shi L, Nie X, Zhu J** (2013) Race/ethnicity, insurance, income and access to care: the influence of health status. Int J Equity Heal 12: 29. doi:10.1186/1475-9276-12-29.

52. **Strecher V** (1997) The health belief model and health behavior. Handbook of health behaviour research, Personal and social determinants. New York: Gochman DS. 71–91.

53. **Pearce N, Davey Smith G** (2003) Is social capital the key to inequalities in health? Am J Public Health 93: 122–129.

54. **Kawachi I, Berkman LF** (2001) Social ties and mental health. J Urban Heal Bull New York Acad Med 78: 458–467. doi:10.1093/jurban/78.3.458.

Annexe 3.3 Typologie des repas en Ile-de-France

PLOS | ONE

RESEARCH ARTICLE

Is There Still a French Eating Model? A Taxonomy of Eating Behaviors in Adults Living in the Paris Metropolitan Area in 2010

Julien Riou[1,2]*, Thomas Lefèvre[1,2,3], Isabelle Parizot[4], Anne Lhuissier[5,6], Pierre Chauvin[1,2]

1 Sorbonne Universités, UPMC Univ Paris 06, UMR_S 1136, Pierre Louis Institute of Epidemiology and Public Health, Department of social epidemiology, F-75013 Paris, France, **2** INSERM, UMR_S 1136, Pierre Louis Institute of Epidemiology and Public Health, Department of social epidemiology, F-75013 Paris, France, **3** AP-HP, Hôpital Jean-Verdier, Department of Forensic Medicine, F-93140 Bondy, France **4** CNRS, UMR 8097, Centre Maurice Halbwachs, Research Team on Social Inequalities, F-75014 Paris, France, **5** INRA, UR1303 ALISS, F-94205 Ivry sur Seine Cedex, France, **6** University of Oxford, Department of Sociology, Manor Road, Oxford OX1 3UQ, United Kingdom

* julien.riou.k@gmail.com

Citation: Riou J, Lefèvre T, Parizot I, Lhuissier A, Chauvin P (2015) Is There Still a French Eating Model? A Taxonomy of Eating Behaviors in Adults Living in the Paris Metropolitan Area in 2010. PLoS ONE 10(3): e0119161. doi:10.1371/journal.pone.0119161

Academic Editor: Hiroaki Matsunami, Duke University, UNITED STATES

Received: July 28, 2014

Accepted: January 11, 2015

Published: March 3, 2015

Data Availability Statement: Data are available from Dryad: doi:10.5061/dryad.5m2q0 (http://datadryad.org/review?doi = doi:10.5061/dryad.5m2q0).

Funding: This study was supported by the National Research Agency (ANR) within the framework of the National Food Research Program (grant number ANR-07-PNRA-002-01; URL: http://www.agence-nationale-recherche.fr). The SIRS survey was also supported by the Institute for Public Health Research (IReSP) (grant number 2008-87; URL: http://www.iresp.net) and the French Interministerial Committee for Urban Affairs (grant number 2009-273; URL:

Abstract

Background

Meal times in France still represent an important moment in everyday life. The model of three rigorously synchronized meals is still followed by a majority of people, while meal frequencies have flattened in other European or North-American countries. We aimed to examine the "French model" of eating behavior by identifying and characterizing distinct meal patterns.

Methods

Analyses were based on data from the SIRS cohort, a representative survey of the adult population in the Paris area. A clustering algorithm was applied to meal variables (number, time, location, with whom the meal is usually shared and activities associated with meals). Regression models were used to investigate associations between patterns and socio-demographic, social environment and perceived food quality variables.

Results

Five different patterns were identified among 2994 participants. The first three types (prevalence 33%, 17% and 24%) followed a three-meal pattern, with differences in locations and social interactions mainly related to time constraints and age. More marked differences were observed in the remaining two types. In the fourth type (prevalence 13%), individuals ate one or two meals per day, often with an irregular schedule, at home and in front of the television. They frequently were unemployed and had lower income. Breakfast skipping, increased snacking and a low adherence to dietary guidelines suggested that this behavior might have health consequences. In the fifth type (12%), people also ate two meals or less per day, possibly with the same consequences on food quality. However, meals were often

http://www.ville.gouv.fr/?le-commissariat-general-a-l,9). 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.

taken outside the home, in social settings, and individuals following this pattern were typically active, integrated, young people, suggesting that this pattern might be an adaptation to a modern urban lifestyle.

Conclusions

While a majority of the population still follows the three-meal pattern, our analysis distinguished two other eating patterns associated with specific sociological profiles.

Introduction

Meal times in France still represent an important moment in everyday life. The three-meal pattern, with breakfast between 7 and 8:30 am, lunch between 12 and 1:30 pm and supper between 7 and 8:30 pm, is still followed by a majority of French people, while meal frequencies have flattened in other European or North-American countries [1–4]. Meals, particularly the breakfast and evening meal, are still often taken at a slow pace, principally at home. Snacking between meals is scarce. Even so, some evolution in meal patterns has been detected, notably a decrease in breakfast frequency, a growing proportion of meals taken alone [3–5], and a simplification of meals in terms of content as shown by Poulain [6].

The three meals a day pattern has been linked to several health benefits, such as a lower prevalence of obesity and a higher consumption of fruits and vegetables [3,7–10]. Nonetheless, some of these associations, particularly with respect to children, have recently been challenged [11]. Meals are also an important occasion for socializing, sharing and consolidation of social ties. The relative preservation of this pattern is not consistent with the idea of destructuration of French eating habits as a consequence of growing globalization and standardization [6,12–14]. Preservation does not necessarily imply plain conservatism, as illustrated by the reactions to the introduction of fast-food chains in the 1970s: France reacted to the expansion of American fast-food by adapting, to a certain extent, traditional national products to the fast-food formula [15]. Eating behaviors are influenced by numerous socio-demographic and behavioral variables, and French society is evolving just as fast as others. Increasing employment of women and general family dynamics, the later age of having a first child, immigration, and shorter lunch breaks all strongly influence eating habits [15–17].

In order to identify typical meal patterns, it is important to take into consideration a wide range of variables and to apply specific powerful statistical approaches, including clustering analysis. Dietary behaviors, with regard to their potential association with obesity, have to date largely been studied in children and adolescents [17]. In this study, we aimed to examine the so-called "French model" of eating behavior in adults in the Paris area, by identifying and characterizing distinct meal patterns.

Methods

Ethics statement

The SIRS cohort (a French acronym for "*Health, inequalities and social ruptures*") study is a collaborative project between the French National Institute for Health and Medical Research (INSERM) and the National Centre for Scientific Research (CNRS), and received legal authorization from two French national authorities for non-biomedical research: the *Comité consultatif sur le traitement de l'information en matière de recherche dans le domaine de la*

santé (CCTIRS) and the *Commission nationale de l'informatique et des libertés* (CNIL). The participants provide their verbal informed consent. Written consent was not necessary because this survey did not fall into the category of biomedical research (as defined by French law).

Study population and design

Analyses were based on data from the 2010 wave of the SIRS cohort, a representative socio-epidemiological survey of the French-speaking adult (≥18) population conducted since 2005 in the Paris metropolitan area (population 6.5 million). The survey employed a stratified, 3-level random sampling procedure. In the first step, fifty census blocks with approximately 2000 inhabitants each were selected, over-representing the poorest neighborhoods. In the second step, sixty households were randomly selected from each surveyed census block. In the final step, one adult was chosen from each household by the birthday method. A questionnaire was administered face-to-face during home visits in 2005 and 2010, and detailed questions concerning meal structures and characteristics were introduced in 2010. In 2010, 47% of the original 2005 respondents were interviewed again face-to-face at home (2.6% had died, 1.8% were too sick to answer our questions, 2.7% were absent during the survey period, 13.9% had moved out of the 50 surveyed census blocks, 18.4% declined to participate, and 13.4% were lost to follow-up). The individuals who could not be reinterviewed were replaced by a random procedure similar to the one used in 2005, up to a final sample size of 60 adults by census block. The refusal rate in the newly contacted individuals was 29% (the same as in 2005). The methodology of the SIRS study has been further described elsewhere [18].

Meal characteristics

Participants were asked about their most common meal-related habits in a typical week. In order to avoid any normative approach, meals were defined as "eating events" considered as meals eaten by the participants themselves [19]. Information concerning meals was collected without reference to a three-meal pattern, by referring to meals by their rank instead of their usual names (for further details, see [4]). We collected data on several meal characteristics, such as number, time, location (home, working place, restaurant), with whom the meal is usually taken (alone, with family members, with colleagues or friends) and activities associated with meals (television, radio, computer, reading, chatting). All those variables were categorized to be included in the clustering algorithm. Meal time was converted into 6 new variables displaying the occurrence of a meal during a time period (using as breaks: 12:30 a.m.; 5:30 a.m.; 10:30 a.m.; 2:30 p.m.; 6:30 p.m. and 10:30 p.m.). The intervals were chosen to correspond to commonly used meal times in France, but were large enough to adapt to diverse situations. Information on the other meal characteristics (location, other participants and activities) was included as proportions of daily meals displaying this characteristic, as it was important to avoid dependency on the number of meals.

Clustering methods

In order to identify the different meal patterns, the partitioning around medoïds (PAM) algorithm with Manhattan distance was used as a reference analysis and applied to all the meal characteristic variables at once [20]. In order to determine the number of clusters, we used a resampling-based method and cluster-robustness approach called consensus clustering [21]. Sensitivity analyses were performed using three other clustering strategies: the same PAM algorithm with Euclidian distance, a *k*-means algorithm, and a hierarchical clustering algorithm. Analyses were conducted with *R* 2.15.3, using the *clusterCons* and *ggplot2* packages [22–24].

Factors associated with meal patterns

Three different categories of factors were explored as being potentially associated with meal patterns. First, we considered social, demographic and economic characteristics such as gender, age (in five classes: 18–29; 30–44; 45–59; 60–74; 75 and over), level of education (in three classes: none or primary; secondary; higher degrees), occupation (five classes: employed; student; unemployed; retired; stay at home), household income per consumption unit (quartiles), living in an underprivileged neighborhood (according to the definition applied by the French government to target urban renewal programs and specific welfare policies), and origin (distinguishing between French, born to two French parents; French, born to at least one foreign parent; foreigner).

The second category depicted the social environment of the participant. It included the household type (four classes: single person household; couple with or without children; single-parent family; household with several unrelated individuals or families), the presence of a child under 16 years of age at home, and the feeling of loneliness.

The third category was made up of food-related characteristics, such as the occurrence of daily snacking (as defined by the participants themselves), dissatisfaction concerning food, level of involvement in meal-related decisions and in meal preparation, whether food quality was considered to be negatively affected by the participant's lifestyle or financial issues. French national public health recommendations of eating five fruit or vegetables and three dairy products per day were also explored.

Associations between each of the socio-demographic and social environment variables and meal patterns were investigated using univariate multinomial regression models. Then, associations between each of the food-related variables and meal patterns were estimated with multinomial logistic regression models adjusted for socio-demographic and social environment characteristics. We reported unadjusted and adjusted odd ratios (OR) with their 95% confidence intervals and p-values. Analyses were conducted with *R* 3.0.3, with the *nnet* package [22,25].

Prevalence estimates

To estimate prevalence in the reference population, some proportions presented in this article were weighted to account for the complex sample design (notably, the design effect associated with cluster sampling and the overrepresentation of poorer neighborhoods) and for the post-stratification adjustment for age and gender according to the general population census data.

Results

We retrieved data on meal characteristics for 2994 of the 3006 participants in the survey. We estimated that 66%, 24% and 8% of the Paris area adult population had three, two and four or more meals, respectively. In most cases, mealtimes matched the three timeslots for breakfast (5:30 a.m. to 10:25 a.m.), lunch (11:30 a.m. to 2:25 p.m.) and dinner (6:30 p.m. to 9:25 p.m.). When only two meals were declared, it was mainly because the participant skipped breakfast. For those who ate four or more meals, the additional meal was generally taken around 4:00 p.m. Further description of meal characteristics for the full sample are presented in table 1.

The five types of meal patterns

Cluster robustness analysis indicated that the optimal number of clusters was five, even though the robustness of the division into four clusters was very close (Fig. 1). The examination of the relationship between the 4- and 5-way classifications revealed that while three groups were

Table 1. Characteristics of meals in the whole sample and according to the five types of meal patterns.

	General		Type 1		Type 2		Type 3		Type 4		Type 5	
	n = 2994		n = 875 (29%)		n = 672 (22%)		n = 698 (23%)		n = 440 (15%)		n = 309 (10%)	
	n	%	%	*p**	%	*p**	%	*p**	%	*p**	%	*p**
Number of meals per day	2994			<0.001		<0.001		<0.001		<0.001		<0.001
1	109	3.6%	0%		0.3%		0.1%		18%		8.7%	
2	680	22.7%	1.6%		5.4%		2.9%		78%		86.4%	
3	1997	66.7%	89.9%		83.6%		88.5%		3.6%		4.5%	
4	189	6.3%	7.3%		10%		7.9%		0.5%		0.3%	
5	17	0.6%	1%		0.7%		0.4%		0%		0%	
6	2	0.1%	0.1%		0%		0.1%		0%		0%	
12:30 a.m. to 5:29 a.m.	2994			0.553		0.367		0.073		0.662		0.296
No	2925	97.7%	97.1%		96.9%		99%		97%		99%	
Yes	69	2.3%	2.9%		3.1%		1%		3%		1%	
5:30 a.m. to 10:29 a.m.	2994			<0.001		<0.001		<0.001		<0.001		<0.001
No	789	26.4%	4.2%		4.8%		2%		93.2%		95.8%	
Yes	2205	73.6%	95.8%		95.2%		98%		6.8%		4.2%	
10:30 a.m. to 2:29 p.m.	2994			<0.001		0.809		0.004		<0.001		0.152
No	248	8.3%	2.2%		7.6%		4.7%		25%		11.3%	
Yes	2746	91.7%	97.8%		92.4%		95.3%		75%		88.7%	
2:30 p.m. to 6:29 p.m.	2994			0.342		0.003		0.81		0.935		0.422
No	2609	87.1%	88.8%		82.6%		88%		87.7%		89.6%	
Yes	385	12.9%	11.2%		17.4%		12%		12.3%		10.4%	
6:30 p.m. to 10:29 p.m.	2994			0.075		0.949		<0.001		<0.001		0.029
No	340	11.4%	8.9%		11.8%		5.7%		21.1%		16.2%	
Yes	2654	88.6%	91.1%		88.2%		94.3%		78.9%		83.8%	
10:30 p.m. to 12:29 a.m.	2994			0.994		0.925		0.015		0.047		0.58
No	2796	93.4%	93.5%		93%		96.1%		90.5%		91.9%	
Yes	198	6.6%	6.5%		7%		3.9%		9.5%		8.1%	
At home	2994			<0.001		<0.001		<0.001		<0.001		<0.001
≤25%	58	1.9%	1.6%		0.7%		0%		0.2%		12.3%	
26–50%	437	14.6%	17.6%		1.8%		0.9%		10.2%		71.2%	
51–75%	845	28.2%	79.1%		10.4%		10.6%		0.5%		2.3%	
>75%	1653	55.2%	1.7%		87%		88.5%		89.1%		14.2%	
At work	2994			<0.001		<0.001		<0.001		<0.001		<0.001
≤25%	2014	67.3%	19.5%		95.7%		98.7%		95.7%		29.1%	
26–50%	877	29.3%	72.2%		4%		1.3%		4.3%		61.5%	
51–75%	78	2.6%	7.9%		0.3%		0%		0%		2.3%	
>75%	25	0.8%	0.3%		0%		0%		0%		7.1%	
In a restaurant	2994			<0.001		<0.001		0.258		0.066		<0.001
≤25%	2747	91.8%	88.2%		97.2%		91.3%		95.5%		85.8%	

(*Continued*)

Table 1. (*Continued*)

	General		Type 1		Type 2		Type 3		Type 4		Type 5	
	n = 2994		n = 875 (29%)		n = 672 (22%)		n = 698 (23%)		n = 440 (15%)		n = 309 (10%)	
	n	%	%	*p**	%	*p**	%	*p**	%	*p**	%	*p**
26–50%	229	7.6%	10.9%		2.5%		8.7%		4.5%		11.7%	
51–75%	8	0.3%	0.8%		0.1%		0%		0%		0%	
>75%	10	0.3%	0.1%		0.1%		0%		0%		2.6%	
Eaten alone	2994			<0.001		<0.001		<0.001		<0.001		<0.001
≤25%	1127	37.6%	31.3%		1.5%		69.6%		32%		69.9%	
26–50%	831	27.8%	40.3%		8.3%		29.1%		30.5%		27.5%	
51–75%	419	14%	27.8%		24.7%		1.1%		0.5%		0%	
>75%	617	20.6%	0.6%		65.5%		0.1%		37%		2.6%	
Shared with family	2994			<0.001		<0.001		<0.001		<0.001		<0.001
≤25%	1089	36.4%	32.6%		76.9%		0.3%		41.6%		33%	
26–50%	716	23.9%	38.9%		15.5%		1.6%		24.3%		49.8%	
51–75%	563	18.8%	28.5%		7.1%		36.4%		1.1%		2.3%	
>75%	626	20.9%	0.1%		0.4%		61.7%		33%		14.9%	
Shared with colleagues/ friends	2994			<0.001		<0.001		<0.001		<0.001		<0.001
≤25%	1869	62.6%	16.1%		92.7%		90.2%		95.9%		17.7%	
26–50%	915	30.6%	69.7%		6%		8.2%		3.4%		63.6%	
51–75%	131	4.4%	12.4%		0.9%		1.4%		0%		2.3%	
>75%	73	2.4%	1.8%		0.4%		0.1%		0.7%		16.4%	
Eaten in front of television	2994			<0.001		<0.001		<0.001		<0.001		<0.001
≤25%	1029	34.4%	43.2%		24.9%		39.4%		18.4%		41.4%	
26–50%	851	28.4%	40.6%		18.8%		18.1%		18.2%		53.1%	
51–75%	535	17.9%	15.3%		33.6%		24.2%		0.7%		1%	
>75%	579	19.3%	0.9%		22.8%		18.3%		62.7%		4.5%	
Eaten in front of computer	2994			<0.001		0.108		<0.001		<0.001		0.464
≤25%	2839	94.8%	92%		95.1%		98.6%		94.8%		93.9%	
26–50%	121	4%	7.1%		3.1%		1.1%		3%		5.5%	
51–75%	18	0.6%	0.9%		1.3%		0.1%		0%		0%	
>75%	16	0.5%	0%		0.4%		0.1%		2.3%		0.6%	
Eaten while reading	2994			0.004		0.066		0.332		0.011		0.27
≤25%	2779	92.8%	91.3%		91.5%		93.7%		94.3%		95.8%	
26–50%	169	5.6%	7.2%		5.7%		5.7%		3.9%		3.6%	
51–75%	24	0.8%	1.5%		1.2%		0.4%		0%		0%	
>75%	22	0.7%	0%		1.6%		0.1%		1.8%		0.6%	
Eaten while listening to the radio	2994			<0.001		<0.001		0.163		<0.001		<0.001
≤25%	2047	68.4%	63.7%		56.3%		64.3%		86.6%		91.3%	
26–50%	656	21.9%	31%		23.8%		24.2%		8%		6.8%	

(*Continued*)

Table 1. (*Continued*)

	General		Type 1		Type 2		Type 3		Type 4		Type 5	
	n = 2994		n = 875 (29%)		n = 672 (22%)		n = 698 (23%)		n = 440 (15%)		n = 309 (10%)	
	n	%	%	*p**	%	*p**	%	*p**	%	*p**	%	*p**
51–75%	156	5.2%	4.9%		9.8%		6.7%		0%		0%	
>75%	135	4.5%	0.5%		10.1%		4.7%		5.5%		1.9%	
Eaten while chatting	2994			<0.001		<0.001		<0.001		<0.001		<0.001
≤25%	1016	33.9%	6.5%		80.2%		19.3%		63%		2.6%	
26–50%	728	24.3%	37.3%		17.3%		14.8%		19.8%		31.2%	
51–75%	580	19.4%	40.6%		1.8%		29.9%		0.2%		1%	
>75%	669	22.4%	15.7%		0.7%		36%		17%		65.3%	

** p-value for comparison to whole sample (chi-square test)*

doi:10.1371/journal.pone.0119161.t001

constantly retrieved in both classifications, the fourth group in the 4-way classification split into two when the 5-way classification was applied. Table 2 provides an example of this phenomenon for a single iteration. For this reason, we used the 5-way classification for all further analyses, while keeping in mind the higher-level proximity between groups 4 and 5.

Cluster robustness using the PAM algorithm with a Euclidian distance also indicated that the optimal number of clusters was five (Fig. 2). The division was very similar to the one obtained using PAM with a Manhattan distance for groups 1–3, while clusters 4 and 5 were now mixed, and a new cluster including individuals eating constantly 4 meals per day appeared. With the *k*-means algorithm, the optimal number of clusters was four, and clusters were less clearly defined: clusters 4 and 5 were mixed again, and intermediate groups mixing individuals from clusters 1–2 and 1–3 appeared. Finally, the hierarchical clustering algorithm was far less specific in determining a robust optimal number of clusters, and classification in five clusters was thus not interpretable.

The five different types of meal pattern are described in table 1. The first type represented 29% of the sample (corresponding to an estimated prevalence of 33% of the adult, French-speaking population of the Paris metropolitan area). We have labelled this group "*3 meals*, *often outside*". Participants commonly had three meals a day, within the usual timeslots. In comparison with the whole population, fewer meals were taken at home and with family members, though more meals were eaten at work or at the restaurant, with colleagues or friends. Activities undertaken during meals were less likely to be watching television, but rather chatting with other eaters.

In the second type, labelled "*3 meals*, *mostly alone at home*" (22% of the sample, corresponding to a prevalence of 17%), the three-meal pattern was also very frequent, though a higher proportion had a fourth meal (10%). Most people (87%) took their meals at home and in a large majority by themselves. A high proportion of these meals were eaten with television or radio, and very few were taken while chatting.

The third type of meal pattern was labelled "*3 meals*, *at home with family*" (23% of sample, corresponding to a prevalence of 24%). The three-meal pattern was generally adopted, and most of the meals were eaten at home, and shared with family members. Activities during meals were similar to the whole population.

The last two types both had a predominant 2-meal pattern, generally within the typical timeslots for lunch and supper, leading to the disappearance of the breakfast. The fourth type

was named "*1 or 2 meals, mostly at home with television*" (15% of sample, corresponding to a prevalence of 13%), and was noteworthy by the large proportion of meals taken while watching television. Even though most meals were taken at home, often with family members, chatting during meals was relatively uncommon. A significant minority of 18% only ate one meal per day. Among these, less than 4% ate a meal in the morning, the median hour for their unique

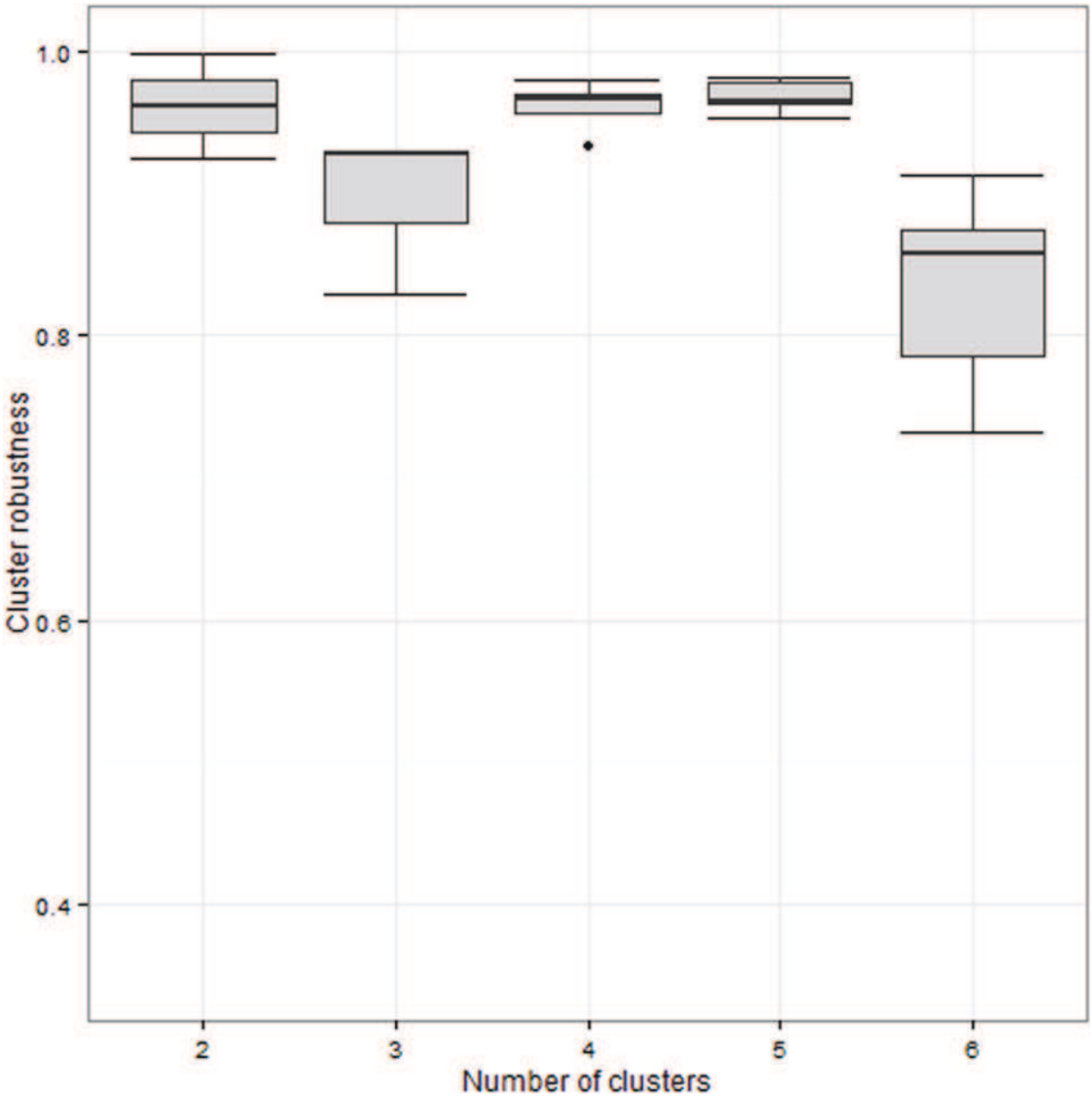


Fig 1. Cluster robustness according to the assumed number of clusters in the dataset (using the PAM algorithm with Manhattan distance). The cluster robustness evaluates the stability of groups while iterating the same clustering method with the same parameters, except for the assumed numbers of clusters in the dataset (from 2 to 6). The black line represents the median value, the bottom and top of the box represent the 1st and 3rd quartiles, and the ends of the whiskers represent minimum and maximum values. Highest median robustness with lowest dispersion was achieved considering 4- and 5-way classification. The examination of the relations between 4- and 5-way classifications revealed that while three groups were constantly retrieved in both cases, the fourth group in the 4-way classification split into two when the 5-way classification was applied (see Table 2).

doi:10.1371/journal.pone.0119161.g001

Table 2. Cross-tabulation of 4- and 5-way classifications using the PAM algorithm with Manhattan distance.

		5-type division					
		1	***2***	***3***	***4***	***5***	***T***
4-type division	*1*	869	0	0	0	27	896
	2	0	670	0	1	0	671
	3	0	0	688	17	0	705
	4	6	2	10	**422**	**282**	722
	T	875	672	698	440	309	2994

This table represents the number of individuals allocated to the different groups when classifying the study population into 4 or 5 groups. While the first 3 groups were constantly retrieved in both classifications, the fourth group in the 4-way classification split into two groups when the 5-way classification was applied.

doi:10.1371/journal.pone.0119161.t002

meal of the day being 7 p.m. (IQR: 4–8 p.m.). Eating one meal per day was not related to food insecurity nor to specific events that occurred during the last week (e.g. fasting, shifted work hours), although these participants more often declared having irregular eating habits (14% vs. 5%, p = 0.007).

The fifth type was named "*2 meals, often outside*" (10% of sample, corresponding to a prevalence of 12%). Most of the participants ate one or two meals per day (8.7% and 86.4%, respectively), with only approximately 5% eating a meal before 10:30 a.m. A very high proportion of meals were eaten away from home, at work or in a restaurant. Moreover, very few were eaten

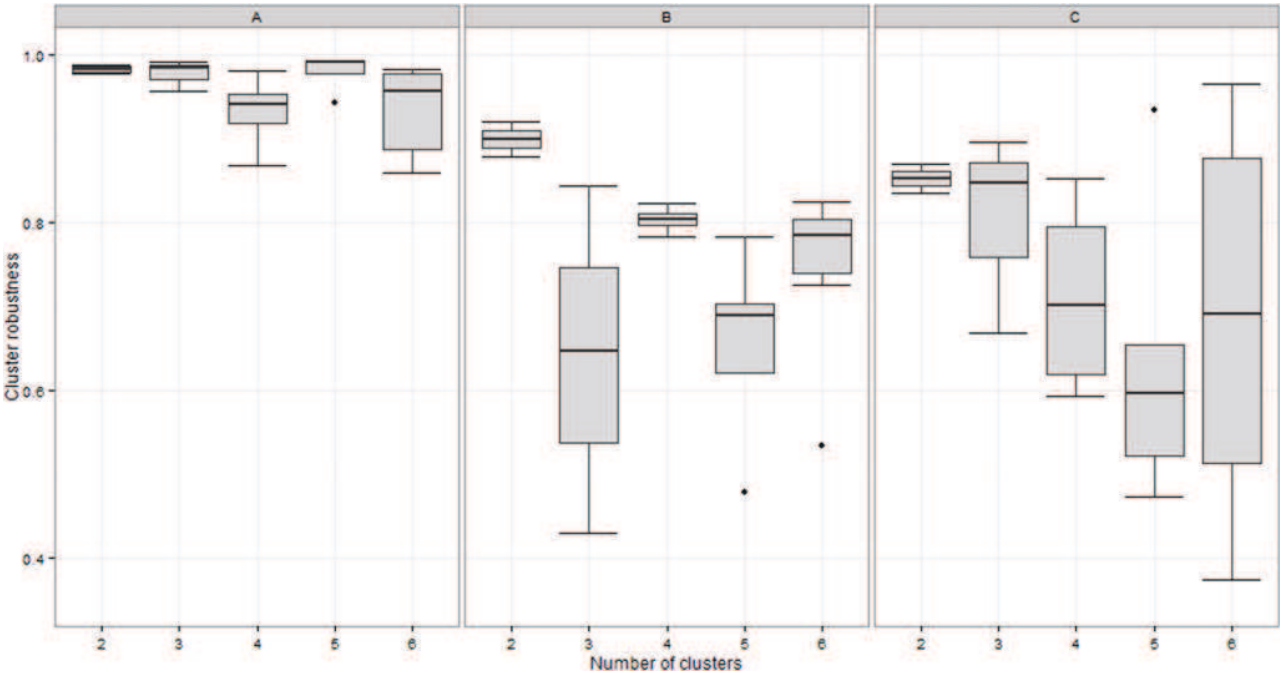


Fig 2. Sensitivity analyses: cluster robustness according to the assumed number of clusters in the dataset using (A) the PAM algorithm with Euclidian distance; (B) the *k*-means algorithm; and (C) a hierarchical clustering algorithm.

doi:10.1371/journal.pone.0119161.g002

alone, as colleagues and friends frequently attended, and the main activity during those meals was chatting.

Socio-demographic and environmental factors associated with meal patterns

Table 3 presents the results of the unadjusted multinomial regression. We chose to consider type 3 ("*3 meals*, *at home with family*") as the reference group, since it was the closest to the traditional French meal pattern.

Compared to the reference group, type 1 ("*3 meals*, *often outside*") included significantly fewer people aged over 60, fewer inactive people (unemployed, retired or staying at home), fewer people with low or intermediate levels of education and fewer people of foreign origin. There was far more diversity in family types (more participants lived alone, in single-parent families or multiple families), but this is most likely explained by the remarkably high prevalence of couples (with or without children) in the reference group (89%).

Belonging to type 2 ("*3 meals*, *mostly alone at home*") was particularly associated with living alone, but also with being female, of an advanced age, having a low or intermediate education level, having a low income and being retired. It was negatively associated with foreign origin. Type 2 was also related with being a student, but this probably reflects the low proportion of students in the reference group.

For participants belonging to type 4 ("*1 or 2 meals*, *mostly at home with television"*), the principal characteristics were being male, of a younger age, having a lower education level, a lower income, fewer children aged under 16 at home and notably living in an underprivileged neighborhood. We also observed that unemployed individuals and people of foreign origin were more likely to belong to this group (although the comparison with type 3 was not significant). Feeling of loneliness was important.

Type 5 ("*2 meals*, *often outside*") was also associated with participants being male, of lower age and a higher sense of loneliness. But unlike type 4, type 5 was connected to higher educational levels.

Finally, the principal characteristics of the reference group ("*3 meals*, *at home with family*") can be deduced from the multinomial regression. This pattern was strongly associated with participants who stayed at home, with a higher income, a nuclear family (couples with or without children) and an almost non-existent sense of loneliness. In this type, students or participants belonging to the 18–29 age group were infrequent.

Food-related factors associated with meal patterns

Table 4 presents the results of a multinomial regression, with adjustment for each previously identified socio-demographic and environmental characteristic. Again, type 3 was chosen as the reference group. Daily snacking was reported significantly less frequently by individuals of type 1, and more frequently by individuals of types 4 and 5, who mostly followed a 2-meal per day pattern. Dissatisfaction concerning food was also more frequent in types 4 and 5. Involvement in meal-related decisions and meal preparation was remarkably frequent in type 2-in which most individuals lived alone. Food quality was considered to be affected by their lifestyle in types 1, 4 and 5, and by financial issues only in type 4. Finally, adherence to the 5-a-day fruit and vegetables guideline was infrequent in all 4 types, reflecting a frequent adherence in the reference group only, but was particularly rare in types 4 and 5 (as was adherence to the 3 dairy products per day guideline in both of these types).

Table 3. Characteristics associated with meal patterns: univariate multinomial logistic regression, with type 3 ("3 meals, at home with family") as reference.

	Type 1: *"3 meals, often outside"* OR [95% CI]	Type 2: *"3 meals, mostly alone at home"* OR [95% CI]	Type 4: *"1 or 2 meals, mostly at home with television"* OR [95% CI]	Type 5: *"2 meals, often outside"* OR [95% CI]	p
Socio-demographic characteristics					
Gender					<0.001
Male	Ref	Ref	Ref	Ref	
Female	1.04 [0.85–1.27]	1.49 [1.19–1.87]	0.75 [0.59–0.96]	0.61 [0.46–0.79]	
Age (years)					<0.001
18–29	Ref	Ref	Ref	Ref	
30–44	0.75 [0.53–1.06]	0.45 [0.29–0.7]	0.62 [0.41–0.93]	0.53 [0.35–0.79]	
45–60	0.77 [0.54–1.09]	0.94 [0.62–1.42]	0.72 [0.48–1.08]	0.49 [0.32–0.75]	
61–74	0.08 [0.05–0.13]	1.01 [0.68–1.51]	0.31 [0.2–0.48]	0.09 [0.05–0.16]	
75+	0.02 [0.01–0.06]	2.17 [1.39–3.4]	0.36 [0.21–0.62]	0.02 [0.01–0.1]	
Educational level					<0.001
College	Ref	Ref	Ref	Ref	
High school	0.59 [0.48–0.74]	1.27 [1–1.59]	1.65 [1.27–2.13]	0.69 [0.52–0.92]	
Primary/none	0.22 [0.14–0.33]	1.46 [1.06–2.01]	1.15 [0.78–1.69]	0.22 [0.12–0.4]	
Occupation					<0.001
Employed	Ref	Ref	Ref	Ref	
Student	1.14 [0.58–2.23]	2.64 [1.26–5.52]	1.55 [0.71–3.4]	1.46 [0.69–3.12]	
Unemployed	0.08 [0.04–0.13]	0.92 [0.6–1.43]	1.21 [0.81–1.8]	0.18 [0.09–0.33]	
Retired	0.03 [0.02–0.04]	1.53 [1.19–1.98]	0.36 [0.27–0.49]	0.06 [0.04–0.09]	
At home/sick	0.01 [0–0.03]	0.45 [0.31–0.65]	0.35 [0.24–0.51]	0.03 [0.01–0.07]	
Income per consumption unit (quartiles)					<0.001
€ 159–1600	Ref	Ref	Ref	Ref	
€ 1601–2500	1.42 [1.03–1.96]	0.51 [0.37–0.69]	0.43 [0.31–0.61]	0.94 [0.63–1.41]	
€ 2501–3900	1.19 [0.86–1.63]	0.33 [0.24–0.45]	0.31 [0.22–0.43]	0.78 [0.52–1.15]	
€ 3901–40,000	1.24 [0.91–1.68]	0.16 [0.12–0.23]	0.16 [0.11–0.23]	0.6 [0.4–0.89]	
Living in an underprivileged neighborhood*					<0.001
No	Ref	Ref	Ref	Ref	
Yes	0.91 [0.71–1.15]	0.94 [0.73–1.21]	1.68 [1.29–2.18]	1.02 [0.75–1.4]	
Parents nationality					<0.001
French, born to two French parents	Ref	Ref	Ref	Ref	
French, born to one foreign parent	0.77 [0.6–0.99]	0.59 [0.45–0.78]	1.27 [0.95–1.7]	0.83 [0.59–1.16]	
Foreigner	0.51 [0.38–0.7]	0.48 [0.34–0.67]	1.3 [0.94–1.79]	0.81 [0.55–1.19]	
Social environment					
Family type					<0.001
Nuclear family	Ref	Ref	Ref	Ref	
One person	52.86 [21.62–129.29]	325.13 [132.42–798.28]	80.27 [32.42–198.72]	53.76 [21.41–135]	
Single parent	3.05 [2.14–4.34]	3.7 [2.39–5.73]	4.68 [3.15–6.95]	3.21 [2.06–5.01]	
Multiple families	1.41 [0.82–2.44]	4.69 [2.69–8.16]	2.88 [1.61–5.14]	2.58 [1.38–4.84]	
Children <16 at home					<0.001
No	Ref	Ref	Ref	Ref	
Yes	0.97 [0.79–1.19]	0.19 [0.15–0.26]	0.67 [0.52–0.86]	0.99 [0.75–1.3]	
Couple					<0.001
Live together	Ref	Ref	Ref	Ref	

(*Continued*)

Table 3. (*Continued*)

	Type 1: *"3 meals, often outside" OR [95% CI]*	Type 2: *"3 meals, mostly alone at home" OR [95% CI]*	Type 4: *"1 or 2 meals, mostly at home with television" OR [95% CI]*	Type 5: *"2 meals, often outside" OR [95% CI]*	*p*
Do not live together	11.36 [5.41–23.82]	20.5 [9.43–44.54]	9.81 [4.35–22.1]	11.22 [4.93–25.57]	
No couple	5.18 [3.95–6.79]	22.84 [17.02–30.65]	8.56 [6.34–11.58]	6.27 [4.51–8.7]	
Solitude					<0.001
Does not feel lonely	Ref	Ref	Ref	Ref	
Feels lonely	1.06 [0.76–1.48]	3.99 [2.95–5.39]	3.15 [2.26–4.39]	2.14 [1.46–3.14]	

**label applied by the French government to target urban renewal programs and specific welfare policies*

doi:10.1371/journal.pone.0119161.t003

Discussion

Using robust clustering methods, we identified five different meal patterns, considering not only frequencies but also timeslots, locations, co-attendants and activities during meals. Meal frequencies have already been described in the SIRS survey [4]. We further explored the meal structures and characteristics in the same representative sample of the general adult, French-speaking population of the Paris metropolitan area. We were able to characterize socio-demographic and environmental factors associated with each pattern, and link them to several food-related behaviors, such as snacking, dissatisfaction concerning food, involvement in meal-related matters, and adherence to general public health policies concerning fruits and vegetables, and dairy products.

Most European countries share the 3-meal pattern with synchronized meal times [1]. Recent research aimed at understanding how daily meal patterns evolve showed that this pattern is still predominant. This is the case in the Nordic countries [14], in Belgium [26], in Italy [27], in Spain and to a lesser extent in the United Kingdom [28]. We showed that the majority (66%) of the adult population of the Paris metropolitan area still eats three meals per day. However, our cluster analysis allowed us to go further and investigate meal patterns in more detail. Our findings indicate that approximately a quarter of the adult population deviate from the traditional French three-meal pattern [29] by omitting breakfast (and sometimes another meal), a behavior that has been associated with low-quality diet and potential consequences for health [30]. Indeed, more than 95% of individuals belonging to types 4 or 5 only ate 1 or 2 meals per day, with a remarkable 18% of individuals of type 4 eating only one meal per day. This result is somewhat close to those from the Nordic survey, where an "unsynchronized" eating pattern was identified and characterized by a late start to the eating day, the displaced timing of eating and a smaller number of eating events. As mentioned by its authors, "most significantly, in all countries the probability of an unsynchronized eating rhythm among the unemployed is approximately double (and even more in Norway) than it is among those in working" [14].

Although our types 4 and 5 presented similar meal rhythms, both having a predominant two-meal pattern, multivariate analysis allowed us to distinguish two different specific meal patterns that would have otherwise been grouped in a single cluster. If types 4 and 5 both concerned rather young and single people, they corresponded to two quite different social profiles. On the one hand, type 4 concerned poorer, less educated, more frequently unemployed individuals who frequently lived in underprivileged neighborhoods and were of foreign origin. Therefore, their meal patterns may be related to different dimensions of social vulnerability. Notably, in the absence of any regular working hours, their day may be less structured and their meals desynchronized or skipped, as shown in some other research on poverty and food

Table 4. Food-related factors associated with meal patterns: univariate multinomial logistic regression with type 3 ("3 meals, at home with family") as reference, with adjustment on socio-demographic and environmental characteristics.

	Type 1:*"3 meals, often outside"* OR [95% CI]	Type 2:*"3 meals, mostly alone at home"* OR [95% CI]	Type 4:*"1 or 2 meals, mostly at home with television"* OR [95% CI]	Type 5:*"2 meals, often outside"* OR [95% CI]	p
Food					
Daily snacking					<0.001
No	Ref	Ref	Ref	Ref	
Yes	0.72 [0.54–0.96]	0.81 [0.59–1.1]	2.51 [1.88–3.35]	2.12 [1.53–2.94]	
Dissatisfaction concerning food					<0.001
No	Ref	Ref	Ref	Ref	
Yes	1.08 [0.72–1.63]	1.25 [0.81–1.91]	2.03 [1.35–3.04]	2.34 [1.5–3.65]	
Involvement in meal-related decisions					0.016
High	Ref	Ref	Ref	Ref	
Low	1.07 [0.75–1.52]	0.54 [0.35–0.82]	0.8 [0.55–1.17]	1.08 [0.71–1.64]	
Food quality affected by lifestyle					<0.001
No	Ref	Ref	Ref	Ref	
Yes	1.48 [1.1–1.99]	1 [0.7–1.42]	1.97 [1.42–2.73]	2.06 [1.46–2.91]	
Food quality affected by financial issues					0.019
No	Ref	Ref	Ref	Ref	
Yes	0.96 [0.68–1.37]	0.92 [0.64–1.33]	1.49 [1.06–2.11]	0.94 [0.62–1.43]	
Participation in food preparation					<0.001
More than once a week	Ref	Ref	Ref	Ref	
Once a week	0.85 [0.5–1.43]	0.45 [0.23–0.88]	0.6 [0.34–1.08]	1.22 [0.68–2.17]	
Once a month	1.42 [0.76–2.64]	0.39 [0.16–0.92]	0.56 [0.27–1.16]	1.25 [0.59–2.64]	
Less or never	1.58 [1.02–2.44]	0.61 [0.37–1.01]	0.77 [0.49–1.22]	1.38 [0.83–2.3]	
Eats 5 fruits and vegetables a day					<0.001
More than once a week	Ref	Ref	Ref	Ref	
Once a week	1.19 [0.82–1.71]	1.35 [0.91–1.99]	1.94 [1.28–2.94]	1.26 [0.79–2.01]	
Once a month	1.39 [0.97–1.98]	1.19 [0.82–1.74]	3.34 [2.31–4.83]	2.26 [1.5–3.43]	
Less or never	1.62 [1.07–2.46]	1.41 [0.92–2.17]	5.07 [3.37–7.64]	2.96 [1.85–4.72]	
Eats 3 dairy products a day					<0.001
More than once a week	Ref	Ref	Ref	Ref	
Once a week	0.82 [0.56–1.22]	1.03 [0.68–1.56]	1.08 [0.71–1.66]	0.9 [0.54–1.5]	
Once a month	1.39 [0.99–1.94]	0.93 [0.65–1.34]	1.54 [1.08–2.2]	2.11 [1.43–3.12]	
Less or never	1.42 [0.95–2.13]	1.37 [0.91–2.06]	2.62 [1.76–3.88]	2.33 [1.46–3.7]	

doi:10.1371/journal.pone.0119161.t004

[31]. Type 5, on the other hand, can be interpreted better as an unsynchronized eating pattern representing a transitional life-phase that will pass when restrictions associated with family life and work start to exert their influence [14]. This interpretation is supported by the characteristics in type 5 subjects, who were young, active people with constraints on their time and a specific urban lifestyle. As for type 1, this emphasizes the strong influence of work and activity on eating patterns.

The existence of types 4 and 5 also raises the question of the subjective meaning of the meal and its highly social associations. Hence, for most people, a solitary meal is a undesirable situation, and may be not declared as a meal at all [32]. Sobal and Nelson have emphasized that "most research about the effects of eating alone on diet and nutrition focuses on the elderly, for whom living and eating alone are prevalent because of the high proportion of widows and widowers". Conversely, our results showed that individuals in type 2-associated with advanced age, lower income, and living alone (maybe as a consequence of widowhood)—strongly adhered to the three-meal schedule (10% even ate 4 meals a day) and took their meals mostly at home, whereas eating alone and meal skipping principally concerned younger people. With regard to the literature about modernization of life styles and their effect on eating habits, the question is whether this phenomenon is transitional or whether it forecasts more sustainable changes in the way new generations eat.

Our results also highlight the issue of eating out, which is not well documented in France [33]. We showed that the three-meal pattern can take different forms according to the socio-demographic characteristics of eaters and the context in which meals are taken. Type 1 and to some extent type 5 offered new results about eating out that indicate that eating out is part of daily eating practices for a large part of the population of the Paris metropolitan area. This behavior can be attributed to work and time constraints, where interactions with friends and colleagues can replace those with family members, especially at lunchtime. However, it above all extends the previous definition of the French model of the three-meal pattern in which meals were typically taken at home. This pattern was formed during the 19th century on the bourgeois model of three meals a day, and gradually spread to society as a whole, becoming a cultural trait so widely shared by all classes that it became normative [29]. If the French still eat mostly at home [34], it is interesting to note that eating out is integrated into the three-meal pattern. In other words, contrary to what the more pessimistic literature would suggest, eating out is not necessarily synonymous of destructuration of eating habits nor of the obesity epidemic [35] but, in contrast, is actually smoothly integrated into the daily eating schedule.

Finally, our results also raise the issue of eating habits of immigrant populations. Paradoxically, individuals of foreign origin typically belonged either to type 4 (characterized by desynchronized eating rhythms and a majority of meals taken in front of the television) or to type 3 (the most consistent with the traditional French model, accounting for 24% of the Paris area population). Few studies have specifically addressed the question of immigrant socialization through food provisioning, cooking, commensality types, tastes and health outcomes in the French context. The available data has focused on food taken at home and has not investigated eating rhythms. However, it has been shown that eating away from home, particularly at school and factory canteens—which shape specific rules for meals (time slots, 3-course meals, dishes, etc.)—are important contexts for eating socialization, with respect to tastes and manners. Eating rhythms and places could thus be equally important as home cooking conditions. This result is even more surprising for immigrants coming from less educated backgrounds, which in France are less likely to follow the three-meal pattern [4]. Further research on this subject is needed to draw more precise conclusions.

This study has some limitations. Clearly, the population of the Paris area is not representative of the whole country in terms of social organization and living conditions. The data are based on declarations from participants that could be affected by a social desirability bias. However, we think that, in the context of the data collection of the SIRS cohort (where each participant was interviewed during at least 1-hour long sessions, and sometimes even more, on a large and varied set of questions about their living conditions, adverse experiences, and very intimate health and biographical events), such a bias might have been reduced since relationships of trust were built between participants and interviewers and food-related questions were

not among the most intimate and sensitive ones addressed. Nonetheless, documentation of meal patterns in a typical week from self-report is clearly limited compared to other, more intensive methods of data collection such as the use of meal diaries over a given period of time. Moreover, only subjective and general perceptions of food quality have been studied. Data on actual nutrient intake would be necessary to further investigate food quality.

Conclusion

Our findings offer new insights into the diversity of meal patterns in the Paris metropolitan area. This study used powerful methods to generate a taxonomy of eating patterns in the largest French metropolitan area. Even if the traditional French model is relatively conserved, we showed that a quarter of the population of the Paris area diverges from it. Social factors such as age, family type, income and occupation are strongly linked to meal patterns and eating behaviors. More specifically, people may divert from the traditional model in two ways that both involve skipping breakfast. In the first group, representing approximately 13% of the population, individuals typically eat one or two meals per day, often with an irregular schedule, at home and in front of the television. Individuals from this group frequently experience unemployment, lower income and living in underserved urban areas. Breakfast skipping, increased snacking, low adherence to dietary guidelines and frequent dissatisfaction about food quality suggest that there might also be health consequences to this meal pattern. In a second group, representing approximately 12% of the population, people also eat two meals or less per day, possibly with the same consequences on food quality. However, in this case, meals are often taken outside of the home, in social settings, and individuals from this group are typically active, integrated, young people, suggesting that this pattern might be an adaptation to a young, modern urban lifestyle. Only longitudinal research would reveal if this 'generational' pattern will change back towards a more traditional one, with ageing, family building or having children.

Author Contributions

Conceived and designed the experiments: JR TL IP PC. Performed the experiments: JR TL IP PC. Analyzed the data: JR. Contributed reagents/materials/analysis tools: TL. Wrote the paper: JR TL IP AL PC.

References

1. Eurostat. How Europeans spend their time—Everyday life of women and men—Data 1998–2002. [Internet]. Luxembourg: Office for Official Publications of the European Communities; 2004. Available: http://ec.europa.eu/eurostat/en/web/products-pocketbooks/-/KS-58-04-998. Accessed 2 February 2015.
2. De Saint Pol T. Le dîner des Français: un synchronisme alimentaire qui se maintient. Econ Stat. 2006; 400: 45–69.
3. Escalon H, Bossard C, Beck F. Baromètre santé nutrition. Saint-Denis: INPES; 2009 p. 424.
4. Lhuissier A, Tichit C, Caillavet F, Cardon P, Masullo A, Martin-Fernandez J, et al. Who still eats three meals a day? Findings from a quantitative survey in the Paris area. Appetite. 2013; 63: 59–69. doi: 10.1016/j.appet.2012.12.012 PMID: 23274963
5. Hébel P. Breakfast in France, a meal on a downhill slide. Consomm Modes Vie. 2013;
6. Poulain JP. The contemporary diet in France: "de-structuration" or from commensalism to "vagabond feeding." Appetite. 2002; 39: 43–55. PMID: 12160564
7. Tavoularis G, Mathé T. The French dietary pattern helps to limit the risk of being fat. Consomm Modes Vie. 2010; 232: 1–4.
8. Fulkerson JA, Larson N, Horning M, Neumark-Sztainer D. A Review of Associations Between Family or Shared Meal Frequency and Dietary and Weight Status Outcomes Across the Lifespan. J Nutr Educ Behav. 2014; 46: 2–19. doi: 10.1016/j.jneb.2013.07.012 PMID: 24054888

9. Sobal J, Hanson K. Family meals and body weight in US adults. Public Health Nutr. 2011; 14: 1555–1562. doi: 10.1017/S1368980011000127 PMID: 21356147
10. Hammons AJ, Fiese BH. Is Frequency of Shared Family Meals Related to the Nutritional Health of Children and Adolescents? Pediatrics. 2011; 127: e1565–e1574. doi: 10.1542/peds.2010-1440 PMID: 21536618
11. Skafida V. The family meal panacea: exploring how different aspects of family meal occurrence, meal habits and meal enjoyment relate to young children's diets. Sociol Health Illn. 2013; 35: 906–923. doi: 10.1111/1467-9566.12007 PMID: 23551143
12. Herpin N. Le repas comme institution: Compte rendu d'une enquête exploratoire. Rev Fr Sociol. 1988; 29: 503–521.
13. Fischler C. L'homnivore: le goût, la cuisine et le corps. Paris: Odile Jacob; 1990.
14. Lund TB, Gronow J. Destructuration or continuity? The daily rhythm of eating in Denmark, Finland, Norway and Sweden in 1997 and 2012. Appetite. 2014; 82: 143–153. doi: 10.1016/j.appet.2014.07.004 PMID: 25017129
15. Fantasia R. Fast food in France. Theory Soc. 1995; 24: 201–243. doi: 10.1007/BF00993397
16. Tovar A, Hennessy E, Must A, Hughes SO, Gute DM, Sliwa S, et al. Feeding styles and evening family meals among recent immigrants. Int J Behav Nutr Phys Act. 2013; 10: 84. doi: 10.1186/1479-5868-10-84 PMID: 23803223
17. Leech RM, McNaughton SA, Timperio A. The clustering of diet, physical activity and sedentary behavior in children and adolescents: a review. Int J Behav Nutr Phys Act. 2014; 11: 4. doi: 10.1186/1479-5868-11-4 PMID: 24450617
18. Martin-Fernandez J, Grillo F, Parizot I, Caillavet F, Chauvin P. Prevalence and socioeconomic and geographical inequalities of household food insecurity in the Paris region, France, 2010. BMC Public Health. 2013; 13: 486. doi: 10.1186/1471-2458-13-486 PMID: 23688296
19. Mäkelä J, Kjaernes U, Pipping Ekström M, L'Orange Fürst E, Gronow J, Holm L. Nordic Meals: Methodological Notes on a Comparative Survey. Appetite. 1999; 32: 73–79. doi: 10.1006/appe.1998.0198 PMID: 9989916
20. Kaufman L, Rousseeuw P. Finding groups in data: an introduction to cluster analysis. 2nd Revised edition. Wiley-Blackwell; 2005. PMID: 17204362
21. Monti S, Tamayo P, Mesirov J, Golub T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Mach Learn. 2003; 52: 91–118. doi: 10.1023/A:1023949509487
22. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2013. Available: http://www.R-project.org/. Accessed 2 February 2015.
23. Simpson T, Armstrong J, Jarman A. Merged consensus clustering to assess and improve class discovery with microarray data. BMC Bioinformatics. 2010; 11: 590. doi: 10.1186/1471-2105-11-590 PMID: 21129181
24. Wickham H. ggplot2: elegant graphics for data analysis [Internet]. Springer; 2009. Available: http://had.co.nz/ggplot2/book. Accessed 2 February 2015.
25. Venables W, Ripley B. Modern Applied Statistics with S [Internet]. Fourth edition. Springer; 2002. Available: http://www.stats.ox.ac.uk/pub/MASS4. Accessed 2 February 2015.
26. Mestdag I, Glorieux I. Change and stability in commensality patterns: a comparative analysis of Belgian time-use data from 1966, 1999 and 2004. Sociol Rev. 2009; 57: 703–726. doi: 10.1111/j.1467-954X.2009.01868.x
27. Monteleone E, Dinella C. Italian meals. Meals in Science and Practice: Interdisciplinary Research and Business Applications. 1 edition. Woodhead Publishing; 2009. p. 704.
28. Díaz-Méndez C, García-Espejo I. Eating practice models in Spain and the United Kingdom: A comparative time-use analysis. Int J Comp Sociol. 2014; 55: 24–44. doi: 10.1177/0020715213519657
29. Grignon C. Rule, fashion, work: The social genesis of the contemporary French pattern of meals. Food Foodways. 1996; 6: 205–241. doi: 10.1080/07409710.1996.9962041
30. Pereira MA, Erickson E, McKee P, Schrankler K, Raatz SK, Lytle LA, et al. Breakfast frequency and quality may affect glycemia and appetite in adults and children. J Nutr. 2011; 141: 163–168. doi: 10.3945/jn.109.114405 PMID: 21123469
31. Caillavet F, Darmon N, Lhuissier A, Regnier F. L'alimentation des populations défavorisées en France. Synthèse des travaux dans les domaines économique, sociologique et nutritionnel. [Internet]. Observatoire National de la Pauvreté et de l'Exclusion Sociale; 2005 2006 p. 44. Available: http://www.opnalim.

org/lalimentation-des-populations-defavorisees-en-france-synthese-des-travaux-dans-les-domaines-economique-sociologique-et-nutritionnel/. Accessed 2 February 2015.

32. Sobal J, Nelson MK. Commensal eating patterns: a community study. Appetite. 2003; 41: 181–190. doi: 10.1016/S0195-6663(03)00078-3 PMID: 14550316

33. Lhuissier A. Anything to declare? Questionnaires and what they tell us: a comparison of "eating out" in national food surveys in France and Britain (1940–2010). Anthropol Food. 2014;S10. Available: http://aof.revues.org/7625. Accessed 2 February 2015.

34. De Saint Pol T, Ricroch L. Le temps de l'alimentation en France. Insee Prem. 2012;1417. Available: http://www.insee.fr/fr/themes/document.asp?ref_id = ip1417. Accessed 2 February 2015.

35. Todd J, Mancino L, Lin B. The Impact of Food Away From Home on Adult Diet Quality [Internet]. U.S. Department of Agriculture, Economic Research Service; 2010. Report No.: ERR-90. Available: http://www.ers.usda.gov/publications/err-economic-research-report/err90.aspx. Accessed 2 February 2015.

Annexe 4. Article soumis dans le cadre de la thèse

Estimation de l'âge chez les adolescents migrants

Elsevier Editorial System(tm) for Forensic Science International
Manuscript Draft

Manuscript Number:

Title: Multivariate methods for the analysis of complex and big data in forensic sciences. Application to age estimation in living persons.

Article Type: Original Research Article

Keywords: Nonlinear dimensionality reduction; clustering; age estimation; multivariate methods; big data

Corresponding Author: Dr. Thomas Lefèvre, M.D., Ph.D.

Corresponding Author's Institution: Université Paris 13

First Author: Thomas Lefèvre, M.D., Ph.D.

Order of Authors: Thomas Lefèvre, M.D., Ph.D.; Patrick Chariot, M.D., Ph.D.; Pierre Chauvin, M.D., Ph.D.

Abstract: Researchers handle increasingly higher dimensional datasets, with many variables to explore. High-dimensional datasets pose several problems, since they are difficult to handle and present unexpected features. As dimensionality increases, classical statistical analysis becomes inoperative. Variables can present redundancy, and the dimensionality of a dataset can be reduced to its lowest possible value. Principal components analysis (PCA) has proven useful to reduce dimensionality but present several shortcomings. Forensic sciences will face the issues specific to an ever growing quantity of data to be integrated. Age estimation in living persons, an unsolved problem so far, could benefit from the integration of various sources of data, e.g. clinical, dental and radiological data.
We present here novel multivariate techniques (nonlinear dimensionality reduction techniques, NLDR), applied to a theoretical example. Results were compared to those of PCA. NLDR techniques were then applied to clinical, dental and radiological data (13 variables) used for age estimation. The correlation dimension of these data was estimated.
NLDR techniques outperformed PCA results. Applying NLDR techniques to data used to estimate age showed that two living persons sharing similar characteristics may present rather different estimated ages. Moreover, data presented a very high informational redundancy, i.e. a correlation dimension of 2. NLDR techniques should be used with or preferred to PCA techniques to analyze complex and big data. Data routinely used for age estimation may not be considered suitable for this purpose. How integrating other data or approaches could improve age estimation in living persons is still uncertain.

Suggested Reviewers: Martina Focardi
Department of Health Sciences-Section of Forensic Medical Sciences, University of Florence, Florence, Italy
federica.del@virgilio.it
Dr Focardi has already published articles on the age estimations in living persons topics and could provide an ethical well-documented point of view on this topic.

Roberto Cameriere
Institute of Legal Medicine, Macerata, Italy

r.cameriere@unimc.it

Richard Bassed
Department of Forensic Medicine, Victorian Institute of Forensic Medicine, Monash University, 57-83 Kavanagh St, Southbank, Melbourne, Australia
richardb@vifm.org
Dr Bassed suggested and studied the interest of a multifactorial approach of the age estimation problem. He also authored a review on the advances in age estimation.

Fabio Corradi
Department of Statistics and Computer Science, University of Florence, Viale Morgagni 59, 50134, Florence, Italy
corradi@ds.uniji.it
Dr Corradi suggested interesting insights of probabilistic approaches for the age estimation in living persons.

Cover Letter

Dear Editor,

Please find attached an original article entitled ‘Multivariate methods for the analysis of complex and big data in forensic sciences. Application to age estimation in living persons’, by Lefèvre, Chariot and Chauvin.

Age estimation is based on the observation of physiological data correlated to chronological age, such as height, fusion of the cartilage or dental eruption. Since none of these data taken separately are accurate enough to predict chronological age, it has been proposed to integrate these various source of data to increase performances.

The classical statistical techniques, e.g. linear regression, used so far to estimate age are of limited performances. Moreover, these techniques will not address issues specific to the increasing amount of data considered.

In this study, we show why the ever growing quantity of data to be integrated is more a problem than a potential solution, and how it is possible to properly deal with complex and big data.

Some recent multivariate techniques can preserve the complex relationships between variables, not only linear correlations. Based on empirical data routinely used in age estimation, our analyses show that i) the integration of various data may not be an efficient strategy since data are very redundant in terms of information, and that ii) these data fail to correctly sort out people in terms of estimated ages i.e. people similar or identical in terms of clinical, dental and radiological data can present very different estimated ages.

The techniques presented here will be of a wider use in the near future. We wish they can be challenged on other age estimation datasets as well as in other forensic fields.

Sincerely.

Thomas Lefèvre, MD PhD

Patrick Chariot, MD PhD

Pierre Chauvin, MD PhD

Multivariate methods for the analysis of complex and big data in forensic sciences. Application to age estimation in living persons.

Thomas Lefèvre[1,2,3,4], Patrick Chariot[3,4], Pierre Chauvin[1,2]

[1]Inserm, UMRS 1136, Pierre Louis Institute of Epidemiology and Public Health, Department of Social Epidemiology, Paris, France

[2]Université Pierre et Marie Curie-Paris 6, UMRS 1136, Paris, France

[3] AP-HP, Hôpital Jean-Verdier, Department of Forensic Medicine, F-93140 Bondy, France

[4] IRIS - Institut de recherches interdisciplinaires sur les enjeux sociaux (INSERM, CNRS, EHESS, Université Paris 13, UMR 8156-723), Bobigny, France

Phone: +33 (0)144738460

Fax: +33 (0)144738663

Institutional address: Inserm, UMRS 1136, 27 rue Chaligny, 75571. Paris Cedex 12, France

E-mail addresses: thomas.lefevre@inserm.fr, patrick.chariot@jvr.aphp.fr, pierre.chauvin@inserm.fr

Corresponding author: Thomas Lefèvre

Highlights

- To date, techniques used for age estimation in living persons are not accurate enough for forensic and scientific purposes.
- The integration of diverse sources of information, e.g. clinical, dental and radiological information, to improve accuracy of age estimation may not be an adequate solution.
- Big data do not need to be so “big” to present specific issues regarding their dimensionality and complexity, that must be dealt with.
- When heterogeneous or big data are considered, recent techniques known as non-linear dimensionality reduction techniques may be preferred to classical techniques such as principal analysis components.

ABSTRACT

Researchers handle increasingly higher dimensional datasets, with many variables to explore. High-dimensional datasets pose several problems, since they are difficult to handle and present unexpected features. As dimensionality increases, classical statistical analysis becomes inoperative. Variables can present redundancy, and the dimensionality of a dataset can be reduced to its lowest possible value. Principal components analysis (PCA) has proven useful to reduce dimensionality but present several shortcomings. Forensic sciences will face the issues specific to an ever growing quantity of data to be integrated. Age estimation in living persons, an unsolved problem so far, could benefit from the integration of various sources of data, e.g. clinical, dental and radiological data.

We present here novel multivariate techniques (nonlinear dimensionality reduction techniques, NLDR), applied to a theoretical example. Results were compared to those of PCA. NLDR techniques were then applied to clinical, dental and radiological data (13 variables) used for age estimation. The correlation dimension of these data was estimated.

NLDR techniques outperformed PCA results. Applying NLDR techniques to data used to estimate age showed that two living persons sharing similar characteristics may present rather different estimated ages. Moreover, data presented a very high informational redundancy, i.e. a correlation dimension of 2.

NLDR techniques should be used with or preferred to PCA techniques to analyze complex and big data. Data routinely used for age estimation may not be considered suitable for this purpose. How integrating other data or approaches could improve age estimation in living persons is still uncertain.

Keywords: Nonlinear dimensionality reduction; clustering; age estimation; multivariate methods; big data

1. Introduction

Scientific research and forensic science are – at least partly – about finding associations between factors and searching for underlying causal relationships between so-called exposure and events of interest. Whether accounting for multiple covariates to control for possible biases or not, classical hypothesis testing and linear regression are extensively used and have proved to be relevant. However, they may not be sufficient to address all issues and analytical needs in forensic sciences [1,2]. When one wants to widen the scope of experimentation in forensic sciences, a key question can be raised: how close is one subgroup of people to another or how different are they? Clustering techniques are used in different fields, such as computer vision and imaging [3], genetics [4] and public health [5]. The international ENCODE project made extensive use of clustering techniques to systematically search for the functionality of “junk” DNA [6]. These methods provide clues for identifying homogeneous groups of people, but they share a common limitation: all of them operate on a “flat” feature space. In the real world, the neighborhood or proximity between two persons may not respect such a geometric assumption and inappropriate shortcuts may appear, falsely linking two people who should not be related to each other. Moreover, researchers increasingly have to handle large datasets, with dozens of potential outcomes and as many explanatory variables of interest. Classical clustering methods and classical statistical tools lose their ability to separate two distinct groups as dimensionality increases, as well as their ability to reach statistical significance. This is known as the “curse of dimensionality” or the “empty phenomenon” [7,8]. According to Lee and Verleysen [8], issues with heterogeneous data can appear as soon as we deal with 10 to 20 or more variables. Data should then be considered as “big data” in many cases. The methods discussed in this paper address two related issues: respecting the intrinsic geometry of data and reducing their dimensionality to make them more comprehensive and even graphically displayable. These methods are called “nonlinear dimensionality reduction” (NLDR) techniques and

appeared at the beginning of the 2000s [9,10]. Since then, variants have been proposed [10,11]. In this article, we first presented the importance of preserving the intrinsic geometry of data. Second, we provided a brief description of a typical NLDR technique and compared its performance with principal components analysis (PCA) and multidimensional scaling (MDS) performance on a theoretical example. Third, we applied NLDR techniques to age estimation in living persons. Estimating the age of a person based on various clinical or non clinical data is an old challenging problem in forensic sciences that is standing still, regardless how often it has been explored [12-16]. Even if the way forensic physicians deal with this topic can be controversial, most experts agree that combining any potential informative data is the best mean to reach accuracy, e.g., combining dental and other radiological data [17-19]. The demonstration of a significant linear trend between different characteristics of a living person and its chronological age is not dubious, but is also poorly contributory to an accurate estimate of the person's age. We actually have no precise idea of how informative are the data usually handled to estimate the age of a living person, and how relevant they are to discriminate a person from another one in terms of age, not to mention to determine if a person reached the legal majority or not. To date, a single study used PCA to estimate the age-at-death [20] and none to estimate the age in living persons. Here, we applied NLDR techniques to empirical forensic data, integrating clinical, dental and radiological data and investigated whether these data could properly and accurately ground age estimation in living people.

2. Methods

2.1 Preserving data geometry and complexity

The difference between a flat space and a more generic, curved space can be likened to the difference between considering the earth to be flat and considering it to be spherical. There is no universal projection method for creating a flat map of the earth with virtually no

geometrical distortion, either angular or metric [21]. In the same way, in the absence of information about how datasets are “physically” structured, it may be inaccurate to assume that they are flat, with no curvature at all. Figure 1 shows how geometry determines whether two data points are neighbors or not. It also emphasizes why the dimensionality of a dataset can be reduced.

2.2 Nonlinear dimensionality reduction techniques versus classical techniques

PCA methods can be used to reduce a dataset dimensionality [10,22] by rearranging the feature space by combining variables into factors. These factors are obtained so that they are not correlated with one another. While PCA-like techniques are effective in many cases, their main limitation lies in their linear assumption. The linearity hypothesis assumes that the entire problem to be addressed can be broken down into elementary sub-problems to which the correct weightings can be added to reconstitute the entire initial problem. PCA-like techniques should not provide fundamentally wrong results, although they may destroy evidence for truth or distort more subtle relationships. Therefore, there is room for more suitable techniques than PCA-like techniques. In contrast to the linear assumption in PCA, these techniques are called “nonlinear dimensionality reduction” (NLDR) techniques. They only consider geometrical proximity, apart from statistical considerations. The objective of NLDR techniques is to build the most “respectful” space in terms of “true” neighborhood. For this, these techniques depend on the construction of geodesic paths. NLDR techniques can preserve the nonlinear associations between variables. Data dimensionality is reduced, while the potential complexity of the associations between the variables is qualitatively preserved. The first NLDR algorithms appeared in the early 2000s, the archetype being ISOMAP [9,10]. Other methods have been proposed, based on the construction of a graph depicting the neighbor relationships for each data point. Given an intrinsic dimension and a dataset as inputs, the ISOMAP algorithm operates in three steps, as given in Table 1. Another class of NLDR techniques preserves the topological complexity and properties of datasets (i.e. their

neighboring relationships: two individuals close to each other in the initial dataset remain close to each other in the reduced dataset), based on neural networks, such as the autoencoders [23-25].

2.3 Limitations and comparisons of three algorithms: an empirical approach

We ran the ISOMAP technique on a theoretical example. We also included PCA and MDS techniques for comparison. Results are provided by a MATLAB program called MANI (see supplementary material). The present theoretical example is usually called the Swiss roll dataset [9]. The Swiss roll is a plane that is rolled up with none of its surfaces touching any other. Its intrinsic dimension is 2 (data points all belong to a 2-dimensional plane). Using dimensionality reduction algorithms, we plan to unfold them to obtain a rectangular plane for the Swiss roll. The execution times for the different algorithms are indicated into brackets. Performance of other NLDR techniques both applied on the Swiss roll dataset as well as applied on two other theoretical datasets can be found in additional data.

2.4 Parameters tuning

Most NLDR techniques require parameter setting. Apart from the estimated intrinsic dimension, there is usually only one tuning parameter to be inputted. ISOMAP requires defining the neighborhood, that is, how many neighbors should be searched for around a data point. If the data are relatively sparse, specifying a high number of neighbors may lead to register some that are actually far from that point. In other words, it can lead the algorithm leaping undesirably from one surface to the next if they are close to each other and if there are not enough neighbor data points.

2.5 Dimensionality reduction with intrinsic geometry preservation

Before reducing the initial dimensionality of a dataset to its intrinsic one, one needs to estimate the intrinsic dimension, without any prior information. The intrinsic dimension of a dataset can be defined as the minimum number of independent variables needed to describe it without information loss [26]. The approaches to retrieving this number from a dataset are based on different conceptions of dimensionality. Camastra proposed the following taxonomy for these different methods: local, global and fractal-based [26]. Several techniques exist for estimating the intrinsic dimension of a dataset. Reviews of these techniques can be found in [26,27]. We give here the example of the correlation dimension, which is also used in system dynamics [28]. It is based on the correlation integral $C_m(r)$, which is defined as:

$$C_m(r) = \lim_{N\to\infty} \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} I(|Xj - Xi| =< r) \qquad \text{(Eq. 1)}$$

where X_i are data points, N the number of data points in the sample, r an arbitrary radius, and I the indicator function (i.e., I(condition) = 1 if the condition is true, 0 if it is not).

The correlation dimension D is then defined as

$$D = \lim_{r\to 0} \frac{\ln(C_m(r))}{\ln(r)} \qquad \text{(Eq. 2)}$$

Techniques other than the correlation dimension exist, such as the nearest-neighbor estimator [26] or the maximum-likelihood estimator [29]. The results of these types of dimensionality estimators applied on four examples of theoretical datasets are reported in Table 2. For each example, we give the intrinsic dimensionality, which is the dimensionality to be retrieved.

2.6 About outliers

Some NLDR techniques are not data-conservative. If an initial dataset consists of 3,000 observations, the processed dataset may contain only 2,500 of them (ISOMAP behaves like this). The reason lies in the graph construction, where distances are estimated and *k* neighbors are searched for. Some points may appear to be not connected to any others and are assigned an “infinite” distance to the other points. In such a case, they are isolated and eliminated from the dataset since they are seen as outliers. If one wishes to keep all data points, more conservative algorithms should be used, e.g., LTSA or autoencoders [10,24,25].

2.7 Application to age estimation

2.7.1 Available data

We applied two NDLR techniques on a previously published dataset [16]. This dataset included clinical, dental and radiological data. Clinical data included geographical origin (country) and sex. Dental data were the eruption of the second and third molars (yes/no for each molar, i.e. 8 distinct variables). Radiological data included the readout of the forensic or radiologist expert regarding the fusion of the distal radius and ulna epiphyses. Additional data were: the alleged age of the person, the estimated age as provided by the Greulich and Pyle atlas, and the age estimated by the forensic examiner, on the basis of all the elements mentioned above. The alleged age was the age provided by the person examined. The radiological age was the age estimated based on the fusion of the distal radius and ulna epiphyses, according to the Greulich and Pyle atlas.

Countries were aggregated into 8 levels of one geographical origin variable (Africa, Asia, Western Europe, Eastern Europe, Middle East, Oceania, North America, South and Central America). Data were obtained over one year, from 1 January to 31 December 2007 in a suburban area near Paris, France. The age assessments were requested by the public

prosecutor's office of Bobigny (Seine-Saint-Denis) for purposes of criminal or asylum proceedings. Examinations were conducted by trained forensic physicians.

2.7.2 Descriptive analyses

Median ages with 10^{th} and 90^{th} percentiles were computed for each case whether 2^{nd} or 3^{rd} molars were present or not; and whether gender was male or female. Differences in medians were assessed with Kruskal-Wallis test. Correlations between the age estimated by the forensic examiner, the skeletal age and the alleged age were also estimated.

2.7.3 Mapping techniques

We aimed at mapping every person for whom clinical and radiological data were complete in a low-dimensional space so that the proximity or similarity of each pair of persons would be preserved. In such a mapping, if clinical and radiological data were representative of the age of a person, then two persons close to each other should have similar or close ages. Since data other than ages were categorical, we first applied a conservative transformation using multiple correspondence analysis (MCA), so that we obtained continuous variables. MCA is equivalent to PCA for discrete variables.

The second step consisted in estimating the intrinsic dimensionality of the dataset. We used the correlation dimension estimator. The third step consisted in applying two different NLDR techniques on these data, i.e. a conservative one, namely an autoencoder which is a kind of neural network [23-25], and a non-conservative one, the ISOMAP algorithm. The final step was the labeling of each person in the mapping provided by the NLDR techniques with their associated ages, i.e. respectively the alleged age, the estimated age according to the Greulich and Pyle atlas and the age estimated by the forensic physician, taken as a clinical, dental and radiological synthesis. So that figures could be readable, a random sample of 40 individuals out of the total number of available individuals was drawn and used for graphical

display. MCA was performed with the statistical R software, and the ISOMAP and autoencoder algorithms run under MATLAB R2009b with the DR toolbox [30].

3. Results

3.1 A theoretical example

Results are reported in Figure 2. PCA was unable to unfold the Swiss roll dataset and behaved like a projection of data on a plane in a certain direction. ISOMAP behaved considerably better and was able to correctly unfold the dataset. In terms of computer time for execution, ISOMAP was time-consuming, since one of its steps consisted in applying a classical MDS algorithm, which was slow. Comparisons with other NLDR techniques and on other datasets (see supplementary material, Figures 1-3) all confirmed that NLDR methods perform better than PCA or MDS.

3.2 Age estimation data

The initial dataset contained 499 people. Data were complete for 233 of them (46.7%). 74.7% were males, and the geographical origins were represented as follows: 96 from Africa (41%), 70 from Middle East (30%), 40 from Asia (17%), 18 from Western Europe (8%), 5 from South America, 3 from Western Europe and 1 from Oceania. The mean alleged age was 16.4 years (standard deviation SD: 2.1, median m [min-max]: 16.5 [9-36]). The mean estimated age according to the Greulich and Pyle atlas (resp. mean estimated age) was 17.8 (SD 1.58, m 18 [9-19.5]) (resp. 17.9 SD 2.08, m 18.5 [9-36]).

Table 3 presents the descriptive results. Individually, almost all variables were associated with significant differences in terms of median age as estimated by the forensic examiner. It was notably the case for the presence of 3rd molars. Estimated ages were strongly correlated to skeletal ages (Spearman's coefficient: 0.92, $p<0.001$) and alleged ages (0.73, $p<0.001$).

The application of MCA on the data resulted in two principal axes accounting for 15.9% and 11.2%, respectively, of the total variance of the data. The ages according to the Greulich and Pyle atlas, the alleged ages and the ages estimated by the forensic examiner were distributed along the principal axes seemingly at random. Ages under and above 18 years were intertwined with each other (data not shown). Otherwise stated, no straight or simple line could separate people younger than 18 from people older than 18 or give a clear linear and ordered trend for ages across the plane.

According to the correlation dimension estimator, the intrinsic dimensionality of the dataset was 2. The results of the application of the ISOMAP and the autoencoder algorithms are displayed in Figures 3 and 4. The non conservative ISOMAP algorithm took into account 101 people out of 233, i.e. 132 were considered not connected to the 101 considered.

Starting from 13 variables routinely used to estimate age in clinical forensic settings, we ended up with an intrinsic data dimensionality of 2. A closer examination of the data obtained either with the ISOMAP or the autoencoder algorithms showed that in fact only 1 dimension could suffice to characterize data. Indeed, some areas of the 2 dimensional plane showed a perfect line, i.e. only one dimension could be used to describe data perfectly.

In the 2-dimensional case, i.e. outside the particular subgroup of people that fitted a line, (figure 4), we found a high dispersion of the individuals across the plane, and no specific portion of this plane seemed to gather people according to a homogeneous age profile, e.g., a region of the plane gathering more specifically individuals aged between 14 and 17 years. The alleged ages, the radiological ages and the ages estimated as a synthesis of the forensic examination were seemingly distributed in a random manner across the plane.

4. Discussion

We have presented the fundamental issues raised by the increase in dataset dimensionality, as well as the need to preserve the intrinsic data geometry as much as possible, in order to

avoid mistakes in analysis. On one hand, it seems necessary to collect more and more data, but on the other, doing so will entail some unexpected pitfalls if we do not change our habits regarding conventional analysis. One way to cope with these issues might be to use NLDR techniques. Although they are still recent techniques, they have acquired some maturity and diversity, which allow making comparisons, especially with linear dimensionality reduction techniques. We can assume that they will be increasingly used in many fields, as has been shown with a few available examples [11,31,32] and forensic sciences should be no exception. Considering more data sources to estimate age will lead to bigger datasets that should be handled and analyzed carefully. Age estimation in living persons can be an opportunity to introduce more appropriate techniques like NLDR techniques.

These methods have limitations. First, like many other techniques, the user has to specify some parameters, which can be a problem if there is no prior information on the intrinsic structure of data. However, only two parameters usually need to be specified, one of which is the estimated intrinsic dimension. Second, there is presently no consensus as to what the best intrinsic dimension estimator is for each situation and objective when these estimators can be used. The correlation dimension estimator is widely used and usually behaves well. Finally, there is still a lack of comparisons between NLDR techniques and classical techniques on real data. However, NLDR manifested its superiority on synthetic datasets [33,34] as well as in the first experiments on real data [31-33].

In our forensic age estimation study, clinical, dental and radiological data that are routinely used to estimate one person's age failed to sort out people so that their similarity according to these characteristics translates into similar estimated ages. It appeared that a problem described by 13 distinct variables was collapsible to a one- or two-dimensional problem, although the classical approach, i.e. with MCA, suggested that two dimensions only accounted for 27.1% of the total variance. This suggests a high redundancy – or equivalently a poor informational value – of the data. Moreover, the locally one-dimensional representation of data implies the possibility of a perfect linear regression. Unfortunately,

figure 3 demonstrated that if this part of data could be depicted by a line, it failed to sort out in a correct order the different ages. This combination of clinical, dental and radiological data could not explain the ages that we registered. Worse, their addition or integration seemed to be useless. It therefore questions the relevancy of searching for more variables to integrate and compare. This new insight to the age estimation problem highlights the fact that the existence of a linear correlation between some characteristics such as radiological features and the chronological age is one kind of evidence, but of limited value for the accurate estimation of one particular person's age. Explaining and predicting facts are distinct, rarely compatible tasks [35]. Our findings do not contradict that such a correlation exists. They merely illustrate that this correlation is neither an efficient way to estimate the age of a person, nor to decide whether a person is less or more than 18 years old.

The goals we aim at – estimating the age, classifying persons with respect to being above 18 years – must not be confounded with our will to understand as accurately as possible the physiology of ageing. Even high standards in molecular analysis have failed short to predict a physiological age more accurately than a characteristic time of one year [36]. If integrating data of different natures to estimate age is laudable, the linear regression techniques used so far [19] are inappropriate since they are above all explanatory techniques (i.e. they distribute the overall variance of data among the variables of interest according to their respective contribution to this variance). They are in no way predictive methods although they are abusively called this. They are useful in a risk factor approach, to identify and quantify independent factor risks. Such modeling approaches cannot be satisfactory in age estimation unless they are very accurate. In the case of deciding whether a person is above 18 years old, we should focus on the best available techniques that present the best performances in terms of classification, and try them on available data. Highly effective techniques exist that should be challenged in the field of forensics, such as the Support Vector Machine (SVM) algorithms [24]. However, such techniques require the previous knowledge of the real chronological age of a subsample of the people of whom we want to

estimate the age. The performances of SVM algorithms far exceed the results that can be expected from linear techniques, such as regression techniques previously presented in age estimation [9] and have been acknowledged in various settings [37,38]. So far, they are gaining a wider acceptance in clinical fields, particularly for improving diagnostic tests [38].

In this suggested way to improve performances in age estimation, not knowing the real age of the examined individuals is a limitation to our work. Therefore, we strongly encourage researchers to duplicate our findings on their own data. Similarly, we had no basic characteristics such as weight and height, nor more sophisticated imaging methods, such as clavicle CT scans or Magnetic Resonance Imaging [39,40]. There is nonetheless another way to cope with age estimation in the case where the real age remains unknown, which can be provided by Bayesian approaches [41,42]. Recently, they also proved promising for classifying individuals based on dental evidence and relying on soft evidence [43,44]. Moreover, Bayesian approaches could take into account the age alleged by the person, which would provide a more ethical approach to this problem. However, whether SVM or Bayesian approaches are chosen, both cases require that researchers gain confidence into these now well-known techniques, can handle and criticize them, and if they prove efficient, use them in their daily practice.

5. Conclusion

The integration of various sources of information to improve accuracy in estimating the age of living persons may be considered cautiously and in accordance to the goal we aim to: estimating the age or classifying persons according a threshold age. The increase of data amounts present specific issues that a forensic scientist should be aware of and that must be dealt with adequate techniques.

Competing interests: We declare that we have no conflicts of interest

Funding: None

List of Abbreviations:

HLLE: Hessian LLE

ISOMAP: Isometric mapping

LLE: Locally Linear Embedding

LTSA: Local Tangent Space Alignment

MCA: Multiple correspondences analysis

MDS: Multidimensional scaling

NLDR: Nonlinear dimensionality reduction

PCA: Principal components analysis.

SVM: Support Vector Machine

References

1 Galea S, Riddle M, Kaplan GA. (2009) Causal thinking and complex system approaches in epidemiology. Int J Epidemiol 39:97–106

2 Reshef DN, Reshef YA, Finucane HK, et al. (2011) Detecting Novel Associations in Large Data Sets. Science 334:1518–1524

3 Connolly AC, Guntupalli JS, Gors J, et al. (2012) The representation of biological classes in the human brain. J Neurosci 32:2608–2618

4 Priyadarshini G, Sarmah R, Chakraborty B, et al. (2012) An effective graph-based clustering technique to identify coherent patterns from gene expression data. Int J Bioinform Res Appl 8:18–37

5 Conry MC, Morgan K, Curry P, et al. (2011) The clustering of health behaviours in Ireland and their relationship with mental health, self-rated health and quality of life. BMC Public Health 11:692

6 ENCODE Project Consortium, et al. (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74

7 Bellman R. (2003) Dynamic programming. Dover Publications, Mineola

8 Lee JA, Verleysen M. (2007) Nonlinear dimensionality reduction. Springer, Berlin

9 Tenenbaum JB, de Silva V, Langford JC. (2000) A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science 290:2319–2323

10 Izenman AJ. (2008) Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning, illustrated edition. Springer-Verlag, New York

11 Gorban AN, Zinovyev A. (2010) Principal manifolds and graphs in practice: from molecular biology to dynamical systems. International Journal of Neural Systems 20:219–232

12 Olze A, Reisinger W, Geserick G, Schmeling A. (2006) Age estimation of unaccompanied minors. Part II. Dental aspects. Forensic Sci Int 159 Suppl 1:S65–67

13 Baumann U, Schulz R, Reisinger W, Heinecke A, Schmeling A, Schmidt S. (2009) Reference study on the time frame for ossification of the distal radius and ulnar epiphyses on the hand radiograph. Forensic Sci Int 191:15–18

14 Mansourvar M, Ismail MA, Raj RG, et al. (2014) The applicability of Greulich and Pyle atlas to assess skeletal age for four ethnic groups. J Forensic Leg Med 22:26–29

15 Bassed RB. (2012) Advances in forensic age estimation. Forensic Science, Medicine, and Pathology 8:194–196

16 Pruvost MO, Boraud C, Chariot P. (2010) Skeletal age determination in adolescents involved in judicial procedures: from evidence-based principles to medical practice. J Med Ethics 36:71–74

17 Schulz R, Schiborr M, Pfeiffer H, Schmidt S, Schmeling A. (2014) Forensic age estimation in living subjects based on ultrasound examination of the ossification of the olecranon. J Forensic Leg Med 22:68–72

18 Manzoor Mughal A, Hassan N, Ahmed A. (2014) Bone Age Assessment Methods: A Critical Review. Pak J Med Sci 30:211–215

19 Bassed RB, Briggs C, Drummer OH. (2011) Age estimation using CT imaging of the third molar tooth, the medial clavicular epiphysis, and the spheno-occipital synchondrosis: a multifactorial approach. Forensic Sci Int 212:273.e1–5

20 Lovejoy CO, Meindl RS, Mensforth RP, and Barton TJ. (1985) Multifactorial determination of skeletal age at death: A method and blind tests of its accuracy. Am J Phys Anthropol 68:1-14

21 Robinson A. (1988) American Cartographic Association. Committee on Map Projections. Choosing a world map: attributes, distortions, classes, aspects. Falls Church

22 Pearson K. (1901) On lines and planes of closest fit to systems of points in space. Philosophical Magazine 2:559–572

23 Hinton GE, Salakhutdinov RR. (2006) Reducing the Dimensionality of Data with Neural Networks. Science 313:504–507

24 Bengio Y, Courville A, Vincent P. (2013) Representation Learning: A Review and New Perspectives. IEEE Trans Pattern Anal Mach Intell 35:1798–1828

25 Bengio Y, Yao L, Alain G, Vincent P. (2013) Generalized Denoising Auto-Encoders as Generative Models. arXiv:13056663 [cs] 2013. (accessed 12 Jan 2015)

26 Camastra F. (2003) Data Dimensionality Estimation Methods: A Survey. Pattern Recognition 36:2945–2954

27 Carter KM, Member S, Raich R, Iii AOH. (2010) On Local Intrinsic Dimension Estimation and Its Applications. IEEE Trans Signal Process 58:650–663

28 Grassberger P, Procaccia I. (1983) Measuring the strangeness of strange attractors. Physica D: Nonlinear Phenomena 9:189–208

29 Levina E, Bickel PJ. (2005) Maximum Likelihood Estimation of Intrinsic Dimension. Advances in Neural Information Processing Systems 17:777–784

30 Matlab Toolbox for Dimensionality Reduction, v0.8.1, 2013, http://homepage.tudelft.nl/19j49/Matlab_Toolbox_for_Dimensionality_Reduction.html. (accessed 12 Jan 2015)

31 Shi J, Luo Z. (2010) Nonlinear dimensionality reduction of gene expression data for visualization and clustering analysis of cancer tissue samples. Computers in Biology and Medicine 40:723–732

32 Weng S, Zhang C, Lin Z, Zhang X. (2005) Mining the structural knowledge of high-dimensional medical data using isomap. Med Biol Eng Comput 43:410–412

33 Gorban AN. (2008) Principal manifolds for data visualization and dimension reduction. UK Springer London, London

34 Van der Maaten LJP, Postma EO, van den Herik HJ. (2009) Dimensionality Reduction: A Comparative Review. Tilburg, Tilburg University Technical Report

35 Thom R. (1999) Prédire n'est pas expliquer. Flammarion, Paris

36 Gibbs WW. (2014) Biomarkers and ageing: The clock-watcher. Nature 508:168–170

37 Howe A, Escalona OJ, Di Maio R, et al. (2014) A support vector machine for predicting defibrillation outcomes from waveform metrics. Resuscitation 85:343-349

38 Garcia Molina JF, Zheng L, Sertdemir, et al. (2014) Incremental learning with SVM for multimodal classification of prostatic adenocarcinoma. PLoS One 9:e93600

39 Krämer JA, Schmidt S, Jürgens Ku, et al. (2014) Forensic age estimation in living individuals using 3.0 T MRI of the distal femur. Int J Legal Med 128:509-514

40 Wittschieber D, Schulz R, Vieth V, et al. (2014) Influence of the examiner's qualification and sources of error during stage determination of the medial clavicular epiphysis by means of computed tomography. Int J Legal Med 128:183–191

41 Pe'er D. (2005) Bayesian network analysis of signaling networks: a primer. Sci STKE, 26;2005(281):pl4

42 Chariot P, Caussinus H. (2014) Age estimation in undocumented migrant adolescents: Medical response to judicial authorities. Presse Med. doi:10.1016/j.lpm.2014.07.017

43 Corradi F, Pinchi V, Barsanti I, et al. (2013) Optimal age classification of young individuals based on dental evidence in civil and criminal proceedings. Int J Legal Med 127:1157-1164

44 Corradi F, Pinchi V, Barsanti I, Garatti S. (2013) Probabilistic classification of age by third molar development: the use of soft evidence. J Forensic Sci. 58:51-59

Table 1: The three steps of the ISOMAP algorithm

Table 2: Intrinsic dimensionality estimators applied to four theoretical examples.

GMST: geodesic minimum spanning tree. Correlation dimension and maximum likelihood estimators succeed in approaching the intrinsic dimension in most, if not all, the examples. Estimators are described in [25,27,28,S2-S4]

A dimension is not necessarily an integer, it can be fractal [S5,S6]. Lines drawn on a sheet of paper can present qualitative differences in terms of their aspects, and yet they share the common property of being one-dimensional objects. Whether one has to locate a specific point on a straight line or on a line consisting of many zigzags, only one unique coordinate is necessary because these two lines are both one-dimensional. The straight line will not fill space significantly, while the zigzag line will occupy more space. It is then possible to define a fractal dimension to characterize these two lines, which will be between 1 and 2, depending on their ability to fill 2-dimensional space.

Table 3: Median ages estimated by the forensic examiner, for each variable (gender, 2^{nd} and 3^{rd} molars).

*: $p<0.01$, **: $p<0.001$. Median age estimated by the forensic examiner is given for each case whether a 2^{nd} or 3^{rd} molar is present or absent, with the 10^{th} and 90^{th} percentiles into brackets and whether the person is male or female.

Legends of figures

Figure 1: Distances over two different spaces: flat space and curved space.

Changes in proximity relationships for a, b and c. Distances that appear to be correct and relevant in the case of a nonstructured space (case A) will turn out to be incorrect if the space is structured and is described by a lower intrinsic dimension (case B): cars cannot go through buildings; they have to use roads.

In case A, the closest person to “b” is “c” because the ambient space is flat (i.e., 2-dimensional), with no specific structure. In case B, the closest person to “b” is “a” because data are intrinsically structured as a spiral. Considering that “b” and “c” belong to the same group on the basis of their proximity leads to a wrong statement if it is measured the same way in both cases. Proximity in case B should be measured in the way in which the dot style distance approximates it. By definition, a distance that measures the shortest path between two points belonging to a specific curved space is called a geodesic distance. In case A, we need a 2-dimensional space to describe data. Two coordinates are needed to locate someone. In case B, when we respect the intrinsic spiral geometry, we only need one degree of freedom: it is merely a line with curvature, and only one coordinate along that line is enough to locate someone.

Figure 2: A theoretical example (Swiss roll) and its expected unfolding, in which a NLDR technique (ISOMAP) is compared with PCA and MDS.

Top: Example of the Swiss roll (left) and its expected unfolding (right)

Bottom: PCA (left), MDS (middle) and ISOMAP (right) unfoldings. Execution times are given in seconds or minutes.

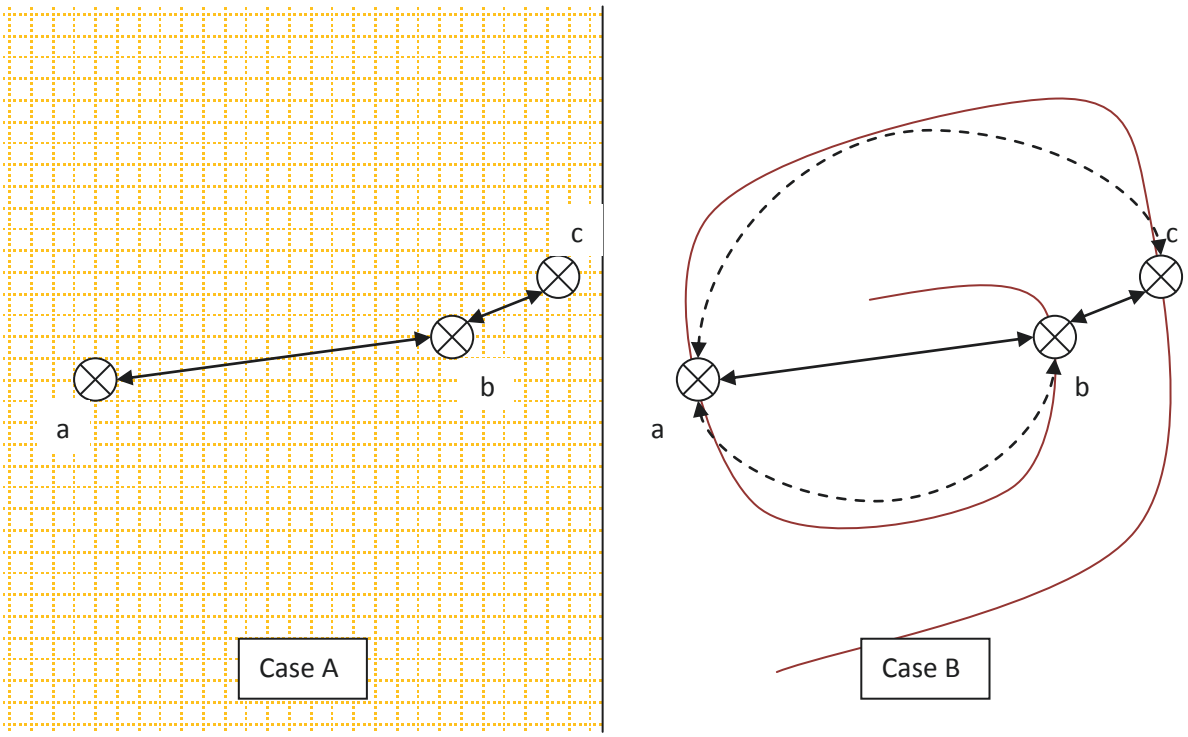
Figure 3: The ISOMAP algorithm applied to age estimation: radiological ages and ages estimated by the forensic examiner are seemingly randomly distributed along a line.

The ISOMAP algorithm identified 101 individuals significantly connected with each other out of 233. The more similar two individuals are in terms of clinical, dental and radiological data, the closer they are to each other. Despite this, two identical or similar individuals can have rather different radiological ages (top) or different ages as estimated by the forensic examiner (bottom). No order seems to be identified along the line. 40 individuals out of 101 have been randomly chosen and reported here.

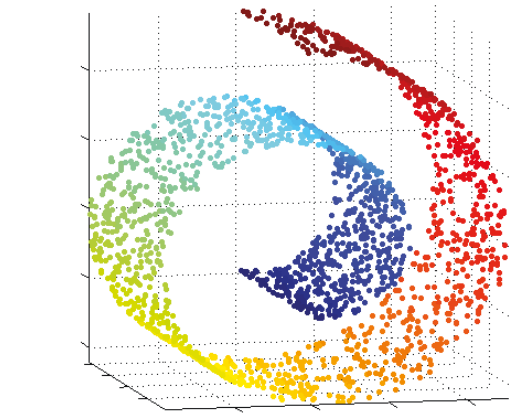
Figure 4: The autoencoder algorithm applied to age estimation: alleged ages (A), radiological ages (B) and ages estimated by the forensic examiner (C) are seemingly randomly distributed across the whole plane.

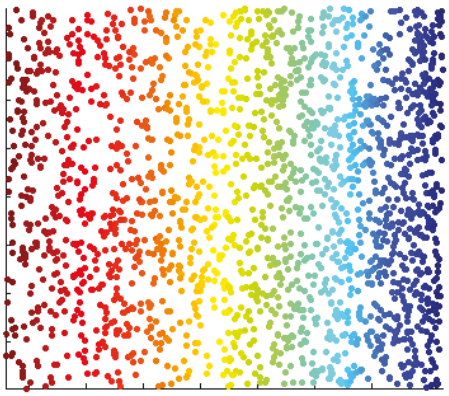
The autoencoder algorithm is a conservative algorithm that keeps all 233 initial individuals and dispatches them across a 2-dimensional space that preserves their topological relationships. The more similar two individuals are in terms of clinical, dental and radiological data, the closer they are to each other. Despite this, two identical or similar individuals can have rather different alleged ages (A). No specific age clustering seems to prevail across the whole plane. The same observation can be done for radiological (B) or forensic estimated ages (C). 50 individuals out of 233 have been randomly chosen and reported here. Only selected points are labelled to ensure readability.

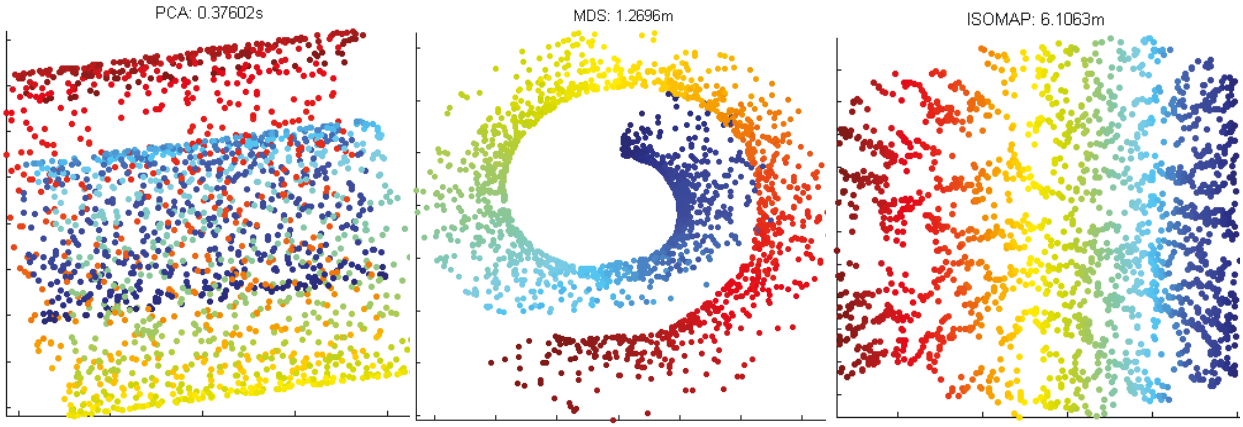
Figure



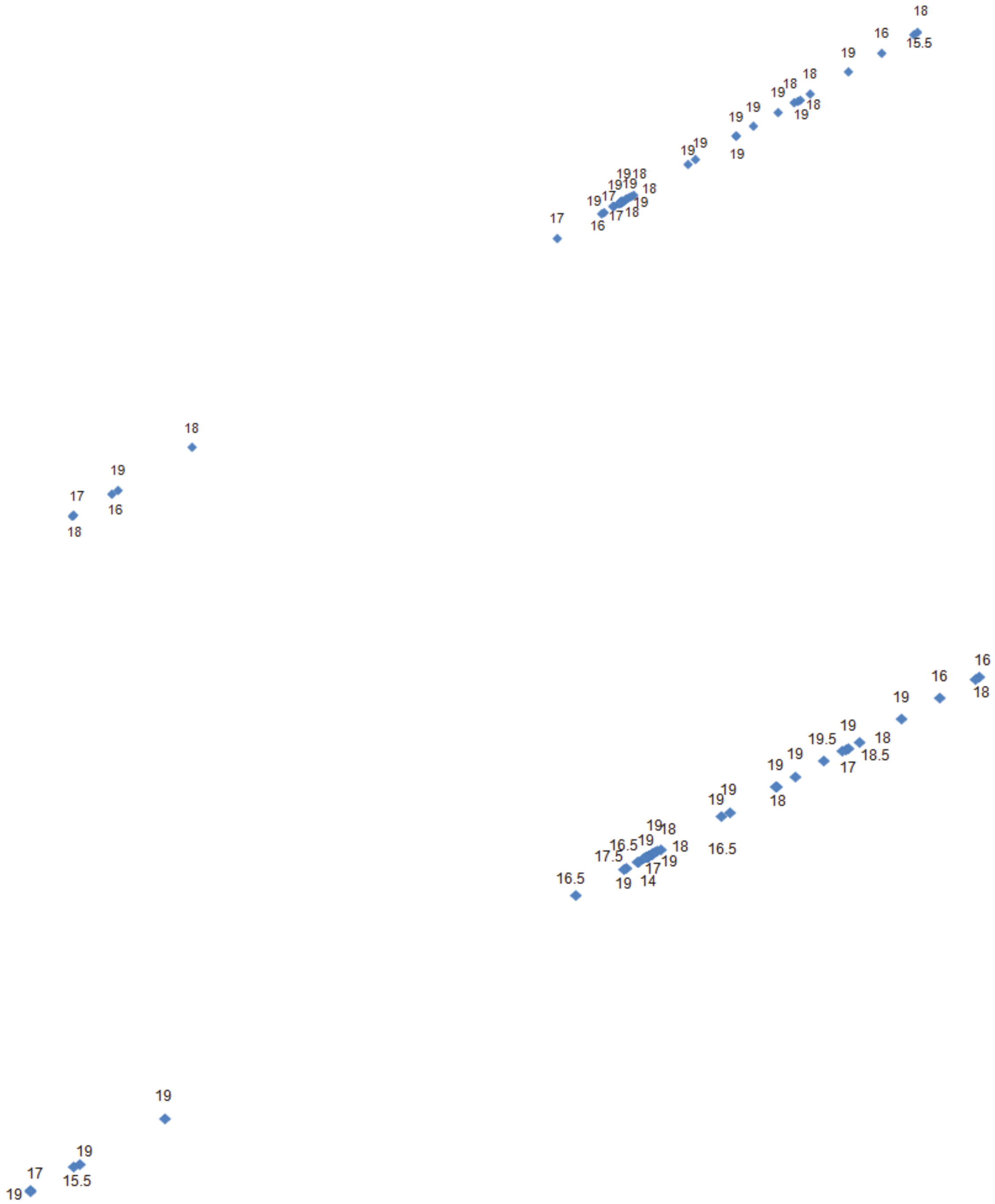
Figure



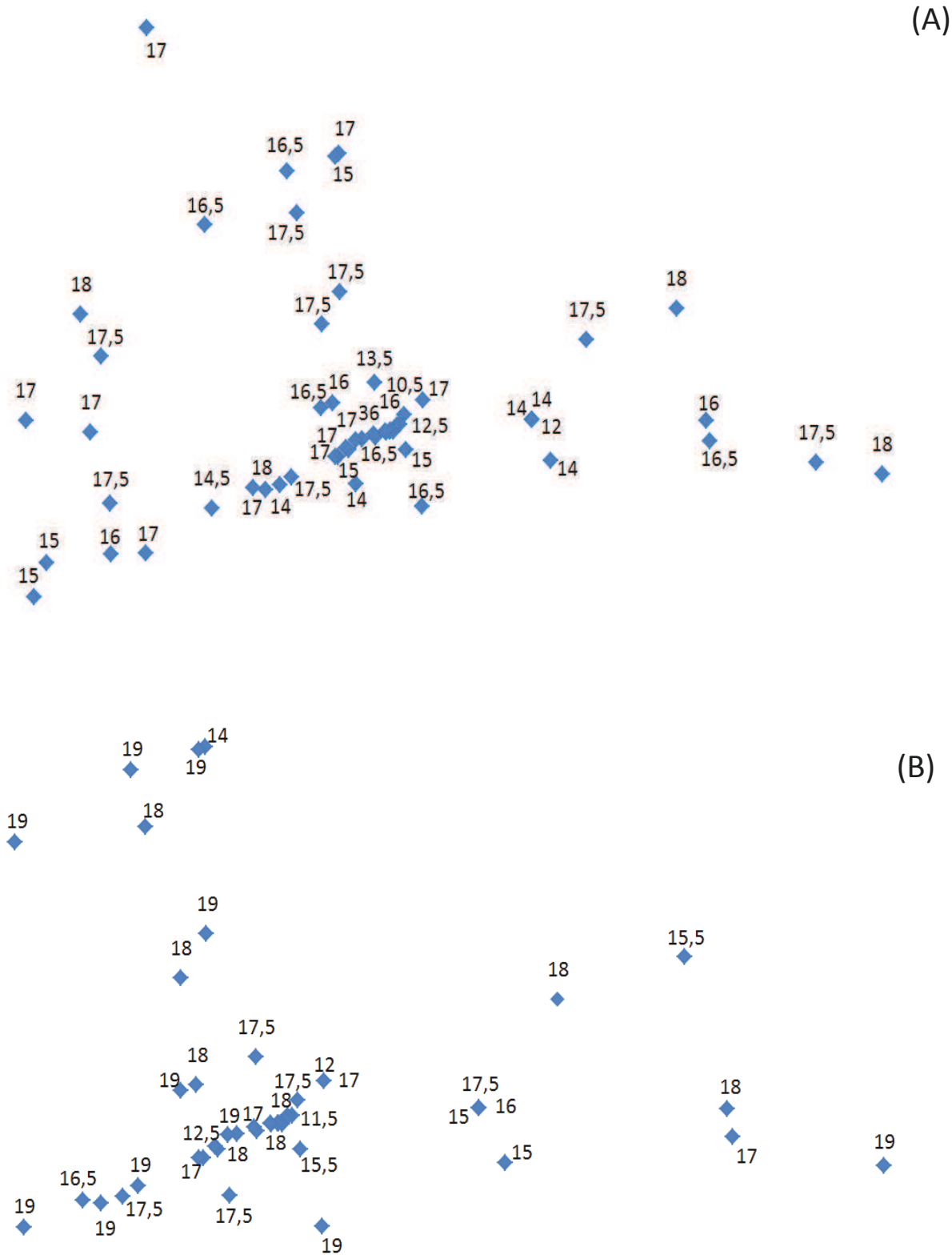




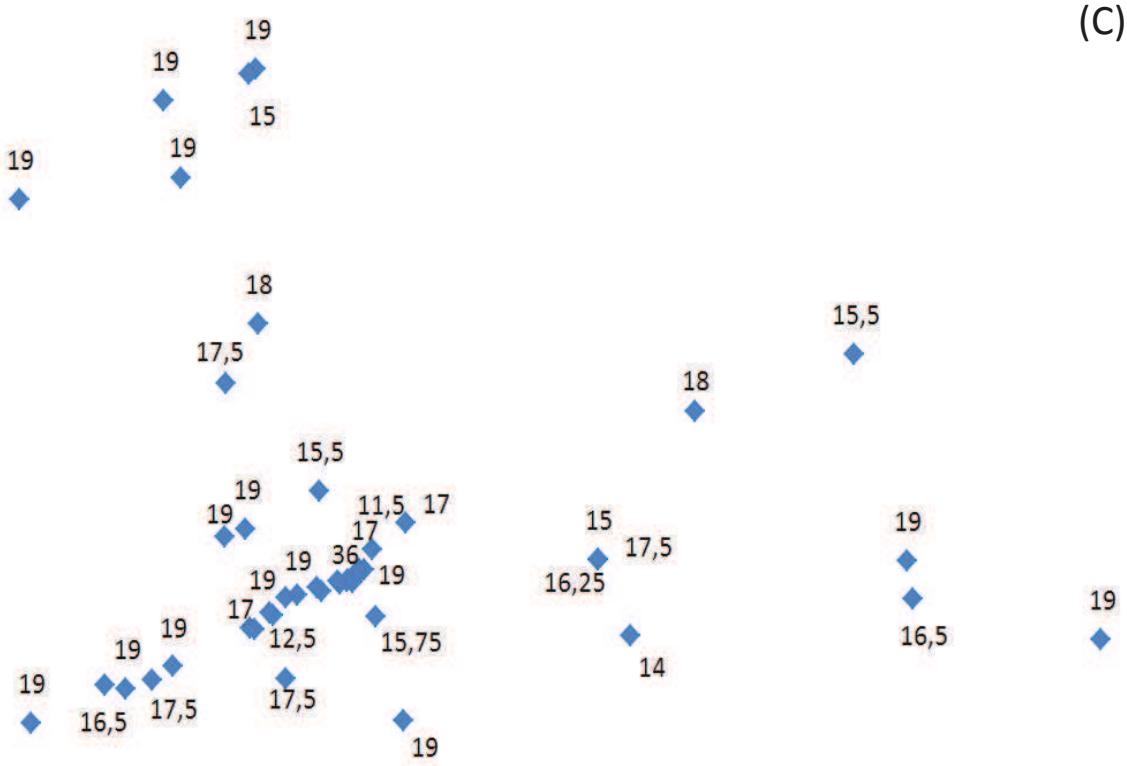
Figure



Figure


(A)
17
16,5
17
15
16,5
17,5
17,5
18
17,5
17,5
18
17,5
18
17,5
13,5
16,5
16
10,5
17
17
17
36
16
12,5
17
17
16,5
15
15
14,5
18
17,5
17
14
15
16,5
14
14
12
14
16
16,5
17,5
18
17,5
16
17
15
15
(B)
19
14
19
19
18
19
18
17,5
18
19
12
17
17,5
18
19
17
11,5
12,5
18
18
17
15,5
16,5
19
17,5
19
17,5
19
15,5
18
17,5
15
16
15
18
17
19

(C)


19
19
19
15
19
18
17,5
15,5
18
15,5
19
19
11,5
17
17
19
36
19
19
17
19
12,5
15,75
19
19
19
19
16,5
17,5
17,5
19
15
17,5
16,25
14
19
16,5
19

Table

Steps	Description
Step 1: construction of a neighborhood graph	For each data point, the algorithm builds the graph (i.e. the connection) between it and its *k*-nearest neighbors. The graph is distance-weighted. This means that each edge is associated with a Euclidean distance, which is measured between the two points (also called “vertices”).
Step 2: computation of geodesic paths for the purpose of building the geodesic distances matrix:	For each pair of points, the geodesic distance is approximated in two ways. For direct neighbors, the geodesic distance is approximated by the Euclidean distance. Otherwise, the shortest path between two points is computed through a classical graph algorithm, such as the Dijkstra algorithm.
Step 3: use of a classical multidimensional scaling (MDS) algorithm to reduce dimensionality:[S1]	The MDS is run on the geodesic distance matrix and delivers the dimensionally reduced dataset.

Table

	Dataset examples			
	Swiss roll	Toroïdal helix	Twin peaks	Infinite
Intrinsic dimension to retrieve	2	1	2	2
Estimators :				
Correlation dimension	1.94	1.47	2.02	2.29
Nearest neighbor dimension	0.56	0.7	0.51	0.44
GMST dimension	1.77	1.38	2.57	2.5
Packing numbers dimension	2.22	1.11	1.31	0.91
PCA eigenvalues dimension	2	2	2	2
Maximum likelihood dimension	1.94	1.5	2.14	2.53

Table

2^{nd} molars	17**	27**	37*	47*
Present	18.5 [15.5; 19.0]	18.5 [15.5; 19.0]	18.5 [15.5; 19.0]	18.0 [15.5; 19.0]
Absent	15.8 [10.5; 19.0]	15.5 [10.0; 17.5]	17.0 [11.0; 19.0]	16.0 [10.0; 19.0]

3^{rd} molars	18**	28**	38**	48**
Present	19.0 [17.5; 19.0]	19.0 [17.5; 19.0]	19.0 [17.5; 19.0]	19.0 [17.5; 19.0]
Absent	17.5 [14.5; 19.0]	17.5 [14.5; 19.0]	17.5 [14.5; 19.0]	17.3 [14.5; 19.0]

Gender**	
Male	19.0 [15.5; 19.0]
Female	17.5 [15.0; 19.0]

Optional e-only supplementary files

Click here to download Optional e-only supplementary files: Supplementary_material.doc

Supplementary material

Supplementary material 1. The empty phenomenon and the curse of dimensionality

The empty phenomenon: In high-dimensional datasets, two phenomena conspire to prevent effective and valid analyses. This is a natural extension of situations that we are used to encountering in lower-dimensional spaces. For instance, consider a genomic study involving 30 patients in whom 100 genes with 2 mutations each are screened for. From a purely mathematical standpoint, some conditions are “empty” because there are not enough patients to cover at least the 2^100 possibilities. This is therefore known as the “empty phenomenon” [7,8].

The curse of dimensionality: metrics like Euclidean distance can behave strangely according to the dimensionality of data. Euclidean distance usually operates correctly in lower-dimensional spaces but not in higher-dimensional ones [7,8]. The capacity of a metric to distinguish between two different points tends to vanish when dimensionality increases. Ultimately, no distinction can be made between points (patients or measures), since they all tend to look the same and each presents roughly the same characteristics (namely, the mean of the measures).

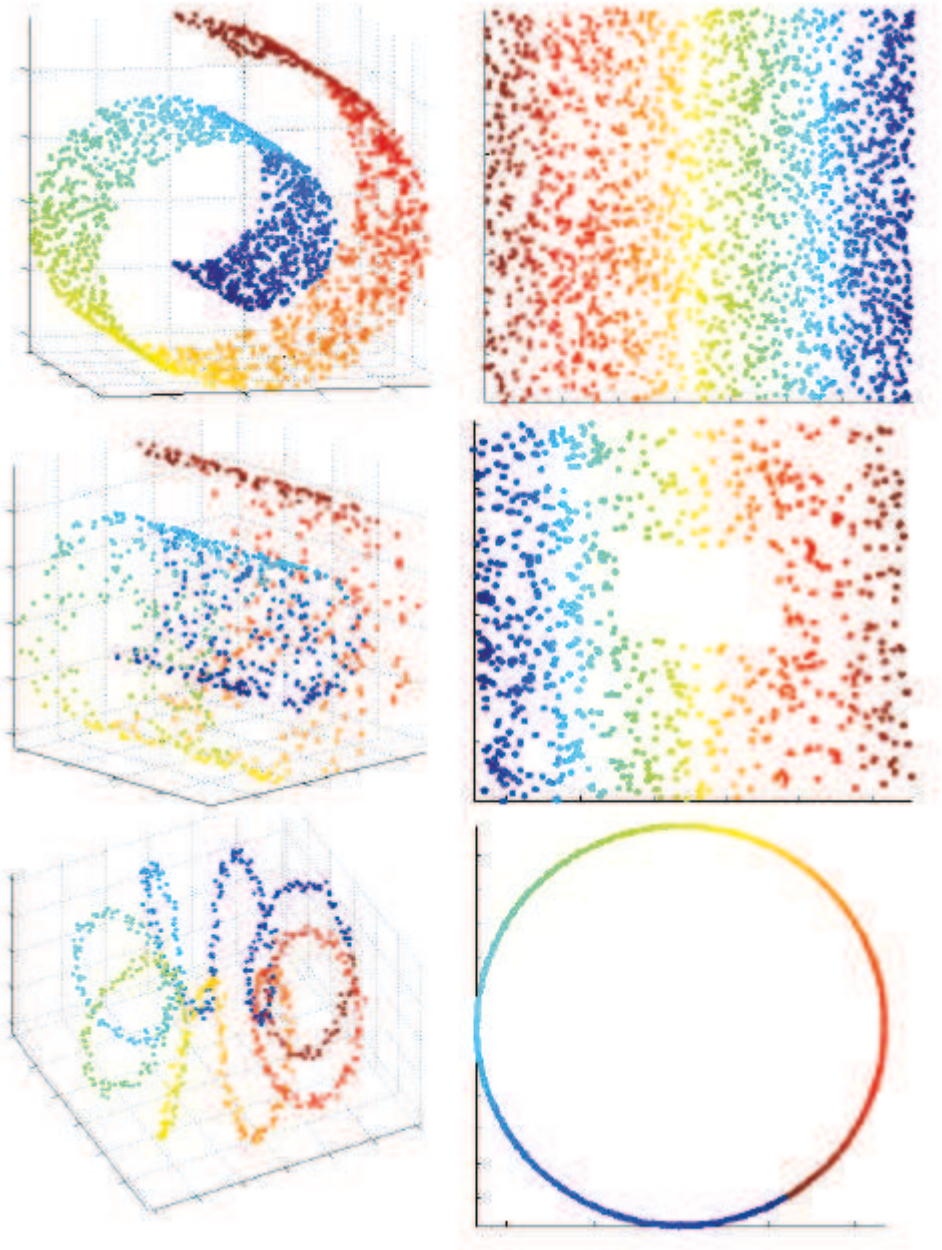
For these two additional reasons, it appears mandatory to stay in low-dimensional settings so that our tools can still operate.

Supplementary material 2. Software resources

NLDR techniques constitute an active research field and none of them has yet been fully integrated into any of the usual statistical packages. Classical MDS is provided with STATA. A quasi-systematic review of software resources is provided in [10]. NLDR techniques are freely available for MATLAB users in three different toolboxes (DR toolbox [29], Shogun toolbox [S7,S8] and a toolbox proposed by Lee and Verleysen [S9]) and in some demonstration code (the MANI demonstrator [S10]). More elementary codes can be found on different algorithm authors’ websites [S11,S12] algorithm. Some code is also available for the R project, such as the RDRToolbox [S13] and the MDR package for R [S14]. Python implementation is also available [S15].

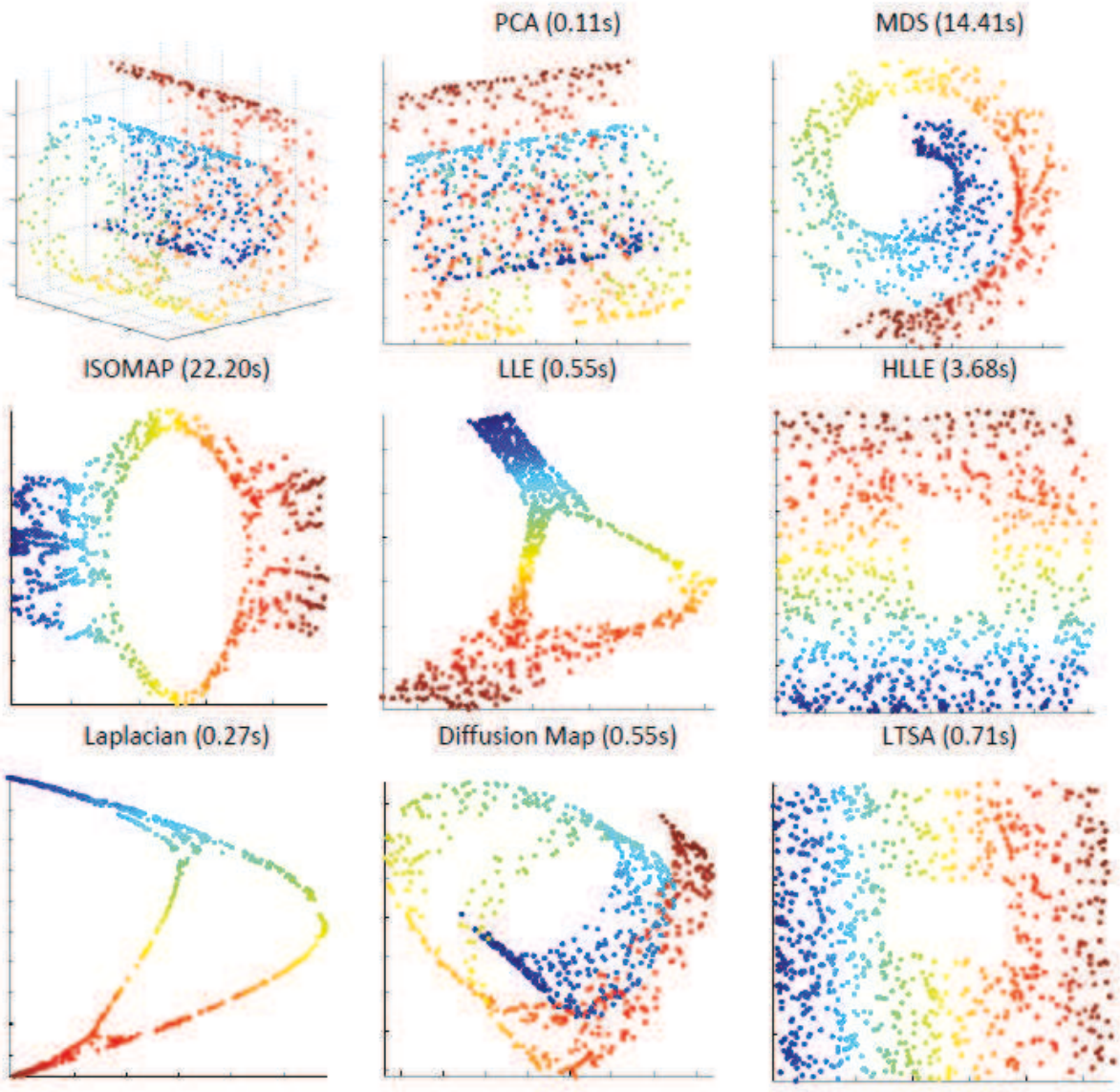
Supplementary figures.

Supplementary figure 1. Results of six NLDR techniques versus PCA and MDS applied to the theoretical example of the Swiss roll (and their execution times in seconds or minutes).



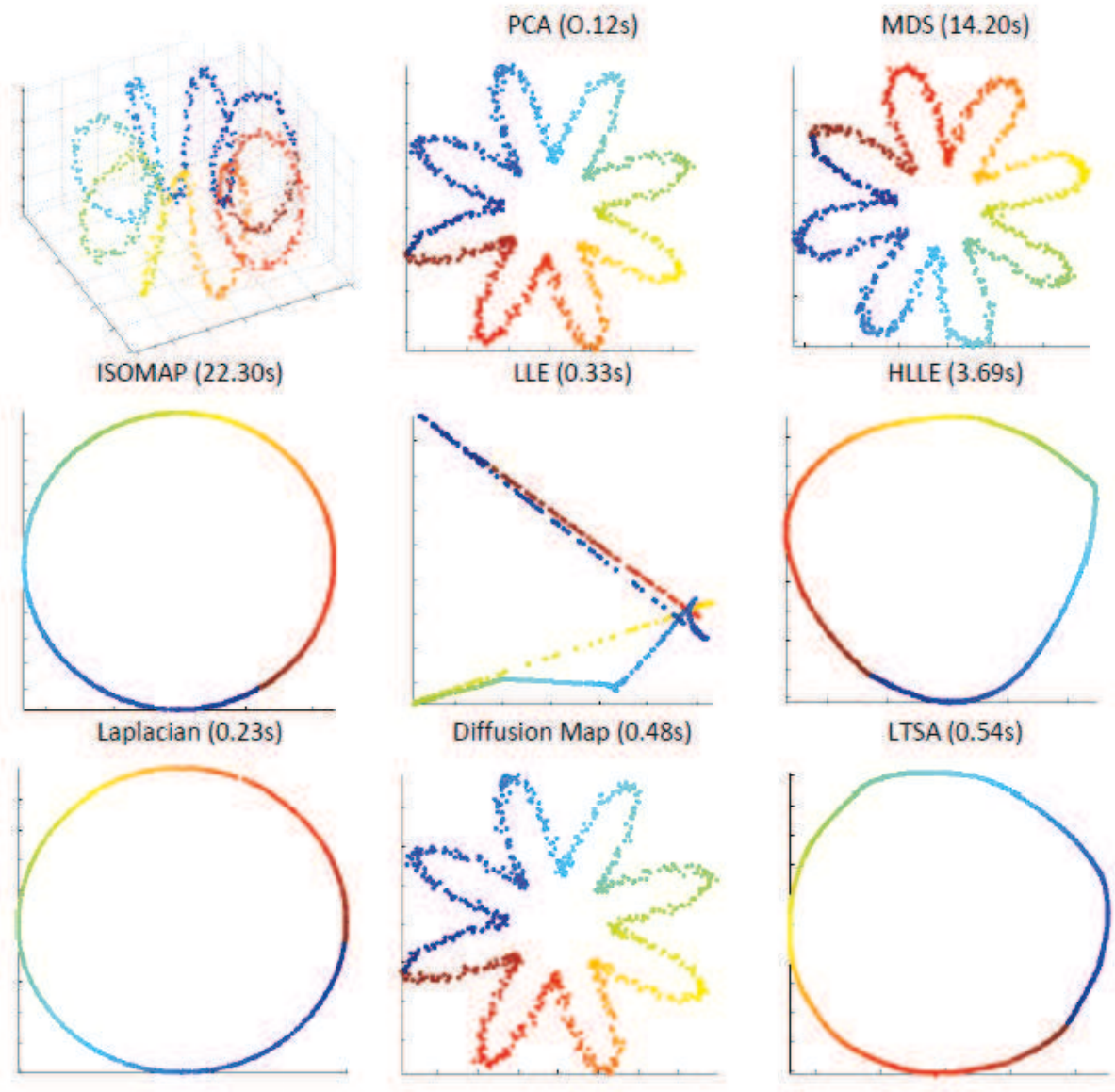
LTSA, Hessian LLE and ISOMAP algorithms perform perfectly on the Swiss roll example and succeed in unfolding the dataset's structure. LTSA appeared by far to be the fastest one. Algorithms other than ISOMAP are documented in [10,S16-S18]

Supplementary figure 2. Results of six NLDR techniques versus PCA and MDS applied to the theoretical example of the Swiss roll with a hole variant (and their execution times in seconds or minutes).



LTSA, Hessian LLE and ISOMAP algorithms perform perfectly on the Swiss roll example and succeed in unfolding the dataset's structure. LTSA appeared by far to be the fastest one. Algorithms other than ISOMAP are documented in [10,S16-S18].

Supplementary figure 3. Six NLDR techniques versus PCA and MDS applied to the theoretical example of the toroidal helix (and their execution times in seconds or minutes).



LTSA, Hessian LLE, Laplacian eigenvalues and ISOMAP algorithms perform perfectly on the Swiss roll example and succeed in unfolding the dataset's structure. LTSA and Laplacian methods were both by far the fastest ones. Algorithms other than ISOMAP are documented in [10,S16-S18].

Supplementary references

[S1] Cox TF, Cox MA. (2001) Multidimensional scaling. Chapman & Hall/CRC, Boca Raton.

[S2] Costa J, Hero A. (2003) Manifold Learning with Geodesic Minimal Spanning Trees. Computing Research Repository - CORR 2003;cs.CV/03 (accessed 12 Jan 2015)

[S3] Kégl B. (2002) Intrinsic Dimension Estimation Using Packing Numbers. Advances in Neural Information Processing Systems 15:681–688.

[S4] Fan M, Gu N, Qiao H, Zhang B. (2010) Intrinsic dimension estimation of data by principal component analysis. Computing Research Repository - CORR 2010;abs/1002.2.

[S5] Mandelbrot B. (1967) How Long Is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension. Science 156:636–638.

[S6] Lopes R, Betrouni N. (2009) Fractal and multifractal analysis: A review. Medical Image Analysis 13:634–649.

[S7] Sonnenburg S, Rätsch G, Henschel S, et al. (2010) The SHOGUN Machine Learning Toolbox. Journal of Machine Learning Research 11:1799−1802.

[S8] Shogun - A Large Scale Machine Learning Toolbox, http://shogun-toolbox.org/. (accessed 12 Jan 2015)

[S9] Machine Learning Group. Applied and fundamental research in machine learning, http://mlg.info.ucl.ac.be/index.php?page=NLDR. (accessed 12 Jan 2015)

[S10] MANI fold Learning Matlab Demo. Todd Wittman, Department of mathematics, University of California, Los Angeles, http://www.math.ucla.edu/~wittman/mani/. (accessed 12 Jan 2015)

[S11] A Global Geometric Framework for Nonlinear Dimensionality Reduction, http://isomap.stanford.edu/. (accessed 12 Jan 2015)

[S12] Locally Linear Embedding, http://www.cs.nyu.edu/~roweis/lle/. (accessed 12 Jan 2015)

[S13] Bartenhagen C (2014). *RDRToolbox: A package for nonlinear dimension reduction with Isomap and LLE.* R package version 1.14.0

[S14] http://cran.r-project.org/web/packages/MDR/index.html (accessed 12 Jan 2015)

[S15] http://scikit-learn.org/stable/modules/manifold.html (accessed 12 Jan 2015)

[S16] Roweis ST. (2000) Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science 290:2323–2326.

[S17] Zhang Z, Zha H. (2005) Principal Manifolds and Nonlinear Dimension Reduction via Local Tangent Space Alignment. SIAM Journal on Scientific Computing 26:313–338.

[S18] Donoho DL. (2003) Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. Proc Natl Acad Sci USA 100:5591–5596.