Méthodes pour cartographier les tendances régionales de la prévalence du VIH à partir des Enquêtes Démographiques et de Santé (EDS)

Joseph Larmarange, Roselyne Vallo, Seydou Yaro, Philippe Msellati, Nicolas Méda

Joseph Larmarange, Roselyne Vallo, Seydou Yaro, Philippe Msellati et Nicolas Méda

Méthodes pour cartographier les tendances régionales de la prévalence du VIH à partir des Enquêtes Démographiques et de Santé (EDS)

Introduction

1 En Afrique subsaharienne, le principal indicateur épidémiologique de suivi de l'épidémie de Sida est la prévalence du VIH, c'est-à-dire la proportion de personnes infectées[1]. C'est à partir de celui-ci que l'ONUSIDA[2] estime, pays par pays, les nombres annuels de nouvelles infections et de décès qu'il publie tous les deux ans dans son rapport sur l'épidémie mondiale. La manière d'estimer les prévalences nationales du VIH a connu plusieurs révisions au cours de ces dernières années (Larmarange 2009) suite au développement, au début des années 2000, d'enquêtes nationales en population générale avec collecte d'échantillons sanguins et dépistage du VIH. Dans la majorité des cas, il s'agit d'Enquêtes Démographiques et de Santé (EDS) auxquelles un module de dépistage du VIH a été ajouté. Les EDS ont permis d'affiner les estimations de la prévalence du VIH des adultes (15-49 ans) tant au niveau national que régional[3] et constituent, le plus souvent, la seule source de données en population générale.

2 Une majorité des EDS collecte également les coordonnées géographiques (longitude et latitude) des zones enquêtées. Dès lors, il est envisageable d'exploiter ces enquêtes afin d'estimer et de représenter les variations spatiales de la prévalence du VIH à une échelle infrarégionale. Une telle cartographie constituerait un outil de santé publique permettant de mettre en évidence les zones les plus touchées par l'épidémie, de guider la mise en œuvre des actions de lutte contre le Sida et de mieux comprendre les écarts observés entre les EDS et la surveillance sentinelle des femmes enceintes[4], autre source majeure de données utilisée pour l'estimation des prévalences (Boerma, Ghys et Walker 2003).

Les Enquêtes Démographiques et de Santé (EDS)

3 Les Enquêtes Démographiques et de Santé (EDS) constituent un important programme d'enquêtes mené dans plus de 75 pays du Sud. Depuis 1984, plus de 200 enquêtes ont été réalisées à intervalles réguliers. Aux questions relatives à la fécondité, au planning familial et à la mortalité infanto-juvénile, ont été progressivement ajoutés, selon les pays et les enquêtes, des modules sur la santé de la mère et de l'enfant, les connaissances et comportements vis-à-vis du VIH/Sida et des infections sexuellement transmissibles (IST), les violences domestiques, les mutilations génitales féminines, les mesures anthropométriques des enfants, les tests d'anémie, la prévalence du VIH, etc. Les questionnaires sont standardisés afin de faciliter les comparaisons dans le temps et entre pays.

4 Elles sont conduites à intervalle régulier par les instituts nationaux de la statistique, avec l'appui technique de Macro International Inc. L'ensemble des rapports finaux ainsi que les bases de données sont disponibles gratuitement sur un site dédié : http://www.measuredhs.com.

5 Les EDS présentent un plan de sondage comparable dans chaque pays. Il s'agit d'enquêtes stratifiées avec un tirage à deux degrés. Le pays est divisé en plusieurs strates, une par région administrative et par milieu de résidence. La base de sondage des unités primaires est composée des zones de dénombrement au dernier recensement général de la population et de l'habitat (RGPH). Au premier degré, les unités primaires ou grappes sont tirées au sort, séparément dans chaque strate, avec une probabilité proportionnelle à leur nombre de ménages ordinaires[5] dans le RGPH. Après un recensement exhaustif des ménages de chaque grappe, un

nombre prédéterminé de ménages est sélectionné au second degré, par tirage au sort simple. Suivant le pays, seule une partie (le tiers ou la moitié) des ménages enquêtés est retenue pour le dépistage du VIH. Tous les adultes (15-49 ans en général) de ces ménages sont alors testés. Afin de tenir compte du plan d'échantillonnage complexe des EDS, chaque base de données contient une variable de pondération, ce qui permet de rendre l'échantillon représentatif au niveau national et régional. Cette variable est proportionnelle à l'inverse de la probabilité de sondage de chaque ménage, c'est-à-dire à la probabilité que le ménage en question soit enquêté. Le Tableau 1 permet de comparer l'échantillonnage de plusieurs EDS récentes.

6 Certaines EDS collectent également par GPS les coordonnées géographiques du centre de chaque grappe enquêtée. Depuis l'arrivée des modules de dépistage du VIH et afin de garantir l'anonymat des personnes enquêtées, ces coordonnées sont, dans les bases de données fournies par Measure DHS, décalées aléatoirement dans un rayon de deux kilomètres en milieu urbain et cinq kilomètres en milieu rural[6].

7 À partir de 2004, des enquêtes similaires aux EDS mais spécifiques à la problématique VIH/Sida avec un questionnaire allégé ont été élaborées : les *AIDS Impact Surveys*[7] (AIS).

Tableau 1 : Échantillonnage pour le dépistage du VIH de 25 EDS ou AIS

Pays	Année	Type	Grappes	Personnes testées pour le VIH 15-49 ans	Nombre moyen de pers. testées par grappe	Prévalence nationale du VIH 15-49 ans (en %)
Burkina Faso	2003	EDS	400	7 244	18,1	1,8
Cameroun	2004	EDS	466	9 900	21,2	5,5
Côte d'Ivoire	2005	AIS	249	8 436	33,9	4,7
Éthiopie	2005	EDS	540	10 540	19,5	1,4
Ghana	2003	EDS	412	9 144	22,2	2,2
Guinée	2005	EDS	297	6 388	21,5	1,5
Kenya	2003	EDS	400	6 001	15,0	6,7
Kenya	2008-09	EDS	400	6 707	16,8	6,3
Lesotho	2004	EDS	405	5 043	12,5	23,4
Liberia	2007	EDS	300	11 733	39,1	1,6
Malawi	2004	EDS	522	5 150	9,9	12,0
Mali	2001	EDS	403	6 475	16,1	1,8
Mali	2006	EDS	407	8 141	20,0	1,3
Niger	2006	EDS	345	7 262	21,0	0,7
Ouganda	2004-05	AIS	417	16 906	40,5	6,4
République Démocratique du Congo	2007	EDS	300	8 504	28,3	1,3
Rwanda	2005	EDS	462	10 016	21,7	3,0
Sénégal	2005	EDS	377	7 503	19,9	0,7
Sierra Leone	2008	EDS	353	6 174	17,5	1,5
Swaziland	2006-07	EDS	275	8 187	29,8	25,9
Tanzanie	2003-04	AIS	345	10 747	31,2	7,0
Tanzanie	2007-08	AIS	475	15 044	31,7	5,7
Zambie	2001-02	EDS	320	3 807	11,9	15,6
Zambie	2007	EDS	320	10 444	34,8	14,3
Zimbabwe	2005-06	EDS	400	12 796	32,0	18,1

EDS : Enquête démographique et de Santé ; AIS : AIDS Impact Survey.
Sources : rapport final de chaque enquête disponible sur http://www.measuredhs.com.

Objectif

8 Si les publications portant sur les EDS sont nombreuses, les analyses spatiales à partir de ces dernières sont plus limitées. À titre d'exemple, sur le site de Measure DHS où plus de

650 articles scientifiques portant sur les données des EDS ont été identifiés, seuls 13 sont classés dans la catégorie « modélisation spatiale »[8]. Les travaux les plus fréquents sont des atlas présentant des cartes choroplèthes (TACAIDS 2006) ; des cartes au niveau national ou régional comme le propose l'outil en ligne *HIVmapper*[9] ; ou bien des analyses multi-niveaux intégrant une ou plusieurs variables géographiques (distance à une route ou à une infrastructure, typologie spatiale, etc.). Cela est facilité par la mise à disposition sur internet, depuis quelques années, de fonds de carte géoréférencés des unités administratives utilisées dans les EDS[10].

9 Les cartes choroplèthes par régions administratives ne sont pas toujours adaptées pour rendre compte de la spatialisation d'un phénomène, les frontières administratives correspondant rarement à des limites propres au phénomène. Par ailleurs, pour les régions fortement peuplées et donc disposant d'un échantillonnage important, les cartes par régions induisent une perte d'information à un niveau plus local : les différences infrarégionales sont masquées.

10 Notre objectif vise donc à estimer, à partir des données des EDS et indépendamment du découpage administratif du territoire, une surface des prévalences mettant en évidence les principales variations spatiales de l'épidémie, tout en conservant une précision locale infrarégionale dans les zones suffisamment enquêtées.

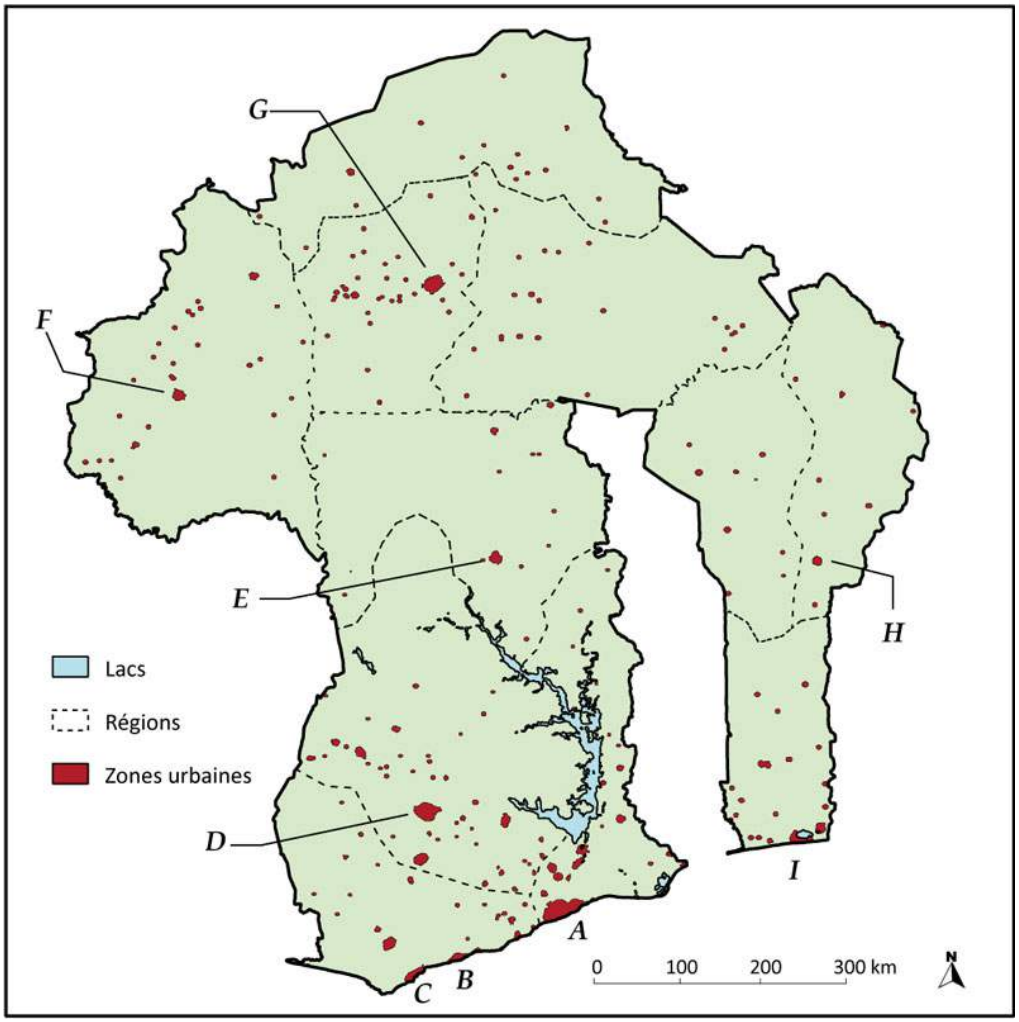
11 Afin de pouvoir tester plusieurs approches méthodologiques, nous avons élaboré un pays fictif à partir duquel nous avons simulé des EDS : il est dès lors possible de comparer la surface des prévalences estimées à partir des données d'enquêtes avec la surface des prévalences d'origine du modèle. Plusieurs approches ont été testées : la première reposant sur un lissage par des cercles de même effectif avant interpolation spatiale, la seconde adaptée des travaux de Davies et Hazelton (2010) à partir d'estimateurs à noyau à fenêtres adaptatives et une troisième, toujours à partir d'estimateurs à noyau, utilisant des cercles de même effectif. Enfin, une application sur des données réelles a été effectuée à partir de l'EDS 2003 du Burkina Faso.

Méthodes

Élaboration d'un pays modèle

12 Le Bénin, le Burkina Faso et le Ghana ont été agrégés pour créer un pays fictif servant de modèle. Le Togo a été volontairement exclu afin d'obtenir une forme concave qui peut complexifier l'estimation, le « creux » ainsi formé n'étant pas enquêté. Une telle forme se retrouve dans les frontières de certains pays tels que le Sénégal. Les données du *Global Rural-Urban Mapping Project* (GRUMP) ont été utilisées pour distribuer la population sur le territoire : densités de population en l'an 2000 avec une résolution de 30 secondes d'arc (CIESEN *et al*. 2005a) et découpage du territoire en milieu urbain et milieu rural (CIESEN *et al*. 2005b). Le territoire a ensuite été divisé en 9 137 unités primaires (7 818 unités rurales et 1 319 unités urbaines) espacées régulièrement selon une résolution moyenne de 2 minutes d'arc en milieu urbain et 5 minutes d'arc en milieu rural. Puis ont été calculées la superficie, la densité moyenne et la population (obtenue en multipliant la densité par la superficie) de chaque unité. Le pays a été divisé en 11 régions administratives et les principaux centres urbains ont été renommés par une lettre, allant de A à I (voir figure 1).

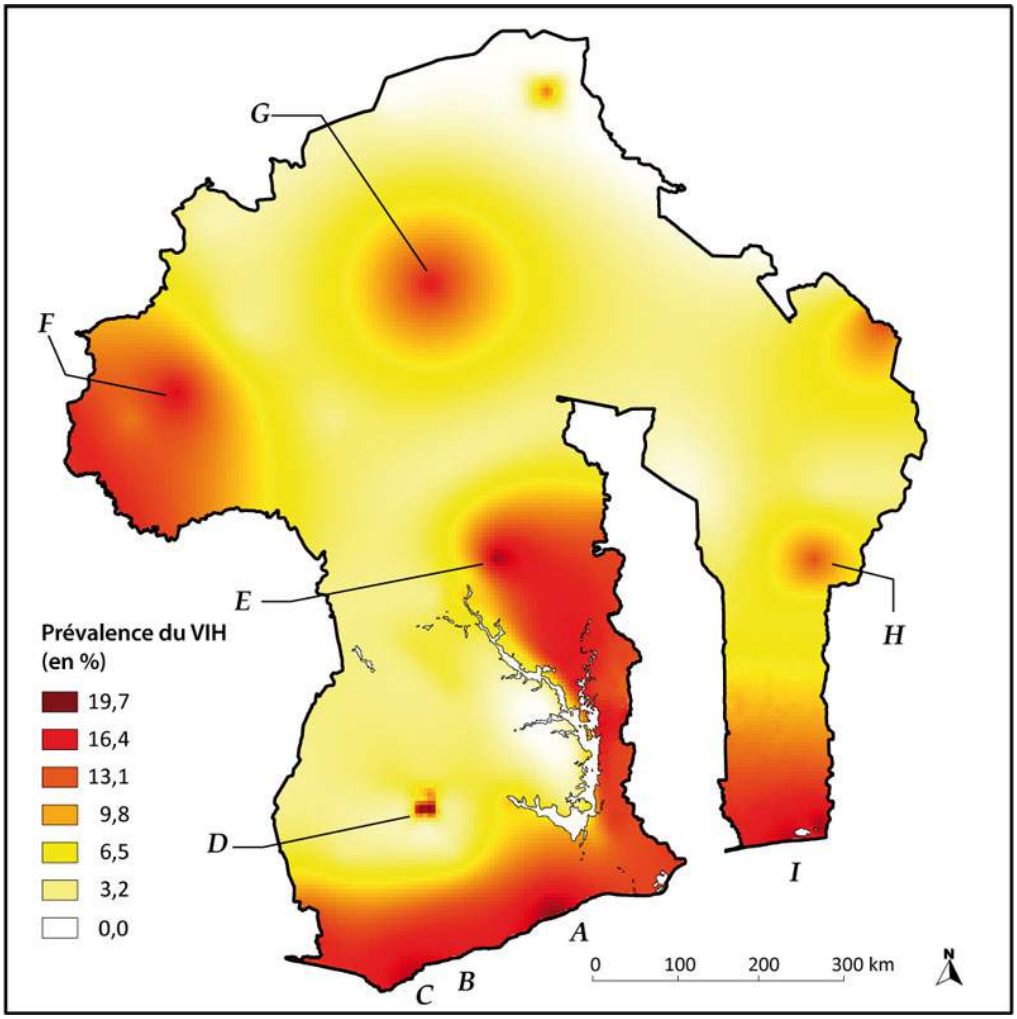
Figure 1 : Zones urbaines et régions du modèle



13 Dans un second temps, nous avons créé une surface de prévalences (figure 2) par interpolation spatiale[11] à partir de points choisis de manière *ad hoc* afin que cette surface présente différents modèles de diffusion : ville importante avec une prévalence concentrée sur celle-ci et faible autour (D) ; ville importante (G) et moyenne (H) avec diffusion progressive ; pic localisé en zone rurale peu peuplée dans le nord du pays ; rupture de continuité de part et d'autre d'un grand lac ; gradient des côtes vers l'intérieur des terres au sud du pays, avec deux grandes agglomérations (A et I) et deux villes moyennes (B et C) ; diffusion depuis la frontière ouest et depuis une ville située de l'autre côté d'une frontière (à l'est).

14 La prévalence nationale s'obtient en faisant la moyenne de la prévalence des unités primaires[12] pondérée par leur population. Afin d'obtenir une prévalence nationale de 10 %, nous avons multiplié la prévalence de chaque unité par un même facteur d'échelle. La surface obtenue a pour caractéristique, de par sa construction, d'être spatialement continue et fortement auto-corrélée.

Figure 2 : Surface des prévalences du VIH dans le modèle (prévalence nationale de 10 %, crée de manière *ad hoc* par interpolation spatiale)



Format des données et simulation d'EDS

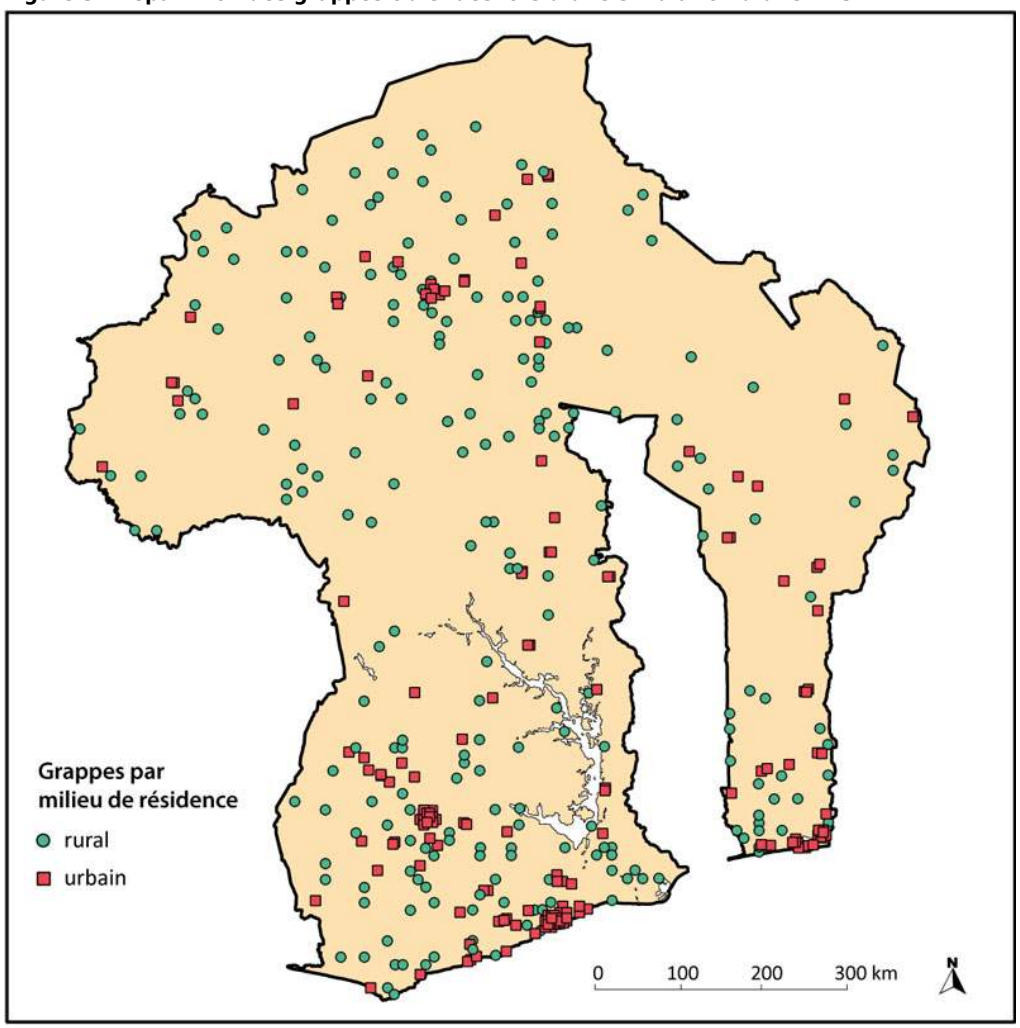
15 Les données des EDS se présentent sous la forme d'un semis de points correspondant aux différentes grappes enquêtées. Dans la mesure où les grappes sont sélectionnées avec une probabilité proportionnelle à leur nombre de ménages, ce semis de points reflète les variations de la densité de population : les points sont nombreux et proches dans les zones fortement peuplées et, à l'inverse, rares et espacés dans les zones faiblement peuplées. Pour chaque individu testé, nous disposons de son statut sérologique, de sa grappe d'appartenance et de son poids statistique. Tous les individus d'une même grappe sont spatialement positionnés sur le même point.

16 Des simulations d'EDS ont été réalisées pour reproduire des données analogues aux enquêtes réelles, à partir de trois paramètres : la prévalence nationale, le nombre total de personnes enquêtées et le nombre de grappes au premier degré. Afin d'obtenir le niveau de prévalence nationale choisi, la prévalence de chacune des unités primaires est multipliée par un même facteur d'échelle adéquat. Chacune des onze régions est divisée en deux strates, l'une urbaine et l'autre rurale. Le nombre de grappes tirées par strate est proportionnel à la population totale de celle-ci[13]. Les grappes sont tirées aléatoirement, strate par strate parmi les unités primaires, avec une probabilité de sondage proportionnelle à leur population.

17 Dans un second temps, l'effectif de personnes enquêtées par grappe est déterminé aléatoirement, selon une loi normale, afin de reproduire la variabilité du nombre de personnes enquêtées par grappe, variabilité que l'on peut observer dans les EDS[14]. Ce nombre est redressé pour que le total corresponde à l'effectif visé. Enfin, le nombre de personnes séropositives de chaque grappe est déterminé aléatoirement selon une loi binomiale. Un facteur de pondération, analogue à celui utilisé dans les EDS, est calculé et appliqué aux individus.

18 Les données générées par simulation sont de nature comparable aux données réelles des Enquêtes Démographiques et de Santé. En effet, la distribution des prévalences observées générées par une simulation présente le même profil que celles observées dans plusieurs EDS (tableau non reproduit). La figure 3 présente la répartition spatiale des grappes obtenues lors d'une simulation donnée, simulation qui nous servira d'exemple dans la suite de cet article.

Figure 3 : Répartition des grappes obtenues lors d'une simulation d'une EDS.



19 Paramètres de la simulation : prévalence nationale de 10 %, 8 000 personnes enquêtées, 400 grappes (en raison d'arrondis, le nombre de grappes effectivement sélectionnées est de 401).

Approche par interpolation spatiale

20 Les techniques d'interpolation spatiale permettent, à partir d'un semis de points, de produire la surface d'un phénomène. Pour chaque point de la carte où la valeur du phénomène n'est pas connue, cette dernière est estimée à partir des points pour lesquels une information est disponible. Ces différentes techniques considèrent que les variations de la variable interpolée sont continues dans l'espace et accordent un poids supérieur aux observations proches par rapport aux observations éloignées, selon l'hypothèse que des points voisins se ressemblent. Ces techniques permettent donc d'estimer un phénomène sur l'ensemble d'un territoire à partir d'une information parcellaire limitée à un nombre fini de points.

21 La formule générale pour déterminer la valeur estimée $\hat{s}(x,y)$ de la surface s au point situé en x,y, connaissant les valeurs de s en n points de coordonnées x_i,y_i, i variant de 1 à n, est la suivante :

Équation 1

$$\hat{s}(x,y) = \frac{\sum_{i=1}^{n} w(d_i)s(x_i,y_i)}{\sum_{i=1}^{n} w(d_i)}$$

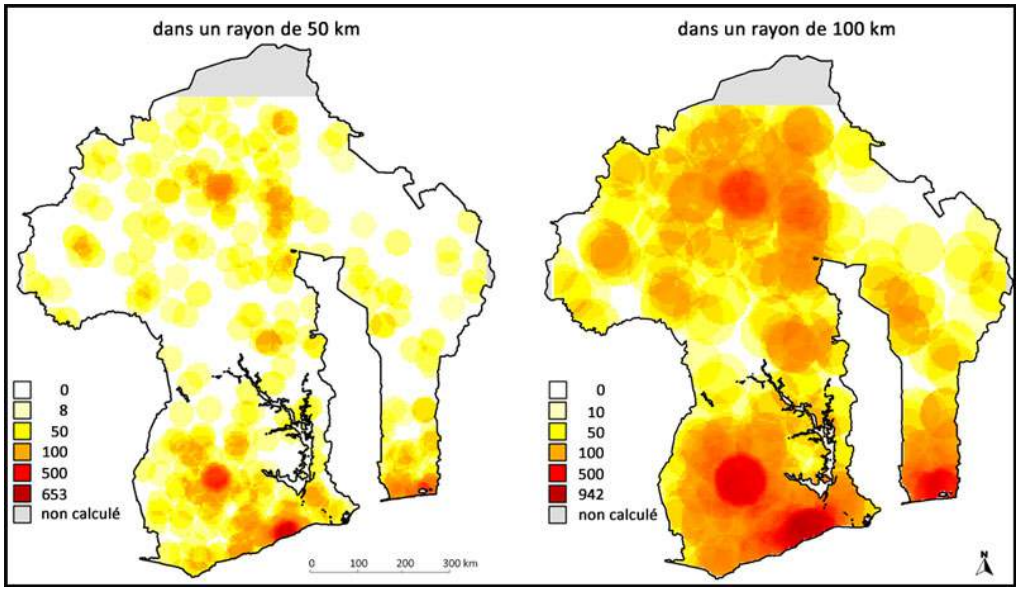
où d_i représente la distance géométrique entre l'observation i et le point situé en x,y et w une fonction de pondération décroissante accordant un poids plus faible aux observations éloignées. La plupart des fonctions de pondération utilisées conservent les valeurs observées, c'est-à-dire que la valeur estimée en un point k connu reste égale à sa valeur observée : $\hat{s}(x_k,y_k) = s(x_k,y_k)$.

22 Pour utiliser ces techniques, il est nécessaire de disposer pour chaque point observé de deux informations : sa position et la valeur de la prévalence en ce point. Or, les prévalences observées pour chaque grappe enquêtée, c'est-à-dire les prévalences calculées à partir des personnes testées dans la grappe, ont une variance et une marge d'erreur élevées. En effet, le nombre de personnes testées par grappe est faible, de 10 à 40 en moyenne (voir tableau 1). De fait, les prévalences observées reflètent plus les variations aléatoires dues à l'échantillonnage que le niveau de l'épidémie.

23 Nous avons déjà présenté ailleurs (Larmarange 2007 ; Larmarange *et al.* 2006) une approche méthodologique visant à lisser la prévalence de chaque grappe au préalable, avant de procéder à une interpolation spatiale classique.

24 S'inspirant de techniques de lissage à partir de moyennes mobiles basées sur des cercles de même rayon (Griffin 1949 ; Krumbein 1956 ; Nettleton 1954), un cercle est tracé autour de chaque grappe et la prévalence de la grappe centrale est dès lors recalculée à partir de l'ensemble des personnes testées situées à l'intérieur du cercle[15]. Une interpolation spatiale peut ensuite être réalisée à partir de ces prévalences lissées.

Figure 4 : Nombre de personnes enquêtées dans un rayon de 50 et 100 kilomètres.



Note : pour chaque point de la carte, on indique le nombre de personnes enquêtées dans un rayon de 50 ou 100 kilomètres autour de ce point, pour la simulation d'EDS de la figure 3.

25 Cependant, le recours à des cercles de même rayon s'avère non pertinent du fait de la répartition très inégale des grappes sur le territoire (figure 4). Il faut déterminer en effet un rayon suffisamment élevé pour que le calcul des prévalences lissées puisse porter sur un nombre suffisant d'individus, en particulier dans les zones où les grappes sont distantes les unes des autres. Mais dans le même temps, dans les zones fortement peuplées et enquêtées, il serait possible d'avoir recours à des cercles plus petits car les effectifs sont largement suffisants. Pour une proportion, la précision de l'estimation est liée au nombre d'observations. Il est alors plus opportun d'avoir recours à une approche par des cercles non pas de même rayon mais de même effectif.

26 Dès lors, une fois un effectif N fixé, les prévalences lissées de chaque grappe sont calculées à partir des observations situées dans un cercle tel que le nombre d'individus testés y soit au moins égal à N. Si l'on note $cum_i(r)$ la fonction donnant l'effectif cumulé des observations situées dans un rayon r autour de la grappe observée i, alors le rayon r_i du cercle de lissage pour cette grappe, ayant fixé un effectif minimum N, correspond à $min[r \mid cum_i(r) \geq N]$.

27 Pour produire une surface des prévalences, les prévalences lissées sont ensuite interpolées spatialement selon la technique du krigeage ordinaire. Le krigeage (Krige 1951 ; Matheron 1963) présente l'avantage de prendre en considération la structure de dépendance spatiale des données (Baillargeon 2005). La variance en fonction de la distance entre les points est mesurée empiriquement sous la forme d'un semi-variogramme qui est ensuite modélisé. Les valeurs inconnues sont alors estimées à partir des valeurs voisines connues, pondérées en fonction de ce semi-variogramme et de façon à obtenir une prévision non biaisée et de variance minimale.

Approches par les estimateurs à noyau

28 Un autre champ de l'analyse géostatistique repose sur l'estimation de surfaces de densité à partir d'estimateurs à noyau (Silverman 1986 ; Wand et Jones 1994). Ces techniques visent à construire une surface à partir d'un semis de points, chaque point représentant un cas observé. La surface obtenue peut être exprimée en nombre de cas par unité de surface (surface d'intensité) ou bien en ramenant son intégrale à l'unité (surface de densité).

29 Une surface de densité est construite autour de chaque cas observé de manière à ce que la densité soit maximale en ce point et diminue à mesure que l'on s'en éloigne. La surface d'intensité estimée correspond alors à la somme de ces surfaces de densité (voir Figure 5 pour un exemple à une dimension). En termes mathématiques, la surface d'intensité $\hat{s}$ au point (x,y) est estimée selon l'expression suivante :

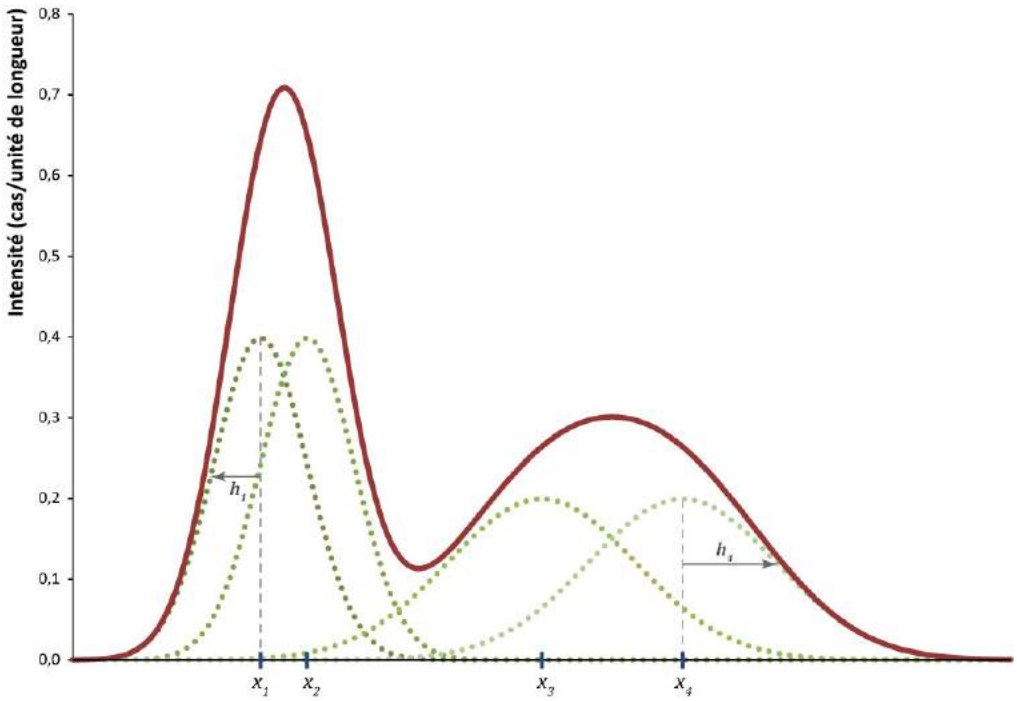
Équation 2

$$\hat{s}(x,y) = \sum_{i=1}^{n} \frac{1}{h_i^2} K\left(\frac{d_i}{h_i}\right)$$

où n est le nombre de cas observés, d_i la distance géométrique entre le cas i et le point situé en (x,y), K une fonction de densité (appelée noyau) ayant une intégrale égale à 1 et h_i la fenêtre utilisée pour le cas i. La surface de densité s'obtient, quant à elle, en divisant la surface d'intensité par le nombre de cas observés (n).

30 La fenêtre permet d'appliquer un degré de lissage plus ou moins important aux données. L'estimation sera à fenêtre fixe s'il s'agit d'une constante ($\forall i,\ h_i = h$) et, à l'inverse, à fenêtre adaptative si h_i varie selon le cas observé ($\exists\ a,b \mid h_a \neq h_b$).

Figure 5 : Exemple de calcul d'une fonction d'intensité avec un estimateur à noyau gaussien à fenêtre adaptative (cas à 1 dimension)



Note : estimation à partir de 4 points observés situés en x_1, x_2, x_3 et x_4. La fonction d'intensité estimée (courbe en trait plein) est la somme des 4 fonctions de densité (courbes en pointillés) calculées, à l'aide d'un noyau gaussien,

pour chaque point observé et centrées en ce point. h_1 et h_4 représentent la largeur de la fenêtre utilisée pour le calcul des fonctions de densité.

31 Plusieurs travaux ont utilisé les estimateurs à noyau en épidémiologie spatiale (Gatrell, Bailey, Diggle et Rowlingson 1996), notamment pour estimer une surface de risques relatifs (Bithell 1990 ; Davies et Hazelton 2010 ; Kelsall et Diggle 1995). La surface des risques relatifs correspond au ratio entre la surface de densité des cas positifs et la surface de densité des cas témoins (la population soumise au risque). Les deux surfaces de densité sont estimées séparément à partir de deux semis de points indépendants. Le semis des cas positifs est le plus souvent issu d'une surveillance épidémiologique. Les cas témoins peuvent être déterminés, pour leur part, de différentes manières : sélection aléatoire à partir d'un répertoire téléphonique pour Gatrell, un échantillon de codes postaux pour Davies et Hazelton…

32 Dans le cas de fenêtres fixes, plusieurs auteurs (Bithell 1990 ; Kelsall et Diggle 1995) suggèrent l'emploi d'une même constante h pour l'estimation des deux surfaces de densité (cas positifs et cas témoins). Plusieurs travaux (Bithell 1990 ; Carlos, Shi, Sargent, Tanski et Berke 2010 ; Davies et Hazelton 2010 ; Diggle, Rowlingson et Su 2005) suggèrent néanmoins que le recours à une fenêtre adaptative s'avère plus pertinente dans le domaine de la santé afin de mieux tenir compte de la distribution spatiale de la population et réduire ainsi le lissage de l'information.

33 La principale difficulté des estimateurs à noyau consiste à choisir une valeur adéquate pour la fenêtre de lissage. Kelsall et Diggle (1995) ont exploré différentes approches de détermination automatique de la valeur de la fenêtre à partir des données dans le cas des fenêtres fixes. D'autres travaux ont porté sur cette question dans le cas de fenêtres adaptatives (Sain 1994 ; 2002) pour l'estimation d'une seule surface mais sans faire d'investigation sur la question du ratio de deux surfaces estimées simultanément.

34 Dans un article récent, Davies et Hazelton (2010) ont proposé une approche pour estimer une surface de risques relatifs en utilisant des fenêtres adaptatives dont les valeurs sont déterminées à partir des données observées. Dans un premier temps, les auteurs déterminent une valeur pilote de la fenêtre, commune pour les cas positifs et les cas témoins, à partir du principe maximal de lissage[16] proposé par Terrell (1990). Puis, ils déterminent les valeurs locales de la fenêtre de lissage, séparément pour les cas positifs et les cas témoins, à partir de cette valeur pilote. Leur approche a été testée sur six situations théoriques différentes et comparée, en utilisant le critère ISE (*Integrated Squarred Error*), avec des estimateurs à fenêtre fixe déterminée selon le principe de lissage maximal : il apparaît dès lors que leur approche à fenêtre adaptative s'est avérée en général plus efficace, en particulier sur des échantillons de grande taille. Celle-ci a par ailleurs été implémentée dans un package nommé *sparr* pour le logiciel de statistiques *R* (Davies, Hazelton et Marshall 2010).

35 Dans notre situation, nous ne cherchons pas à produire une surface de risques relatifs (ratio de deux surfaces de densité) mais une surface des prévalences (ratio de deux surfaces d'intensité). Les fonctions fournies dans le package *sparr* permettent de calculer à la fois les surfaces de densité et les surfaces d'intensité. Nous avons donc testé l'approche proposée par Davies et Hazelton adaptée de manière à calculer le ratio de deux surfaces d'intensité au lieu du ratio de deux surfaces de densité.

36 Les travaux de Davies et Hazelton portent sur deux semis de points indépendants pour les cas positifs et les cas témoins, chacun relevant d'un échantillonnage aléatoire simple : la localisation des cas positifs diffère donc de celle des cas témoins. Or, dans le cadre des EDS, les données sont issues d'un échantillonnage à deux degrés et les semis de points correspondent à l'échantillonnage au premier degré. Les cas positifs (personnes testées positives au VIH) et les cas témoins (personnes testées quel que soit le résultat) d'une même grappe ont la même localisation spatiale. Dès lors, l'utilisation d'une même fenêtre hi pour les cas d'une même grappe semble plus appropriée.

37 Nous avons donc testé une autre approche à partir d'estimateurs à noyau à fenêtres adaptatives de manière à ce que la fenêtre utilisée pour les cas d'une même grappe ne dépende que de leur localisation et, plus précisément, du nombre d'observations dans le voisinage de la grappe. Concernant l'estimation de la surface d'intensité des cas observés, le principe est similaire à la technique des plus proches voisins décrite par Silverman (1986) ou encore Altman (1992) et testée par Bithell (1990). Un effectif N d'observations minimum est fixé et le rayon h_i de la

fenêtre de lissage est alors proportionnel au rayon du cercle à tracer autour de la grappe afin de capturer cet effectif minimum. Il s'agit de fait des rayons r_i des cercles de lissage de même effectif décrits précédemment dans le cadre de l'approche par interpolation spatiale. Dans la situation particulière des données EDS, les cas témoins d'une même grappe se voient attribuer une même fenêtre : si les points *i* et *j* appartiennent à la même grappe *k*, alors $h_i = h_j = \lambda r_k$ (λ étant un facteur d'échelle). Pour les cas positifs, nous appliquons la même fenêtre que celle calculée pour les cas témoins de la même grappe, à savoir λr_k.

38 Plusieurs fonctions de densité peuvent être utilisées pour le noyau K. Il est couramment admis que le choix d'une telle fonction est moins important que celui de la taille de la fenêtre. Davies et Hazelton (2010) précisent que les noyaux Gaussiens (utilisant la loi normale) sont fréquemment utilisés pour l'estimation de surfaces à deux dimensions bien que le recours à des noyaux à étendue finie[17] (comme la fonction *biweight*) soit également courant. Si les noyaux à étendue finie ont un avantage en théorie pour des fenêtres adaptatives, en pratique le noyau Gaussien s'avère plus adapté lorsque la répartition des points est très inégale, notamment dans les régions où le nombre d'observations est faible.

39 Nous avons donc retenu le noyau Gaussien en utilisant un facteur d'échelle λ de 0,5 : la fenêtre h_i d'un cas localisé dans la grappe *k* vaut donc $h_i = r_k/2$. Cela induit que 86 % de l'intensité du noyau est située à l'intérieur du cercle de rayon r_k.

Choix d'un indicateur de comparaison

40 Afin de comparer les surfaces de prévalences estimées avec la surface de prévalences du modèle, nous avons utilisé l'indicateur MISD (*Mean Integrated Squared Difference*) (Anderson et Titterington 1997), un indicateur analogue au MISE (*Mean Integrated Squared Error*) (Wand et Jones 1994). Le MISD correspond à l'espérance du carré des différences en chaque point de la carte. Soit deux surfaces spatiales $\hat{s}_1(x,y)$ et $\hat{s}_2(x,y)$. Le MISD entre $\hat{s}_1$ et $\hat{s}_2$ se calcule comme suit[18] :

Équation 3

$$MISD_{\hat{s}_1,\hat{s}_2} = E\int [\hat{s}_1(x,y) - \hat{s}_2(x,y)]^2\,dx\,dy$$

41 Le MISD peut-être approximé à partir d'une grille fine de *p* points régulièrement espacés :

Équation 4

$$MISD_{\hat{s}_1,\hat{s}_2} \approx \frac{1}{p}\sum_{k=1}^{p}[\hat{s}_1(x_k,y_k) - \hat{s}_2(x_k,y_k)]^2$$

42 Le MISD permet ainsi de quantifier les écarts en chaque point entre la surface estimée et celle du modèle. Dès lors, la meilleure estimation sera celle minimisant le MISD.

Logiciels utilisés

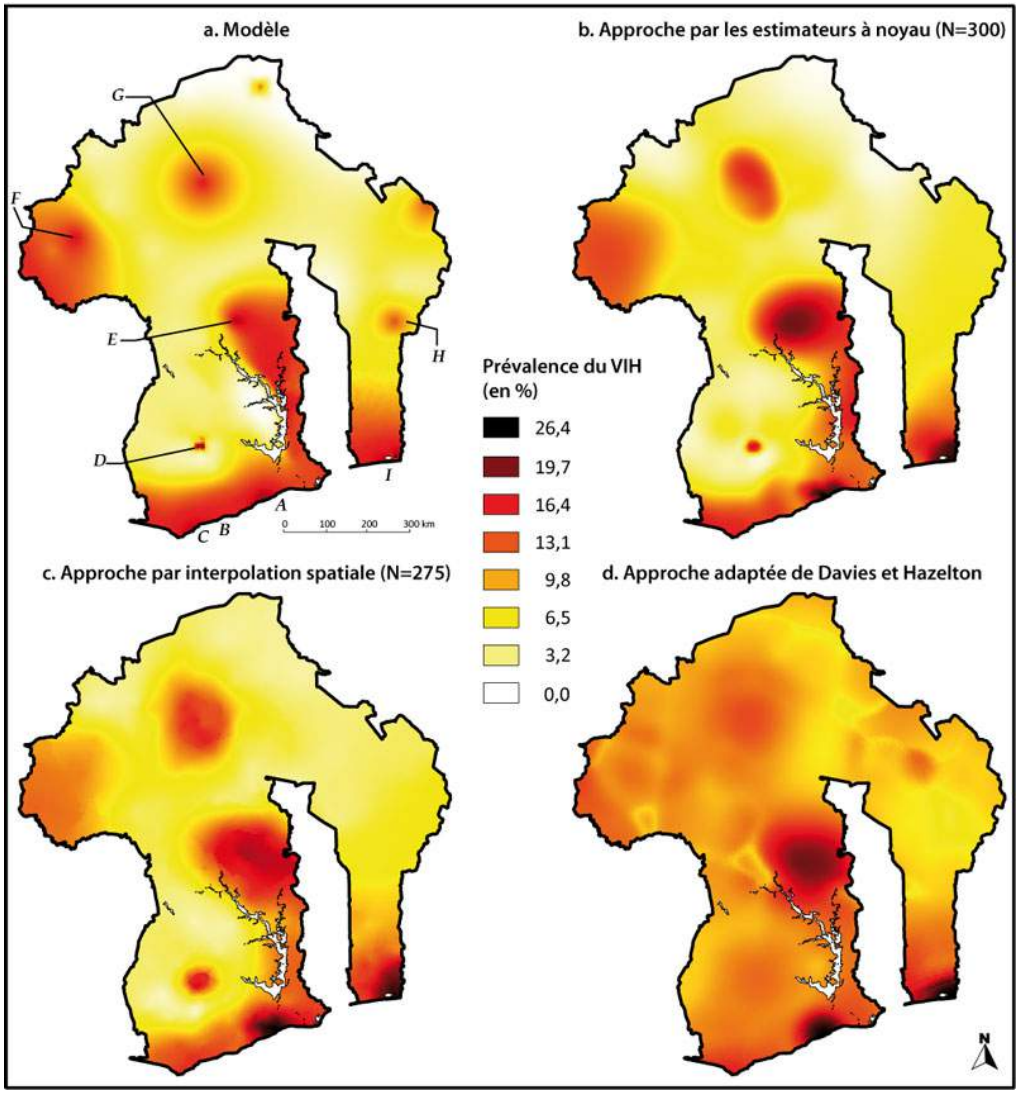
43 Les calculs ont été effectués sous le logiciel de statistique *R* (R Development Core Team 2007). Nous avons utilisé la fonction *krige* du package *gstat* (Pebesma 2004) pour l'interpolation spatiale par krigeage ordinaire. L'approche adaptée de Davies et Hazelton a été mise en œuvre grâce au package *sparr* (Davies, Hazelton et Marshall 2010) développé par les auteurs. Pour l'approche par les estimateurs à noyau, nous avons eu recours à la fonction *KernSur* du package *GenKern* (Lucy et Aykroyd 2010). Enfin, nous avons développé nos propres fonctions, disponibles dans un package nommé *prevR*[19].

44 Les différentes cartes présentes dans cet article ont été dessinées avec le logiciel *Quantum GIS*[20].

Résultats

Comparaison des différentes approches

Figure 6 : Surface des prévalences du modèle et surfaces estimées des prévalences selon trois approches différentes à partir d'une même simulation d'EDS.

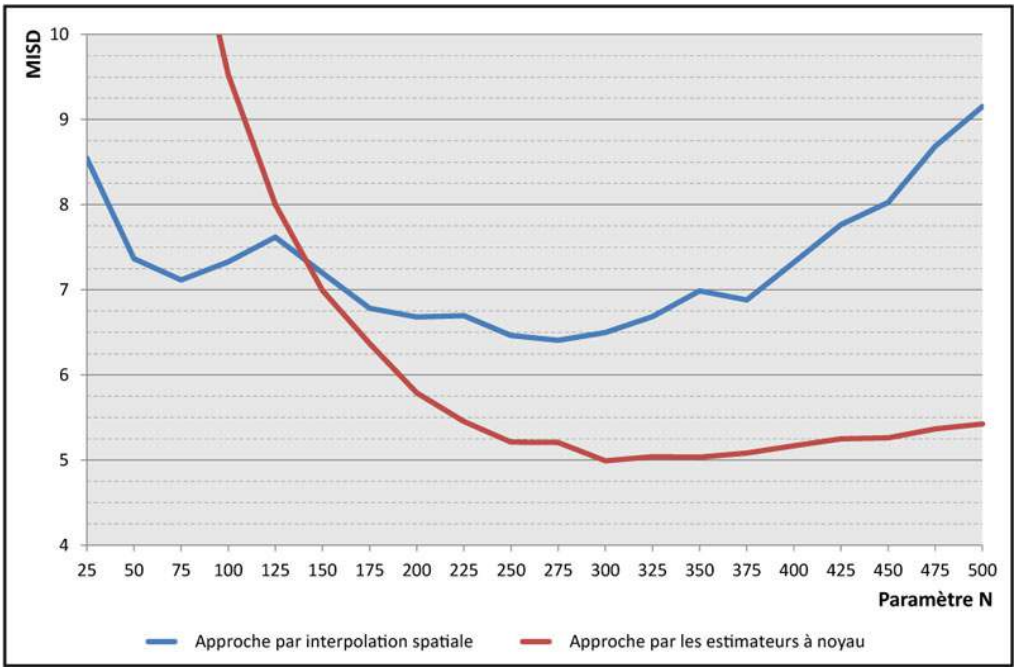


Note : l'échelle colorimétrique est identique à celle de la figure 2.

45 Les figures 6.b, 6.c et 6.d représentent les surfaces des prévalences estimées selon les trois approches. Les figures 8.b, 8.c et 8.d permettent, quant à elles, de visualiser les écarts entre surfaces des prévalences estimées et surface des prévalences du modèle : chaque point de la carte correspond à la différence mathématique entre la prévalence estimée et celle du modèle, soit $\hat{s}_e(x,y) - s_m(x,y)$. Le MISD correspond de fait à la moyenne du carré de ces écarts.

46 L'approche adaptée de Davies et Hazelton (figure 5.d) aboutit à un MISD de 28,1, soit largement plus que les MISD obtenus avec les approches à partir des cercles de même effectif (interpolation spatiale ou estimateurs à noyau, voir figure 7). Les prévalences sont surestimées d'au moins cinq points sur la majeure partie de la surface (figure 8.d). Le lissage produit des « zébrures » sur la surface des prévalences dans les zones faiblement enquêtées (à l'ouest de E, au sud de F, entre G et H…). Les prévalences estimées sont fortement lissées autour de D et entre G et F, ne reproduisant pas les variations de la surface du modèle.

Figure 7 : MISD selon différentes valeurs de N



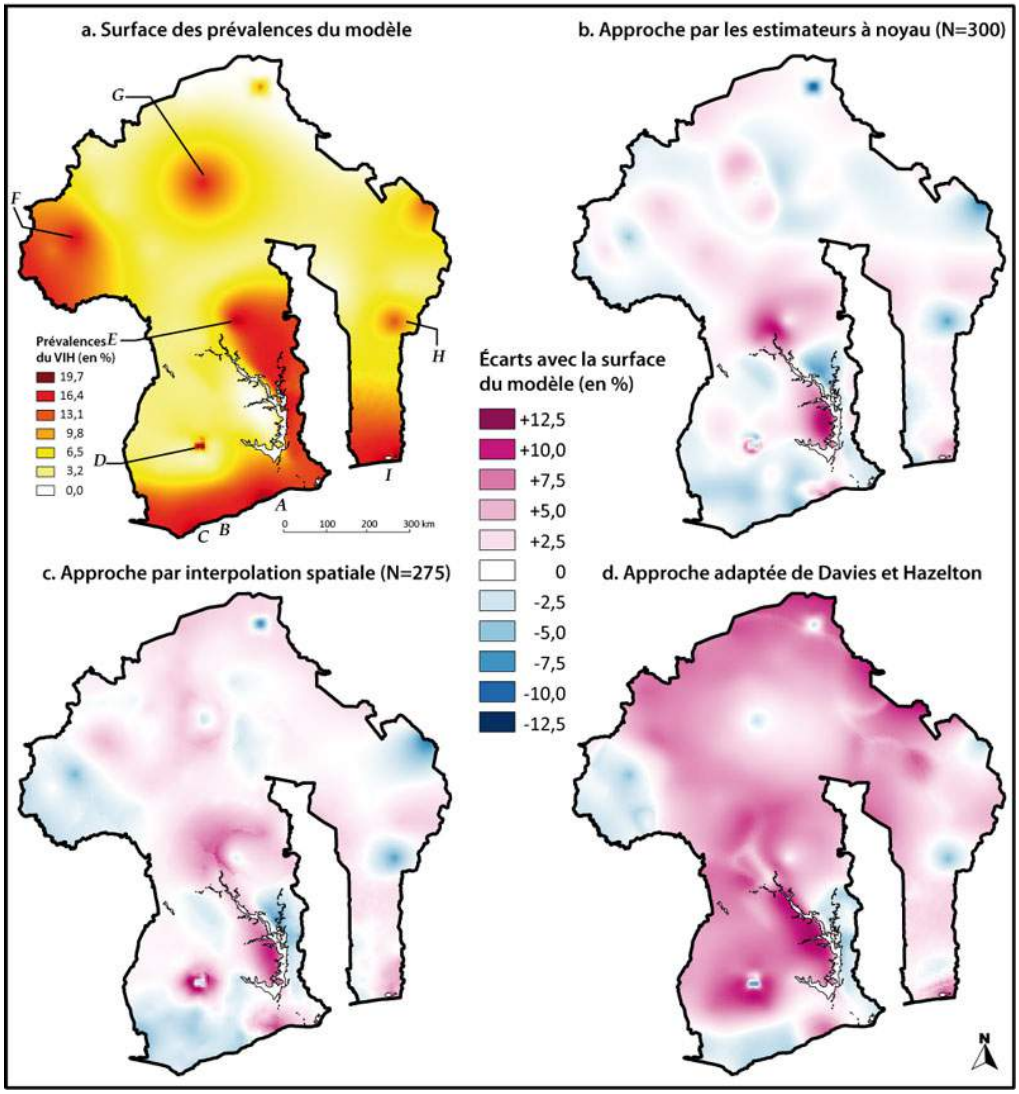
Note : voir tableau annexe 1 pour le détail des valeurs.

47 Les deux approches reposant sur les cercles de même effectif ont un comportement similaire : l'indicateur MISD diminue avec l'augmentation du paramètre *N* jusqu'à atteindre un minima avant que le lissage ne devienne trop important et induise une augmentation du MISD (voir figure 7). Aux petites valeurs de *N*, l'approche par les estimateurs à noyau produit un MISD supérieur à celui de l'approche par interpolation spatiale, mais sa décroissance du MISD s'avère plus rapide. Ainsi, le MISD se minimise à 5,0, correspondant à une valeur de *N* de 300, pour l'approche par les estimateurs à noyau, tandis que le MISD minimum obtenu avec l'approche par interpolation spatiale est de 6,4 avec *N* égal à 275.

48 Les figures 6.b et 6.c présentent les surfaces des prévalences estimées avec ces valeurs optimums de *N*. Globalement, les principales variations de la surface des prévalences du modèle ont été reconstituées. Le gradient de la côte sud vers le nord est reproduit, les contrastes étant accentués du fait d'une surestimation (figures 8.b et 8.c) au niveau des agglomérations A et I. L'agglomération D présente toujours une prévalence concentrée par rapport à son voisinage, cette « concentration » étant mieux rendue par les estimateurs à noyau. De même, ces derniers reproduisent mieux le gradient à la frontière ouest et la diffusion des prévalences autour de l'agglomération G. De part et d'autre du grand lac, où une rupture nette avait été introduite dans le modèle, la prévalence a été surestimée à l'ouest et sous-estimée à l'est, les deux approches ne prenant pas en compte les frontières naturelles.

49 Enfin, les variations situées dans les zones faiblement enquêtées n'ont pu être reproduites : le pic épidémique localisé dans le nord du pays, la diffusion autour de l'agglomération H et depuis la frontière est.

Figure 8 : Surface des prévalences du modèle et écarts (différence mathématique) entre les surfaces estimées selon les trois approches et la surface du modèle

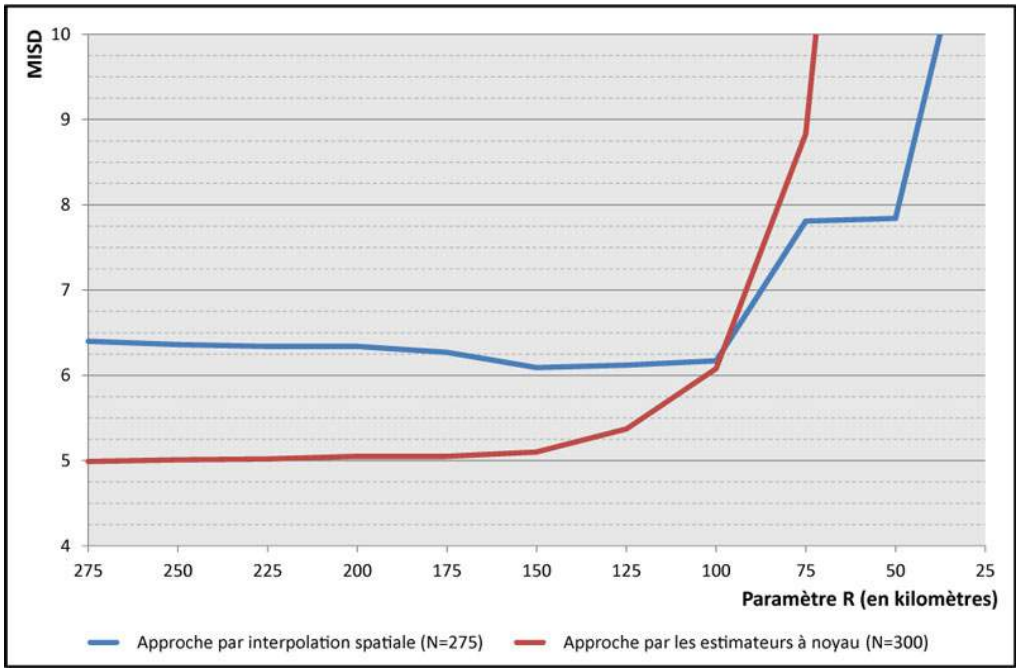


Limitation de la taille des cercles de lissage : ajout d'un paramètre *R*

50 Le recours à un lissage par des cercles de même effectif permet d'estimer les prévalences à partir d'un nombre d'observations suffisantes et induit un lissage spatial variable selon les régions. Ainsi, dans les zones denses en observation, le rayon des cercles de lissage est relativement petit. Par contre, dans les zones peu enquêtées, notamment le long des frontières, le lissage fait intervenir des grappes éloignées les unes des autres et le rayon des cercles de lissage augmente de manière importante (voir Tableau annexe 2).

51 Dans des travaux précédents (Larmarange 2007 ; Larmarange *et al.* 2006), nous avions suggéré la possibilité d'ajouter un second paramètre nommé *R* et correspondant à un rayon maximum des cercles de lissage. Ce second paramètre n'impacte de fait que les grappes situées dans des zones peu enquêtées : si l'effectif *N* n'est pas atteint dans un rayon inférieur à *R*, alors le rayon du cercle de lissage est fixé à *R*.

Figure 9 : MISD pour différentes valeurs du paramètre R (N étant fixé)



Note : la taille maximum des cercles de lissage étant de 269 kilomètres en l'absence du paramètre *R* (voir Tableau annexe 2), les surfaces des prévalences obtenues avec une valeur de *R* de 275 kilomètres sont identiques aux surfaces obtenues en l'absence du paramètre *R*.

52 La Figure 9 présente les MISD obtenus pour différentes valeurs du paramètre *R* (le paramètre *N* étant fixe et correspondant au MISD minimal obtenu en l'absence du paramètre *R*, soit 275 pour l'approche par interpolation spatiale et 300 pour l'approche par les estimateurs à noyau).

53 Si pour l'approche par interpolation spatiale l'ajout du paramètre *R* permet d'obtenir une baisse légère du MISD (minimum obtenu pour *R* égal à 150 kilomètres), il n'y a pas de gain pour l'approche par les estimateurs à noyau. Dans les deux cas, une faible valeur de *R* augmente substantiellement le MISD.

54 Il apparaît donc que l'approche la plus efficace, pour cette simulation donnée, s'avère être l'approche par les estimateurs à noyau à fenêtres adaptatives de même effectif. Pour cette dernière, l'ajout d'un rayon maximum aux cercles de lissage ne permet pas d'améliorer la surface des prévalences estimée.

Discussion

Choix du paramètre *N*

55 L'approche par les estimateurs à noyau à fenêtres adaptatives de même effectif permet de reproduire les principales variations de la surface des prévalences du modèle. La principale difficulté pour appliquer cette méthode à des données réelles consiste à déterminer la valeur adéquate du paramètre *N* à utiliser. Dans le cadre des simulations d'EDS, il a été possible de calculer un MISD dans la mesure où la surface des prévalences à estimer était connue. Or, pour une application à des données réelles, cette surface des prévalences est inconnue et il n'est donc pas possible de calculer un MISD.

56 Altman (1992) suggère que l'une des possibilités pour déterminer la valeur dd'un paramètre de lissage consiste à réaliser plusieurs estimations avec plusieurs valeurs puis à sélectionner subjectivement celle répondant le mieux à ce qui est attendu. Si l'on souhaite mettre en évidence les tendances principales du phénomène, alors une valeur élevée du paramètre de lissage sera pertinente. À l'inverse, si l'on souhaite explorer les extrema locaux, une valeur faible du paramètre sera à privilégier. Un choix subjectif du paramètre de lissage offre une grande flexibilité et un regard compréhensif sur les données. Cependant, il rappelle qu'une méthode objective de sélection du paramètre de lissage peut être préférable pour produire une technique de lissage automatique ou pour une meilleure consistance des résultats entre différents investigateurs.

57 L'approche adaptée de Davies et Hazelton a l'avantage de proposer une sélection automatique de la taille des fenêtres à partir des données disponibles. Cependant, la surface des prévalences produite s'avère, dans le cas présent, peu satisfaisante avec un MISD relativement élevé. Cela tient notamment au fait que l'approche de Davies et Hazelton a été développée pour des situations où les distributions des deux semis de points (cas positifs et cas témoins) étaient indépendantes, ce qui n'est pas le cas des données des EDS.

58 Pour les approches à partir des cercles de même effectif, il apparaît que le MISD varie relativement peu autour de son minimum (voir Figure 7). Il est donc raisonnable de considérer que les surfaces obtenues, avec des valeurs de *N* situées dans un intervalle de plus ou moins 50 autour de la valeur de *N* minimisant le MISD, sont acceptables.

59 Dans des travaux précédents (Larmarange 2007 ; Larmarange *et al*. 2006), nous avons essayé de modéliser la valeur optimale de *N* (notée N_o) en fonction de la prévalence nationale observée (*p*), du nombre de personnes testées (*n)* et du nombre de grappes enquêtées (*g*), soient les trois paramètres utilisés pour simuler une EDS. Pour cela, nous avons simulé 22 000 EDS avec différentes valeurs de ces trois paramètres et calculé pour chaque simulation la valeur optimale de *N*. Pour des raisons de temps et de puissance informatique de calcul, le critère de détermination de la valeur optimale de *N* pour une simulation donnée n'était pas la minimisation du MISD calculé sur l'ensemble de la surface des prévalences. L'indicateur retenu a été la minimisation de la moyenne des écarts absolus, entre prévalence estimée et prévalence du modèle, calculés sur les seules grappes enquêtées. Par rapport au MISD calculé sur l'ensemble de la surface des prévalences, cet indicateur accorde un poids plus important aux régions plus densément peuplées (dans la mesure où elles concentrent plus de grappes enquêtées). L'utilisation des écarts absolus majore les grappes pour lesquelles les écarts sont faibles par rapport aux carrés des écarts qui majorent les écarts importants. De fait, les valeurs optimales de *N* calculées avec cet indicateur s'avèrent plus faibles que les valeurs optimales calculées par minimisation du MISD. La modélisation des résultats obtenus a produit la relation suivante :

Équation 5

$$N_o = 2{,}688 \bullet n^{0{,}419} \bullet p^{-0{,}361} \bullet g^{0{,}037} - 91{,}011$$

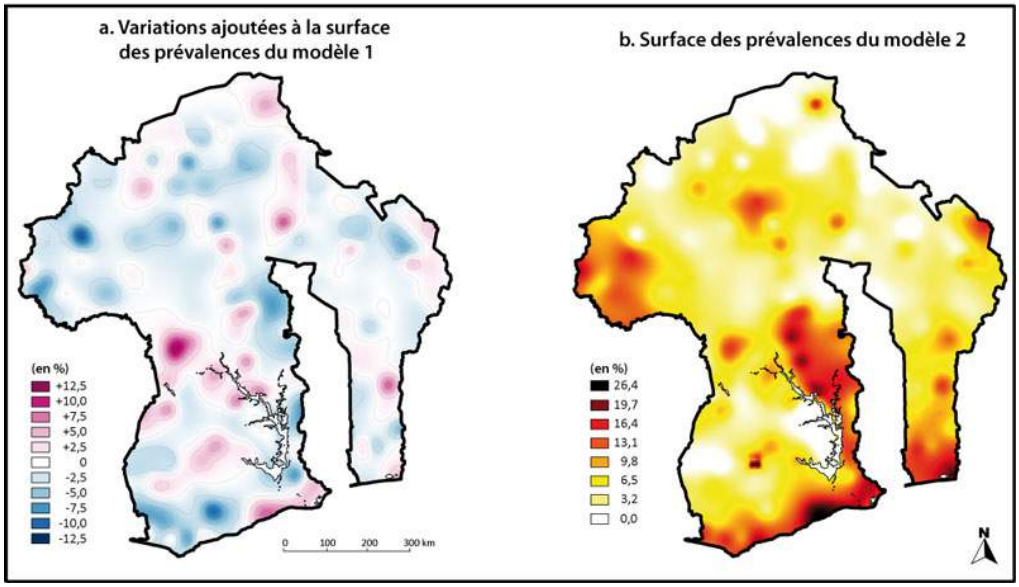
60 Ce résultat n'a qu'une simple valeur indicative. En effet, il est dépendant de la surface des prévalences imposée au modèle. D'autres surfaces des prévalences auraient produit d'autres valeurs optimales. Néanmoins, cette équation peut être utilisée afin de guider le choix du paramètre *N* dans le cadre d'une application à des données réelles, en fournissant un ordre de grandeur.

Surfaces des prévalences estimées et surfaces de tendances

61 La surface des prévalences du modèle est fortement organisée. Elle a été construite de manière *ad hoc* afin d'être multi-polarisée et de présenter des gradients continus et réguliers. Dès lors, il est logique que cette structure puisse être reproduite dans ses principales lignes à partir d'un échantillon suffisant.

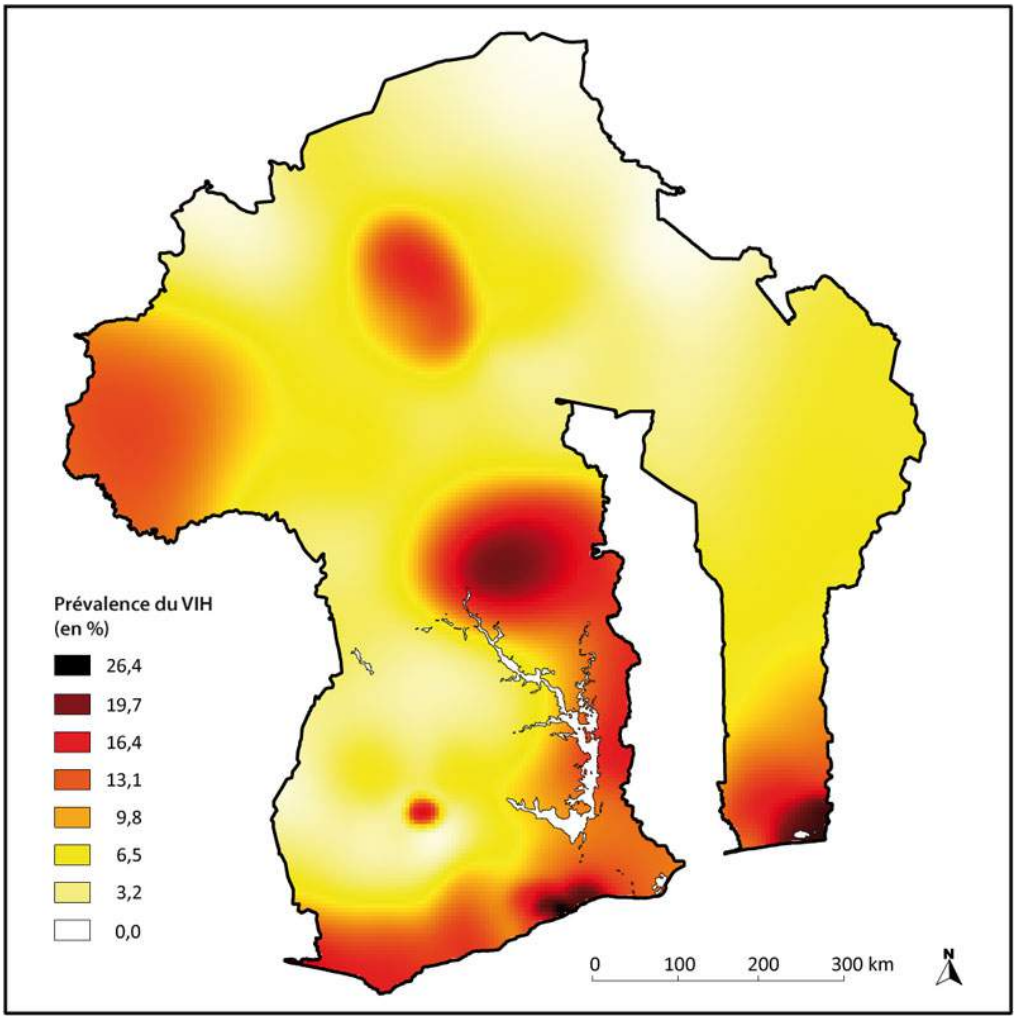
62 Afin de tester les résultats obtenus à partir d'une surface initiale de prévalences moins régulière, nous avons élaboré un second modèle en ajoutant à la surface des prévalences du premier modèle des variations aléatoires localisées. Une surface de variations (figure 10.a) a été générée à partir de 800 points sélectionnés aléatoirement, ayant chacun une fenêtre d'action définie aléatoirement et une contribution positive ou négative déterminée aléatoirement. Une correction a été effectuée afin que la surface des prévalences du second modèle[21] (figure 10.b) ne présente pas de prévalences négatives et que la prévalence nationale (tenant compte de la densité de population) soit toujours égale à 10 %. Cette nouvelle surface présente de fait de nombreuses irrégularités tout en ayant une structure spatiale sous-jacente (celle du premier modèle).

Figure 10 : Variations aléatoires ajoutées au premier modèle et surface des prévalences du second modèle



63 Une nouvelle EDS a été simulée à partir de ce second modèle[22]. La figure 11 présente la surface des prévalences estimée en utilisant l'approche par les estimateurs à noyau et une valeur de *N* égale à 300. Cette surface est relativement similaire à celle obtenue en appliquant la même méthode à une simulation d'EDS effectuée à partir du premier modèle (figure 6.b). Il apparaît donc que cette méthode de lissage filtre les variations locales pour mettre en évidence les variations régionales sous-jacentes.

Figure 11 : Surface des prévalences estimée selon l'approche par les estimateurs à noyau (N=300) à partir d'une simulation d'EDS effectuée sur le second modèle



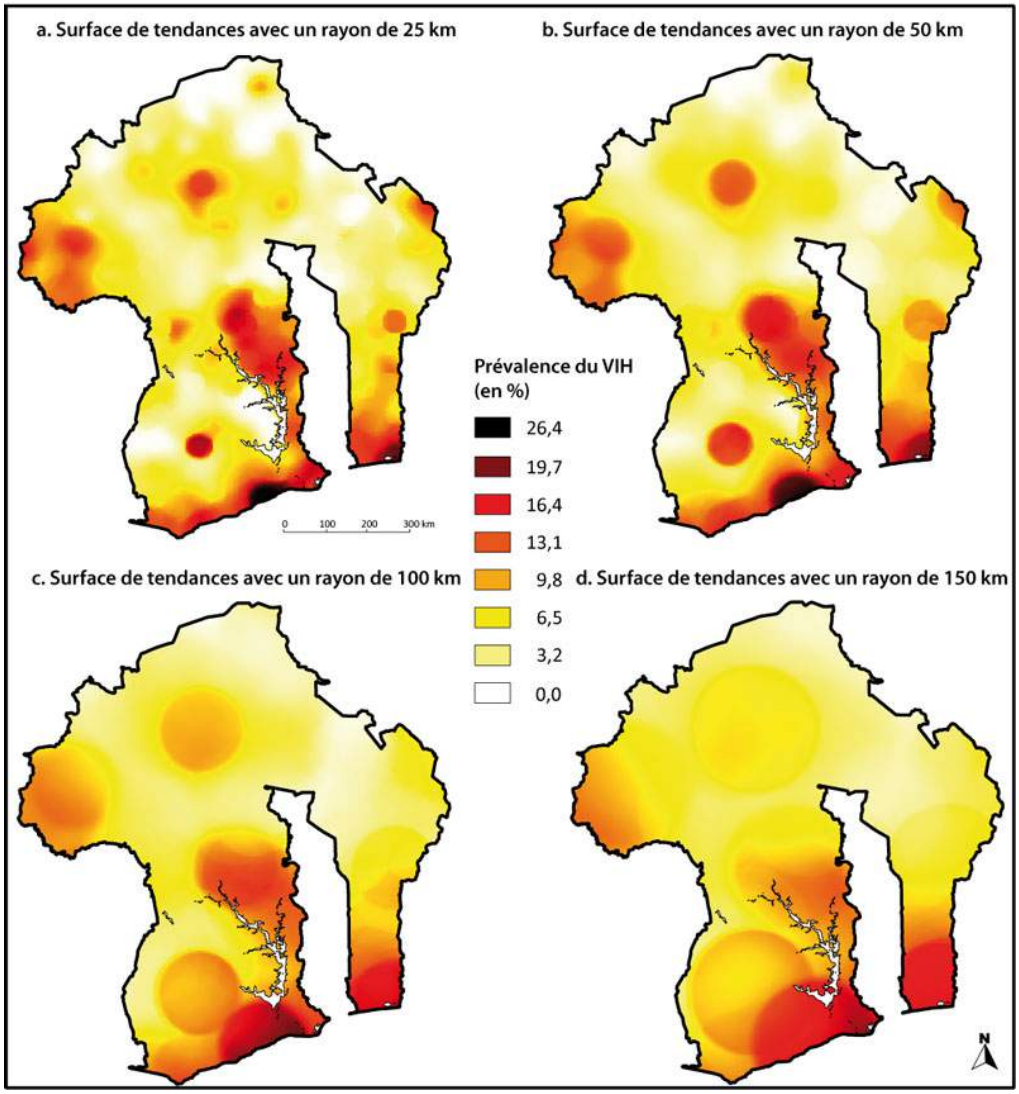
64 Ce résultat évoque les analyses en composantes d'échelle (Chorley et Haggett 1965 ; Griffin 1949 ; Haggett 1973 ; Krumbein 1956) qui se sont développées à partir du milieu du XX^{e} siècle. Ces techniques de filtrage cartographique s'appliquent à des surfaces connues et visent à décomposer la surface du phénomène étudié comme étant la somme d'une surface de tendances régionales et d'une surface de résidus locaux. La surface des tendances régionales est calculée de manière à filtrer les détails locaux et à mettre en évidence les principales variations du phénomène. La surface des résidus locaux est simplement la différence entre la surface des données brutes et celle des tendances régionales.

65 Une des méthodes de filtrage cartographique pour le calcul des tendances régionales est la méthode dite des « cercles » (Griffin 1949 ; Krumbein 1956 ; Nettleton 1954) qui s'apparente à une moyenne mobile spatiale. Elle consiste, pour chaque point de la grille, à tracer un cercle de rayon fixe autour de ce point puis à calculer l'indicateur au sein de la surface définie par le dit cercle.

66 La figure 12 présente différentes surfaces de tendance calculées à partir de la surface des prévalences du second modèle en utilisant la méthode des cercles avec des rayons de 25 à 150 kilomètres. Elles font apparaître progressivement la structure spatiale sous-jacente du modèle.

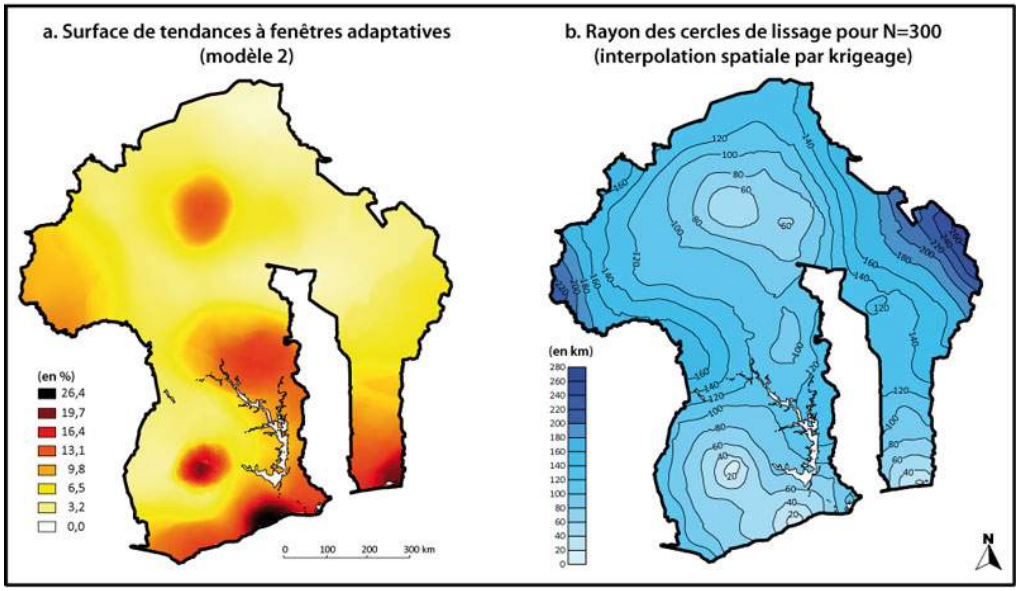
67 La surface des prévalences estimées à partir d'une simulation d'EDS (figure 11) s'avère plus proche de ces surfaces de tendances que de la surface initiale des prévalences du modèle. En effet, le MISD entre la surface estimée et celle du second modèle est de 10,3 alors qu'il est de 9,1 (respectivement 7,5 6,9 et 10,6) entre la surface estimée et la surface de tendances utilisant un rayon de 25 kilomètres (respectivement 50, 100 et 150 kilomètres).

Figure 12 : Surfaces de tendances calculées à partir du second modèle en utilisant des cercles de 25, 50, 100 et 150 kilomètres de rayon



68 Les cercles de lissages utilisés pour définir la valeur des fenêtres de l'approche par les estimateurs à noyau sont similaires à ceux de la méthode des « cercles » mais utilisent des fenêtres adaptatives de même effectif et non de rayon fixe. Cette notion de fenêtre adaptative peut être appliquée au calcul d'une surface de tendances. Il importe pour cela de déterminer le rayon des fenêtres à utiliser en chaque point de la grille. À partir du semis des grappes enquêtées dans la simulation d'EDS, pour lesquelles un rayon de lissage a été déterminé (pour une valeur de *N* donné), une surface de rayons est générée en procédant à une interpolation spatiale (figure 13.b). Une surface de tendances à fenêtres adaptatives (figure 13.a) est alors calculée en utilisant pour chaque point de la grille une fenêtre définie par cette surface de rayons. Le résultat obtenu présente visuellement des similarités avec la surface des prévalences estimées selon l'approche par les estimateurs à noyau (figure 11). Le MISD entre ces deux surfaces n'est que de 6,0.

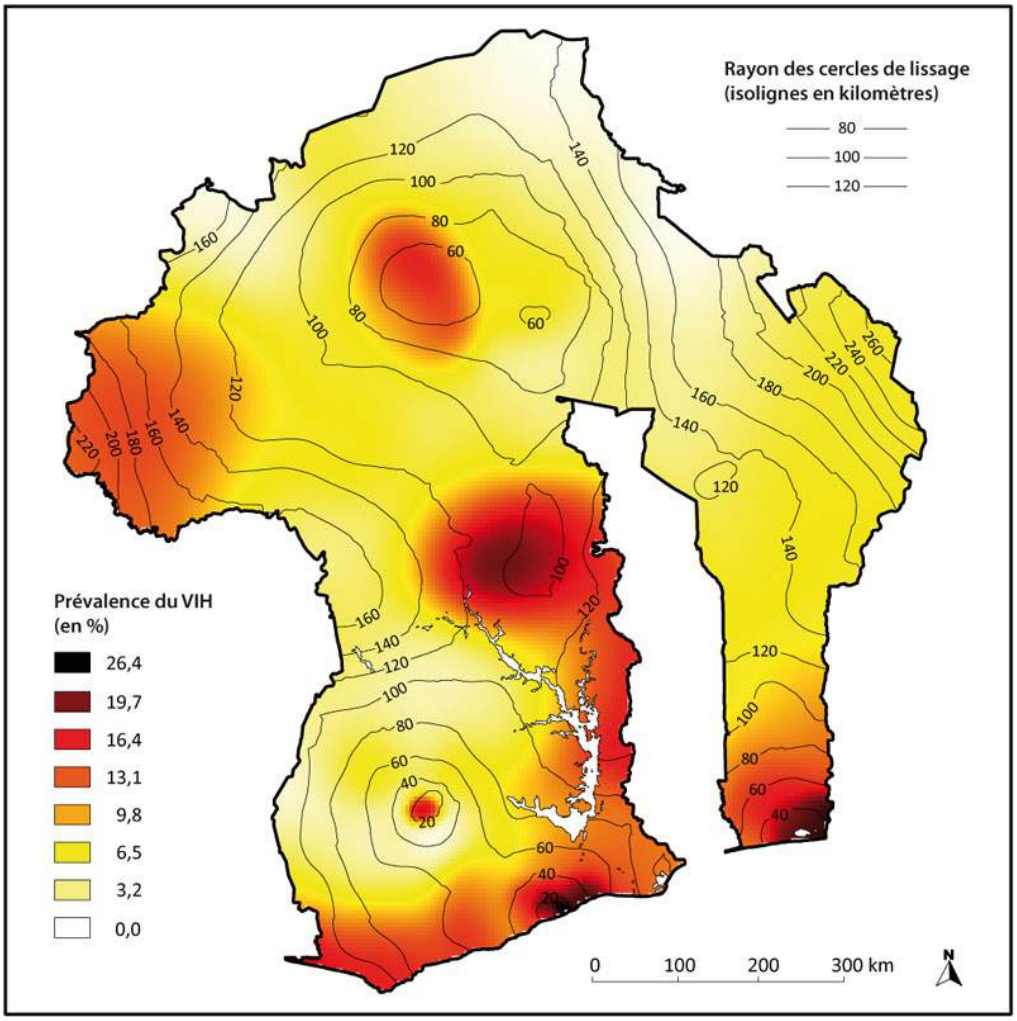
Figure 13 : Surface de tendances utilisant des fenêtres adaptatives et interpolation spatiale du rayon des cercles de lissage (N=300)



69 L'approche par les estimateurs à noyau à fenêtres adaptatives de même effectif ne permet pas de reproduire les variations locales de la surface des prévalences du modèle. Une perte d'information est inévitable du fait de l'échantillonnage des EDS. Cette approche ne permet pas de reproduire la surface réelle des prévalences.

70 Cependant, la surface des prévalences estimées traduit bien une certaine réalité des épidémies et s'apparente à une surface de tendances à fenêtres adaptatives en mettant à jour les variations régionales sous-jacentes du phénomène. Il importe dès lors de prendre en considération les variations du rayon des cercles de lissage pour interpréter la surface estimée des prévalences. Afin de faciliter la lecture des cartes, il est possible de superposer les deux informations en affichant les isolignes[23] des rayons des cercles de lissage sur la surface des prévalences (figure 14).

Figure 14 : Surface des prévalences estimée selon l'approche par les estimateurs à noyau (*N*=300) à partir d'une simulation d'EDS effectuée sur le second modèle et isolignes des rayons des cercles de lissages

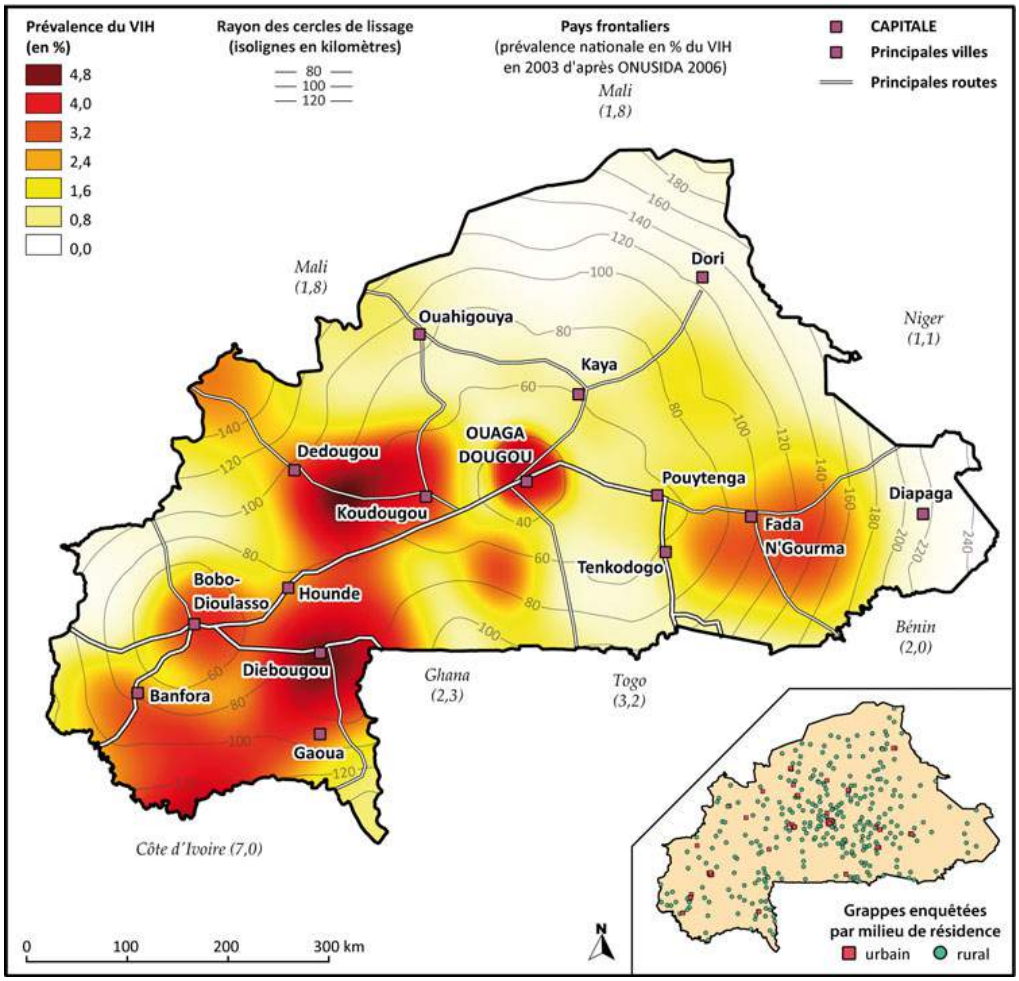


Application à des données réelles : l'EDS 2003 du Burkina Faso

71 Nous avons appliqué l'approche par les estimateurs à noyau avec fenêtres adaptatives de même effectif aux données de l'EDS 2003 du Burkina Faso. Cette enquête a testé 7 244 personnes (15-49 ans) réparties en 400 grappes. La prévalence nationale du VIH mesurée dans l'enquête est de 1,8 % (voir Tableau 1).

72 La distribution des grappes enquêtées est représentée dans un encart sur la figure 15. La carte principale présente la surface des prévalences estimées avec un paramètre *N* de 500, valeur choisie d'après l'équation 5[24]. Sont mentionnés les pays frontaliers du Burkina Faso ainsi que la prévalence nationale du VIH à fin 2003, de chacun d'eux, selon les estimations de l'ONUSIDA (2006). Les isolignes des rayons des cercles de lissage ont été ajoutés ainsi que les principales routes et agglomérations urbaines du Burkina Faso.

Figure 15 : Tendances régionales de la prévalence du VIH (15-49 ans), estimées avec l'approche par les estimateurs à noyau (N=500), sur les données de l'EDS 2003 du Burkina Faso



73 La surface des prévalences estimée à partir de l'EDS s'avère relativement cohérente. Tout d'abord l'épidémie est plus importante dans le sud-ouest du pays, bordant des pays à forte prévalence (Côte d'Ivoire, Ghana), que dans le nord-est sahélien, peu peuplé et bordant des pays à faible prévalence (Mali, Niger). Les zones les plus touchées se situent en majorité autour des principales agglomérations (Ougadougou, Bobo-Dioulasso) et le long des principaux axes routiers vers la Côte d'Ivoire et le Mali (via Dedougou). La région entre Diebougou et Gaoua, présentant un pic épidémique, est connue comme une zone importante d'orpaillage, impliquant une migration saisonnière masculine relativement importante et un commerce du sexe non négligeable. Enfin, les régions du sud-ouest sont également celles ayant connu les plus forts taux de rapatriés de Côte d'Ivoire fin 2002, début 2003 (SP/CONASUR 2004).

74 Le rapprochement effectué entre zones à prévalence élevée, zones migratoires, zone d'orpaillage et principaux centres urbains ne peut permettre d'établir un lien entre ces différents phénomènes, la simple proximité géographique n'ayant pas valeur de preuve. Cependant, ces résultats sont cohérents avec les travaux de Georges Rémy (1999) en Afrique subsaharienne montrant que "*les villes, notamment les plus grandes, sont spécialement exposées à l'infection à toutes les étapes de sa dynamique… Mais des sites ruraux sont également vulnérables. Ils se distinguent par leur participation à des activités économiques variées, à caractère monétaire : centres miniers, étapes routières, périmètres agro-industriels, marchés.*"

Conclusion

75 L'échantillonnage des enquêtes démographiques et de santé est élaboré pour permettre une certaine précision du calcul des prévalences du VIH au niveau national et régional. Les effectifs

s'avèrent cependant trop faibles pour une estimation précise des prévalences au niveau de chaque grappe enquêtée ou à un niveau local fin.

76 L'utilisation de fenêtres adaptatives de même effectif permet d'effectuer un lissage s'adaptant à la grande irrégularité de la distribution spatiale des grappes enquêtées, ces dernières étant sélectionnées en fonction de la répartition de la population. Les cartes générées sont ainsi relativement précises dans les zones densément peuplées tout en étant plus fortement lissées dans les régions faiblement enquêtées.

77 La simulation d'EDS à partir d'un pays fictif a mis en évidence qu'une approche par les estimateurs à noyau à fenêtres adaptives de même effectif s'avérait plus efficace, dans le cas présent, à une interpolation spatiale de prévalences préalablement lissées à partir de ces mêmes cercles de lissages. De même, il n'y a pas de gain significatif à l'ajout d'un rayon maximum comme second paramètre de lissage.

78 Le second modèle utilisé a montré que, si les variations locales de l'épidémie étaient filtrées par ce type de techniques, la composante régionale des variations spatiales des prévalences était globalement reconstituée, les surfaces estimées des prévalences pouvant alors s'interpréter comme des surfaces de tendances régionales à fenêtres adaptatives. Une telle surface, par construction, est forcément spatialement continue et auto-corrélée et ne présume en rien des discontinuités et des variations locales éventuelles de la surface réelle de l'épidémie qui reste inaccessible à partir des données EDS.

79 Cette approche a pour inconvénient de ne pas fournir de technique automatisée pour sélectionner une valeur optimale du paramètre de lissage. Si des travaux pointus de recherche sont menés dans ce domaine, il en existe relativement peu concernant le ratio de deux surfaces de densité. Nous avons testé une de ces approches, adaptée des travaux de Davies et Hazelton (2010). Elle s'est avérée peu efficace dans le cas présent. Les données EDS sont particulières dans le sens où la localisation des données observées résulte d'un échantillonnage en grappes à deux degrés et non d'un échantillonnage aléatoire simple, le semis de points des cas positifs n'étant de fait pas indépendant du semis de points des cas observés.

80 Néanmoins, l'équation 5, bien que déterminée à partir de simulations effectuées sur le pays fictif et ne pouvant en toute rigueur être généralisée à d'autres situations, fournit un ordre de grandeur du paramètre utilisable en pratique. L'application effectuée sur les données de l'EDS 2003 du Burkina Faso a permis de produire une surface des tendances régionales de la prévalence plausible.

81 Si l'interprétation d'une telle carte doit être prudente, elle fournit cependant une indication descriptive sur la situation épidémique au sein d'un pays, indépendante du découpage administratif du territoire. Il s'agit bien là d'un outil de visualisation des principales variations spatiales du phénomène et d'identification des régions les plus touchées. Bien que les EDS soient insuffisantes pour mener une analyse des déterminants spatiaux des épidémies de VIH, elles permettent d'en esquisser un premier portrait, en l'absence d'enquêtes plus spécifiques ayant une meilleure couverture géographique.

Ce travail a bénéficié du support financier de l'Agence Nationale de Recherche sur le VIH/ Sida et les hépatites virales (projet ANRS 12114).

Lecture : pour 60 % des grappes, le rayon du cercle de lissage est inférieur ou égal à 87 kilomètres.

Bibliographie

Altman N.S., 1992, "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression", *The American Statistician*, vol. 46, No. 3, 175-185.

Anderson N.H., Titterington D.M., 1997, "Some Methods for Investigating Spatial Clustering, with Epidemiological Applications", *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, vol. 160, No. 1, 87-105.

Baillargeon S., 2005, *Le krigeage : revue de la théorie et application à l'interpolation spatiale de données de précipitations,* mémoire présenté pour l'obtention du grade de Maître ès Sciences (M.Sc.), Université de Laval, Faculté des Sciences et de Génie, Québec, disponible en ligne à http://www.theses.ulaval.ca/2005/22636/22636.pdf.

Bithell J.F., 1990, “An application of density estimation to geographical epidemiology”, *Statistics in Medicine*, vol. 9, No. 6, 691-701. doi:10.1002/sim.4780090616

Boerma J.T., Ghys P.D., Walker N., 2003, “Estimates of HIV-1 prevalence from national population-based surveys as a new gold standard”, *Lancet*, vol. 362, No 9399, 1929-31. doi: 10.1016/ S0140-6736(03)14967-7

Carlos H.A., Shi X., Sargent J., Tanski S., Berke E.M., 2010, “Density estimation and adaptive bandwidths: A primer for public health practitioners”, *International Journal of Health Geographics*, vol. 9, No. 1, 39. doi:10.1186/1476-072X-9-39

Center for International Earth Science Information Network (CIESIN) of Columbia University, 2005a, *Global Rural-Urban Mapping Project (GRUMP), Alpha Version: Population Density Grids*, disponible en ligne à http://sedac.ciesin.columbia.edu/gpw.

Center for International Earth Science Information Network (CIESIN) of Columbia University, 2005b, *Global Rural-Urban Mapping Project (GRUMP), Alpha Version: Urban Extents*, disponible en ligne à http://sedac.ciesin.columbia.edu/gpw.

Chorley R.J., Haggett P., 1965, “Trend-Surface Mapping in Geographical Research”, *Transactions of the Institute of British Geographers*, No. 37, 47-67. doi:10.2307/621689

Davies T.M., Hazelton M.L., 2010, “Adaptive kernel estimation of spatial relative risk”, *Statistics in Medicine*, vol. 29, No. 23, 2423-2437. doi:10.1002/sim.3995

Davies T.M., Hazelton M.L., Marshall J.C., 2010, “sparr: Analyzing spatial relative risk using fixed and adaptive kernel density estimation in R”. *Journal of Statistical Software*, en cours d’impression.

Diggle P., Rowlingson B., Su T., 2005, “Point process methodology for on-line spatio-temporal disease surveillance”, *Environmetrics*, vol. 16, No. 5, 423-434. doi:10.1002/env.712

Gatrell A.C., Bailey T.C., Diggle P.J., Rowlingson B.S., 1996, “Spatial Point Pattern Analysis and Its Application in Geographical Epidemiology”, *Transactions of the Institute of British Geographers*, New Series, vol. 21, No. 1, 256-274.

Griffin W.R., 1949, “Residual Gravity in Theory and Practice”, *Geophysics*, vol. 14, No. 1, 39-56.

Haggett P., 1973, *L’Analyse spatiale en géographie humaine*, Paris, Collection U, Armand Colin.

Kelsall J.E., Diggle P.J., 1995, “Kernel Estimation of Relative Risk”, *Bernoulli*, vol. 1, No. ½, 3-16.

Krige D., 1951, “A Statistical Approach to Some Basic Mine Valuation Problems on the Witwatersrand”, *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, vol. 52, No. 6, 119-139.

Krumbein W.C., 1956, “Regional and local components in facies maps”, *AAPG Bulletin*, vol. 40, No. 9, 2163-2194.

Larmarange J., 2007, *Prévalences du VIH en Afrique : validité d’une mesure,* thèse de doctorat en démographie, Université Paris Descartes, disponible en ligne à http://tel.archives-ouvertes.fr/ tel-00320283/fr/.

Larmarange J., 2009, “Prévalences du VIH en Afrique sub-saharienne : Historique d’une estimation”, *Médecine Sciences: M/S*, vol. 25, No. 1, 87-92.

Larmarange J., Yaro S., Vallo R., Msellati P., Méda N., Ferry B., 2006, “Cartographier les données des Enquêtes Démographiques et de Santé à partir des coordonnées des zones d’enquête”, *Chaire Quételet 2006*, Louvain-la-Neuve, disponible en ligne à http://www.uclouvain.be/cps/ucl/doc/demo/documents/ Larmarange_et_al_light.pdf.

Lucy D., Aykroyd R., 2010, *GenKern: Functions for generating and manipulating binned kernel density estimates*, disponible en ligne à http://CRAN.R-project.org/package=GenKern.

Matheron G., 1963, *Traite´de géostatistique appliqueé, Tome II : le krigeage*, Mémoires du Bureau de recherches géologiques et minières, Paris, Editions Technip.

Nettleton L.L., 1954, “Regionals, residuals and structures”, *Geophysics*, vol. 19, No. 1, 1-22. doi:10.1190/1.1427966

ONUSIDA, 2006, *Rapport 2006 sur l’épidémie mondiale de SIDA,* No. ONUSIDA/06.20F, Genève, ONUSIDA, disponible en ligne à http://www.unaids.org/fr/KnowledgeCentre/HIVData/ GlobalReport/2006/default.asp.

Pebesma E.J., 2004, “Multivariable geostatistics in S: the gstat package”, *Computers & Geosciences*, vol. 30, 683-691.

R Development Core Team, 2007, *R: A language and environment for statistical computing*, Vienne, R Foundation for Statistical Computing, disponible en ligne à http://www.R-project.org.

Rémy G., 1999, “L’Infection à VIH1 en Afrique subsaharienne : la priorité urbaine reconsidérée”, *Médecine d’Afrique Noire*, vol. 46, No. 8-9, 388-393.

Sain S.R., 1994, *Adaptive Kernel density Estimation,* thèse de doctorat, Houston, Texas, Rice University.

Sain S.R., 2002, “Multivariate locally adaptive density estimation”, *Computational Statistics & Data Analysis*, vol. 39, No. 2, 165-186. doi:10.1016/S0167-9473(01)00053-6

Silverman B., 1986, *Density estimation for statistics and data analysis*, Monographs on statistics and applied probability, London, Chapman and Hall.

SP/CONASUR, 2004, *Analyse des données sur les rapatriés de Côte d’Ivoire,* Ouagadougou, Comité National de Secours d’Urgence et de Réhabilitation.

TACAIDS, 2006, *Tanzania Atlas of HIV/AIDS Indicators 2003-2004*, Dar es Salaam, TACAIDS, NBS, NACP, ORC Macro, disponible en ligne à http://www.measuredhs.com/pubs/pdf/GS5/GS5.pdf.

Terrell G.R., 1990, “The Maximal Smoothing Principle in Density Estimation”, *Journal of the American Statistical Association*, vol. 85, No. 410, 470-477.

Wand M.P., Jones M.C., 1994, *Kernel Smoothing,* Monographs on statistics and applied probability, London, Chapman & Hall/CRC.

Annexe

Tableau annexe 1 : MISD selon différentes valeurs de N pour l’approche par interpolation spatiale et l’approche selon les estimateurs à noyau

N	Interpolation spatiale	Estimateurs à noyau
25	8,54	35,50
50	7,37	18,85
75	7,12	11,92
100	7,33	9,52
125	7,62	8,00
150	7,20	6,99
175	6,78	6,37
200	6,68	5,79
225	6,70	5,45
250	6,46	5,21
275	6,41	5,21
300	6,50	4,99
325	6,68	5,04
350	6,99	5,04
375	6,88	5,08
400	7,32	5,17
425	7,77	5,25
450	8,03	5,26
475	8,69	5,37
500	9,15	5,42

Tableau annexe 2 : Quantiles des rayons des cercles de lissage pour N=300

Quantile	50 %	55 %	60 %	65 %	70 %	75 %	80 %	85 %	90 %	95 %	Max
Valeur (km)	74	80	87	93	102	108	114	129	138	159	269

Notes

1 En épidémiologie, la prévalence d’une pathologie peut être exprimée à la fois de manière absolue (nombre de cas) ou relative (proportion de personnes infectées parmi la population d’étude). Dans la suite de cet article, nous utiliserons systématiquement le terme de prévalence dans son acception relative.

2 Programme commun des Nations unies sur le VIH/Sida.

3 Grandes régions administratives d'un pays.

4 La surveillance sentinelle des femmes enceintes consiste à sélectionner certaines cliniques prénatales réparties sur le territoire national puis à prélever et tester un échantillon sanguin pour chaque femme se présentant à sa première visite prénatale. Ces enquêtes, relativement peu coûteuses et aisées à mettre en œuvre, sont effectuées annuellement dans nombre de pays depuis la fin des années 1980 et le début des années 1990.

5 Sont exclues des ménages dit « ordinaires » les personnes vivant en institution (prisons, hôpitaux, casernes, internat, couvents…). La définition des ménages ordinaires peut varier selon le recensement.

6 http://www.measuredhs.com/aboutsurveys/gis/methodology.cfm, page consultée le 15 septembre 2010.

7 http://www.measuredhs.com/aboutsurveys/ais/start.cfm, page consultée le 15 septembre 2010.

8 http://www.measuredhs.com/pubs/articles/start.cfm?selected=2, page consultée le 15 septembre 2010.

9 http://www.hivmapper.com/

10 HIV Spatial Data Repository : http://www.hivspatialdata.net/

11 Cette technique est explicitée plus loin.

12 Plus précisément, la prévalence interpolée au centroïde de chaque unité primaire a été appliquée à l'ensemble de l'unité primaire.

13 En raison d'arrondis dans le calcul du nombre de grappes à tirer par strate, il peut arriver que le total de grappes sélectionnées diffère légèrement (± 1) du nombre visé.

14 Par ailleurs, le nombre de personnes testées par grappe n'est pas auto-corrélé spatialement (Larmarange, 2007).

15 En tenant compte du poids relatif de chaque individu.

16 Le principe maximal de lissage de Terrell, dans le domaine des noyaux à fenêtre fixe, consiste à utiliser la fenêtre la plus grande parmi une famille de fenêtres optimales estimée à partir de la variance observée de l'échantillon.

17 Le noyau Gaussien produit une surface de densité s'étendant sur l'ensemble de la surface ($\forall (x,y)$, $K(d_i/h_i) > 0$) tandis que les noyaux à étendue finie ont une densité nulle au-delà de la fenêtre (di > hi $\Rightarrow K(d_i/h_i) = 0$).

18 Cette équation ne diffère de celle du MISE que par le fait que l'on compare deux surfaces connues au lieu d'une surface estimée avec une surface de densité inconnue.

19 En ligne à l'adresse http://www.ceped.org/prevR.

20 Version 1.5.0-Tethys, http://www.qgis.org.

21 Obtenue par addition de la surface des prévalences du premier modèle et de la surface des variations.

22 Avec les mêmes paramètres de simulation : prévalence nationale de 10 %, 8000 personnes enquêtées réparties en 400 grappes.

23 Isolignes calculées à partir de la surface des rayons obtenue par interpolation spatiale des rayons des cercles de lissage des grappes enquêtées.

24 L'Error: Reference source not foundéquation 5 appliquée aux données de l'EDS 2003 fournit une valeur optimale de N de 502 que nous avons arrondie à 500.

Pour citer cet article

Référence électronique

Joseph Larmarange, Roselyne Vallo, Seydou Yaro, Philippe Msellati et Nicolas Méda, « Méthodes pour cartographier les tendances régionales de la prévalence du VIH à partir des Enquêtes Démographiques et de Santé (EDS) », *Cybergeo : European Journal of Geography* [En ligne], Systèmes, Modélisation, Géostatistiques, article 539, mis en ligne le 16 juin 2011, consulté le 17 septembre 2012. URL : http://cybergeo.revues.org/23782 ; DOI : 10.4000/cybergeo.23782

À propos des auteurs

Joseph Larmarange
CEPED (UMR 196 Paris Descartes INED IRD), IRD, France.
joseph.larmarange@ceped.org
Roselyne Vallo

Université de Montpellier I / INSERM U 1058 / Départements d'information médicale et d'anatomie cytologie pathologiques, CHU Montpellier, France.
roselyne_vallo@yahoo.fr
Seydou Yaro
Centre Muraz, Burkina Faso.
yaro_seydou@yahoo.com
Philippe Msellati
UMI 233 IRD/Université de Montpellier I, France.
philippe.msellati@ird.fr
Nicolas Méda
Centre Muraz, Burkina Faso.
nmeda.muraz@fasonet.bf

Résumés

Pour de nombreux pays, en particulier en Afrique subsaharienne, les Enquêtes Démographiques et de Santé (EDS) constituent la principale estimation de la prévalence du VIH au niveau national et en population générale. Plusieurs EDS collectent la longitude et la latitude des grappes enquêtées.
Dans cet article, nous présentons trois approches méthodologiques pour cartographier les variations spatiales de la prévalence du VIH à partir des EDS. Ces approches sont appliquées à des simulations d'EDS échantillonnées à partir d'un pays modèle. Les surfaces estimées sont alors comparées à la surface initiale du modèle.
Nous montrons qu'une méthode utilisant des estimateurs à noyau à fenêtres adaptatives de même effectif permet d'estimer les principales tendances régionales des épidémies. Son application aux données de l'EDS 2003 du Burkina Faso fournit une image plausible de la situation épidémiologique dans ce pays.

Methods for mapping regional trends of HIV prevalence from Demographic and Health Surveys (DHS)

For many countries, in particular in sub-Saharan Africa, Demographic and Health Surveys (DHS) are the main estimates of HIV prevalence at national level in general population. Several DHS collect longitude and latitude of surveyed clusters.
In this paper, we present three methodological approaches for mapping spatial variations of HIV prevalence from DHS. These approaches are applied to DHS simulation sampled from a model country. The estimated surfaces are then compared with the initial surface of the model.
We show that a method using kernel estimators with adaptive bandwidths of the same number of observed people allows estimating main regional trends of the epidemics. Its application to data from 2003 DHS of Burkina Faso give a plausible picture of the epidemiological situation in this country.

Entrées d'index

Mots-clés : enquêtes démographiques et de santé, interpolation, interpolation par noyaux, méthodologie, pays en développement, tendances régionales, VIH
Keywords : demographic and health surveys, developing countries, HIV, interpolation, kernel interpolation, methodology, regional trends

Cybergeo : European Journal of Geography

Systèmes, Modélisation, Géostatistiques

Joseph Larmarange, Roselyne Vallo, Seydou Yaro, Philippe Msellati et Nicolas Méda

Méthodes pour cartographier les tendances régionales de la prévalence du VIH à partir des Enquêtes Démographiques et de Santé (EDS)

Référence électronique
Joseph Larmarange, Roselyne Vallo, Seydou Yaro, Philippe Msellati et Nicolas Méda, « Méthodes pour cartographier les tendances régionales de la prévalence du VIH à partir des Enquêtes Démographiques et de Santé (EDS) », *Cybergeo : European Journal of Geography* [En ligne], Systèmes, Modélisation, Géostatistiques, article 539, mis en ligne le 16 juin 2011, consulté le 17 septembre 2012. URL : http://cybergeo.revues.org/23782 ; DOI : 10.4000/cybergeo.23782

Éditeur : CNRS-UMR Géographie-cités 8504
http://cybergeo.revues.org
http://www.revues.org

Document accessible en ligne sur :
http://cybergeo.revues.org/23782
Document généré automatiquement le 17 septembre 2012.