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

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

Introduction

1 The main epidemiological indicator for tracking the AIDS epidemic in sub-Saharan Africa is HIV prevalence, i.e. the proportion of people infected[1]. This is the figure UNAIDS[2] uses in its two-yearly report on the global epidemic to estimate the annual number of new infections and deaths in each country. The way national HIV prevalence is estimated has been revised a number of times in recent years (Larmarange, 2009) following the development in the early 2000s of national surveys in the general population with collection of blood samples and HIV screening. Most of them are Demographic and Health Surveys (DHS) to which an HIV screening module has been added. The DHSs have made it possible to refine estimates of HIV prevalence among adults (aged 15-49) both nationally and regionally[3] and are usually the only source of data in the general population.

2 Most DHSs also collect geographical coordinates (longitude and latitude) for the survey zones. So it is theoretically possible to use these surveys to estimate and represent spatial variations in HIV prevalence at a scale lower than the region. This sort of mapping would be a public health tool for revealing the zones worst affected by the epidemic, guiding the implementation of AIDS programmes and gaining a better understanding of the discrepancies observed between the DHSs and the sentinel surveillance of pregnant women[4], another major source of data used in estimating prevalence figures (Boerma, Ghys and Walker, 2003).

Demographic and Health Surveys (DHSs)

3 Demographic and Health Surveys (DHSs) are a major survey programme carried out in over 75 countries in the global South. Since 1984, more than 200 surveys have been held at regular intervals. Questions about fertility, family planning and infant-child mortality have gradually been supplemented, depending on the country and survey, by modules on mother and child health, knowledge and behaviour concerning HIV/AIDS and sexually transmitted infections (STIs), domestic violence, female genital cutting, child growth measurements, anaemia tests, HIV prevalence, etc. The questionnaires are standardised so that comparisons can be made between countries and over time.

4 The surveys are held at regular intervals by national statistical institutes with technical support from Macro International, Inc. All the final reports and databases are available free online from a dedicated site: http://www.measuredhs.com.

5 The DHSs in each country use a similar stratified two-stage sample design. The country is divided into a number of strata, one per administrative region and urban-rural place of residence. The master sample of primary units is made up of the enumeration areas of the most recent population census. In the first stage, the primary units or clusters are randomly selected, separately in each stratum, with a probability proportional to their number of ordinary households[5] in the census. After an exhaustive enumeration of the households in each cluster, a predetermined number of households are selected in the second stage using a simple random sampling. Depending on the country, only part (one-third or one-half) of these households are then eligible for the HIV survey. To allow for the complex DHS sample design, each database contains a weighting variable making the sample representative at national and regional level. This variable is proportional to the inverse sampling probability of each household, namely the probability that the household will be surveyed. Table 1 compares the samples of various recent DHSs.

6 Some DHSs also use GPS to collect the geographical coordinates of each survey cluster. Since HIV screening modules have been introduced, in order to ensure the anonymity of respondents, these coordinates are randomly offset by up to two kilometres in urban areas and five kilometres in rural areas[6].

7 From 2004, surveys specific to HIV/AIDS but similar to the DHS were designed, with a shorter questionnaire: the AIDS Indicator Surveys[7] (AISs).

Table 1: Samples for HIV screening from 25 DHSs or AISs

Country	Year	Type	Clusters	Individuals aged 15-49 tested for HIV	Average number tested per cluster	National HIV prevalence, age 15-49 (%)
Burkina Faso	2003	DHS	400	7,244	18.1	1.8
Cameroun	2004	DHS	466	9,900	21.2	5.5
Côte d'Ivoire	2005	AIS	249	8,436	33.9	4.7
Democratic Republic of Congo	2007	DHS	300	8,504	28.3	1.3
Ethiopia	2005	DHS	540	10,540	19.5	1.4
Ghana	2003	DHS	412	9,144	22.2	2.2
Guinea	2005	DHS	297	6,388	21.5	1.5
Kenya	2003	DHS	400	6,001	15.0	6.7
Kenya	2008-09	DHS	400	6,707	16.8	6.3
Lesotho	2004	DHS	405	5,043	12.5	23.4
Liberia	2007	DHS	300	11,733	39.1	1.6
Malawi	2004	DHS	522	5,150	9.9	12.0
Mali	2001	DHS	403	6,475	16.1	1.8
Mali	2006	DHS	407	8,141	20.0	1.3
Niger	2006	DHS	345	7,262	21.0	0.7
Rwanda	2005	DHS	462	10,016	21.7	3.0
Senegal	2005	DHS	377	7,503	19.9	0.7
Sierra Leone	2008	DHS	353	6,174	17.5	1.5
Swaziland	2006-07	DHS	275	8,187	29.8	25.9
Tanzania	2003-04	AIS	345	10,747	31.2	7.0
Tanzania	2007-08	AIS	475	15,044	31.7	5.7
Uganda	2004-05	AIS	417	16,906	40.5	6.4
Zambia	2001-02	DHS	320	3,807	11.9	15.6
Zambia	2007	DHS	320	10,444	34.8	14.3
Zimbabwe	2005-06	DHS	400	12 796	32.0	18.1

AIS: AIDS Indicator Survey (or AIDS Impact Survey); DHS: Demographic and Health Survey
Sources: survey final reports available at http://www.measuredhs.com.

Objective

8 Although there are many publications concerning the DHSs, spatial analysis based on them are less numerous. Of more than 650 scientific papers identified on the MeasureDHS website as relating to DHS data, only 13 are classified under "spatial modelling"[8]. Most of them are atlases of choropleth maps (TACAIDS, 2006); national or regional maps, as with the online tool HIVmapper[9]; or multi-level analysis including one or more geographical variables (distance to a road or infrastructure, spatial typology, etc.). This has been made easier in recent years by the uploading of georeferenced map underlays of the administrative units used in the DHSs[10].

9 Chloropleth maps by administrative region are not always appropriate for displaying the spatialisation of a phenomenon, since administrative boundaries rarely correspond to variations inherent in that phenomenon. Furthermore, for densely populated regions that are consequently extensively sampled, maps by region lead to a loss of information at a more local level: infraregional differences are obscured.

10 Our objective is thus to estimate from DHS data, irrespective of administrative divisions, a prevalence surface that reveals the main spatial variations in the epidemic, while retaining an infraregional local accuracy for the adequately surveyed areas.

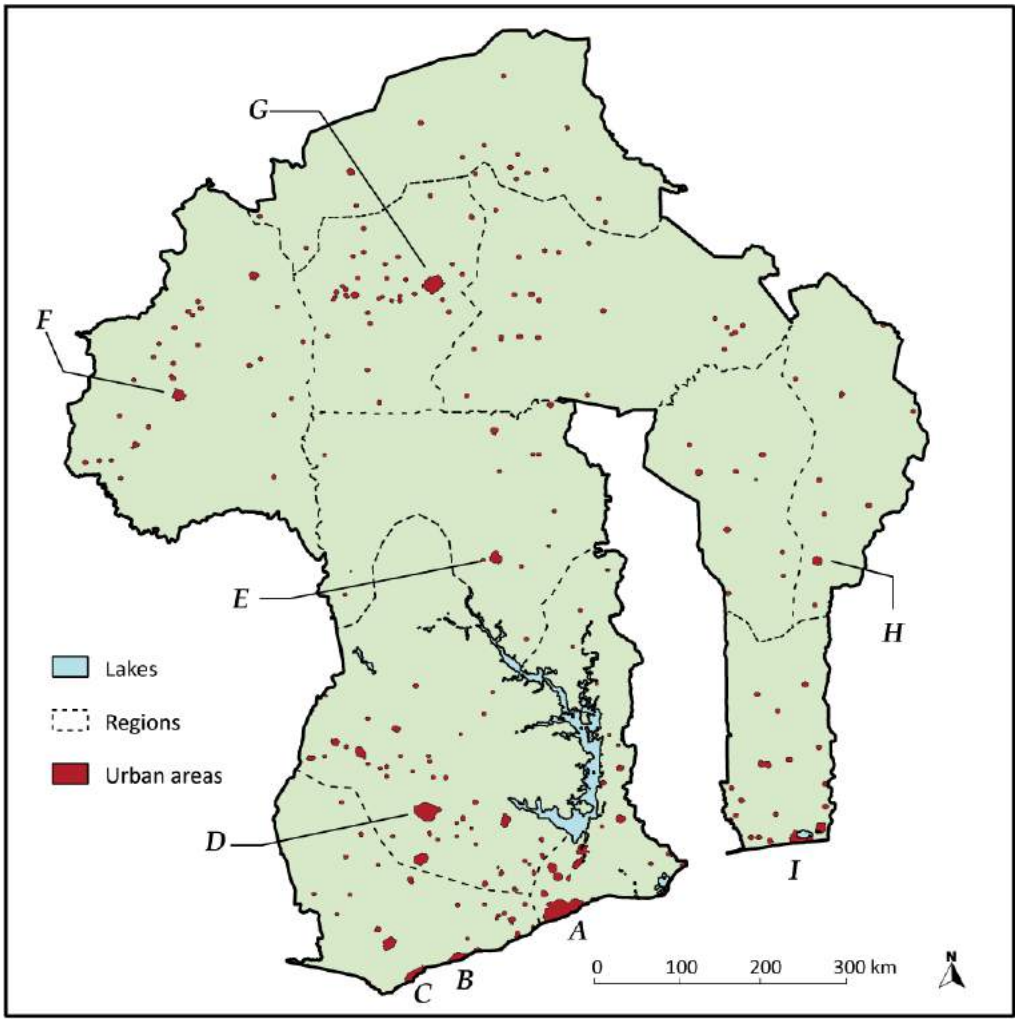
11 In order to test various methodological approaches, we devised a fictitious country for which we simulated DHSs: this makes it possible to compare the prevalence surface estimated from survey data with the model's original prevalence surface. Three approaches were tested: a spatial smoothing based pn circles of equal number of persons surveyed before spatial interpolation; an adaptation of Davies and Hazelton (2010) using kernel estimators with adaptive bandwidths; and a kernel estimation with adaptive bandwidths calculated from circles of equal number of persons surveyed. Finally, these approaches were applied to real data from Burkina Faso's 2003 DHS.

Methods

Devising a model country

12 Benin, Burkina Faso and Ghana were combined to create a fictitious country to be used as a model. Togo was deliberately excluded in order to obtain a concave shape to make the estimation more complex, since the "missing" part was not surveyed. A shape of this sort can be found in the outline of countries such as Senegal. Data from the Global Rural-Urban Mapping Project (GRUMP) were used to distribute the population over the territory: population densities from the year 2000 at a resolution of 30 arcseconds (CIESEN *et al.,* 2005a) and territorial divisions in urban and rural areas (CIESEN *et al.,* 2005b). The territory was then divided into 9,137 primary units (7,818 rural and 1,319 urban) distributed regularly at an average resolution of 2 arcminutes in urban areas and 5 arcminutes in rural areas. Then the area, average density and population (obtained by multiplying density by area) of each unit were calculated. The country was divided into 11 administrative units and the main urban centres were identified by the letters A to I (see Figure 1).

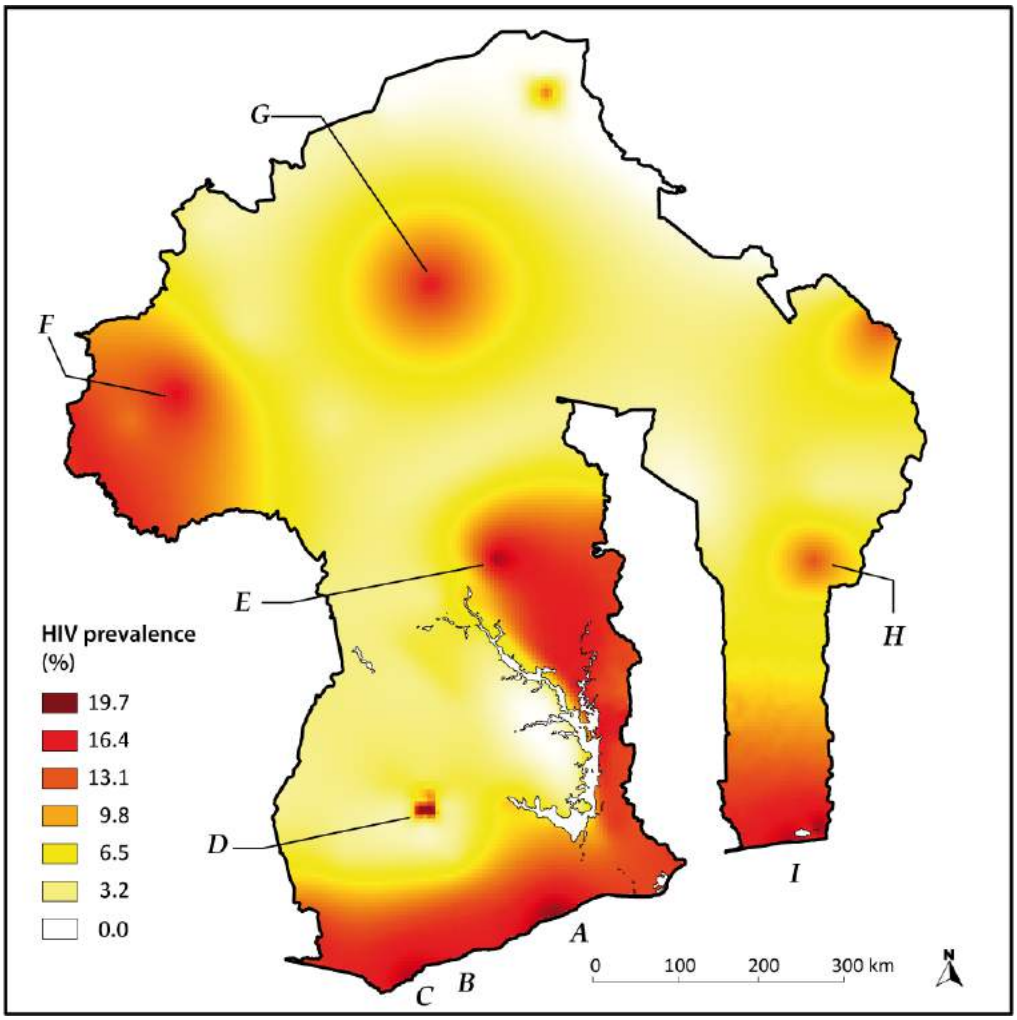
Figure 1: Urban areas and regions in the model



13 Next, we created a prevalence surface (Figure 2) by spatial interpolation[11] from points chosen *ad hoc* for the surface to present various diffusion patterns: major town with prevalence concentrated within it and low outside (D); major and mid-sized towns (G and H) with gradual diffusion; localised peak in sparsely populated rural area in the north; discontinuity between the shores of a major lake; gradient inland from the coast in the south, with two major conurbations (A and I) and two mid-sized towns (B and C); diffusion from the western border and a city across the border to the east.

14 National prevalence was obtained from the average prevalence of the primary units[12] weighted by population. In order to obtain a national prevalence of 10% we multiplied the prevalence in each unit by the same scale factor. The surface obtained is so constructed as to be spatially continuous and highly self-correlated.

Figure 2: Surface of HIV prevalence in the model (national prevalence of 10%, created *ad hoc* by spatial interpolation)



Data format and DHS simulation

15 DHS data come in the form of a scatter plot corresponding to the various clusters surveyed. Since the clusters are selected with a probability proportional to their number of households, this scatter plot reflects variations in population density: the points are numerous and close together in densely populated areas and, conversely, few and far between in sparsely populated areas. For each individual tested we have their serological status, cluster membership and statistical weight. All the individuals in a cluster are spatially located in the same point.

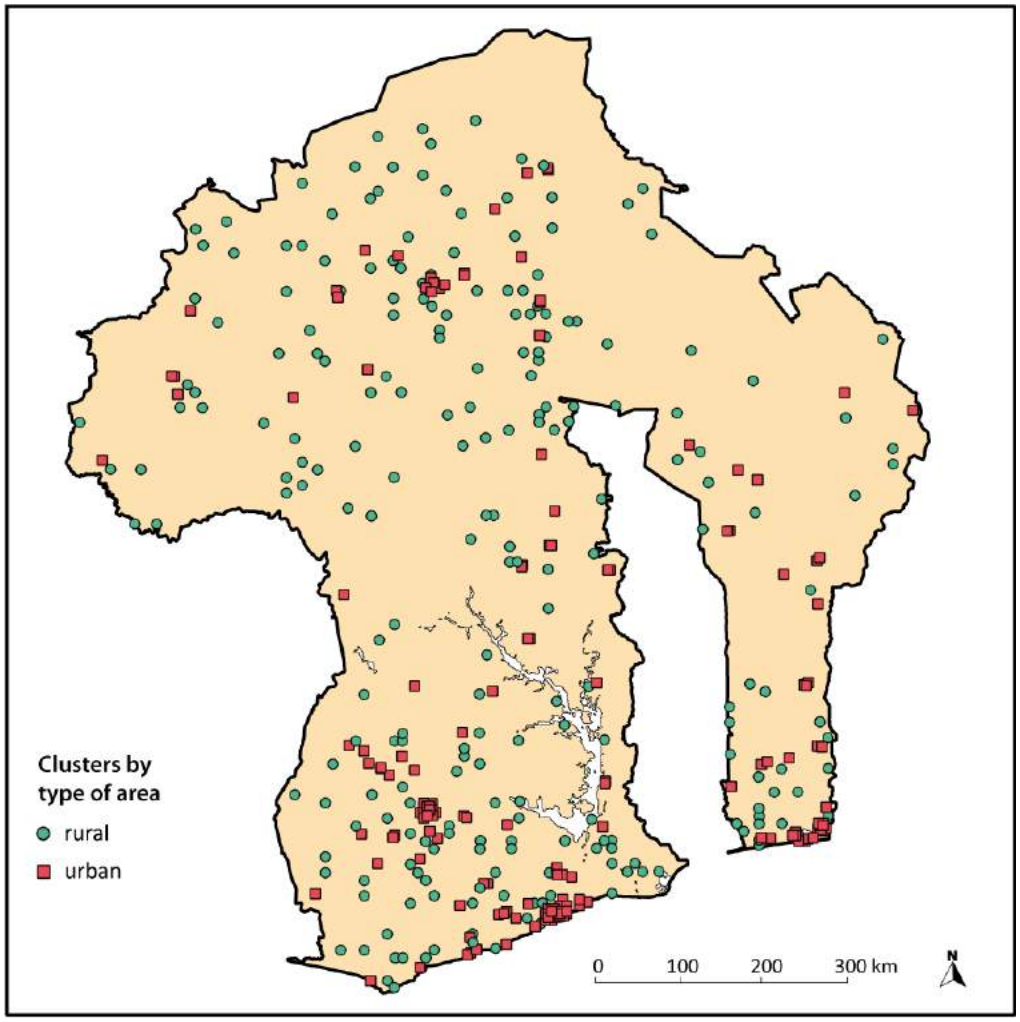
16 DHS simulations were carried out to reproduce data comparable with actual surveys, using three parameters: national prevalence, total number of persons surveyed and the number of first-level clusters. In order to obtain the chosen value for national prevalence, the prevalence in each of the primary units is multiplied by the same appropriate scale factor. Each of the eleven regions is divided into two strata, urban and rural. The number of clusters selected by stratum is proportional to their total population[13]. The clusters are selected randomly, stratum by stratum from the primary units, with a sampling probability proportional to their population.

17 Next, the number of persons surveyed per cluster is determined randomly by a normal distribution, in order to reproduce the variability in the number of persons surveyed per cluster that can be observed in DHSs[14]. The number is corrected to ensure that the total corresponds to the target figure. Then the number of HIV-positive people in each cluster is determined randomly by a binomial distribution. A weighting factor, similar to that used for the DHSs, is calculated and applied to individuals.

18 The data generated by simulation are comparable to the actual data from the Demographic and Health Surveys. The reason is that the distribution of observed prevalence generated

by simulation presents the same profile as those observed in a number of DHSs (table not reproduced). Figure 3 shows the spatial distribution of the clusters obtained by the simulation we use as an example in this article.

Figure 3: Distribution of clusters obtained by simulating a DHS.



Simulation parameters: national prevalence 10%, 8,000 people surveyed, 400 clusters (rounding means that the actual figure is 401).

Spatial interpolation approach

19 Spatial interpolation techniques use a scatter plot to produce the surface of a phenomenon. For each point on the map where the value of the phenomenon is unknown, that value is estimated from those points for which the information is available. These various techniques assume that the variations of the interpolated variable are spatially continuous and give greater weight to nearer observations over more distant ones, under the hypothesis that neighbouring points are similar. These techniques are used to estimate a phenomenon over an entire territory on the basis of fragmentary information restricted to a finite number of points.

20 The general formula for determining the estimated value $\hat{s}(x,y)$ of surface s at point x,y, given the values of s at n points xi,yi where i varies from 1 to n, is as follows:

21 Equation 1

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

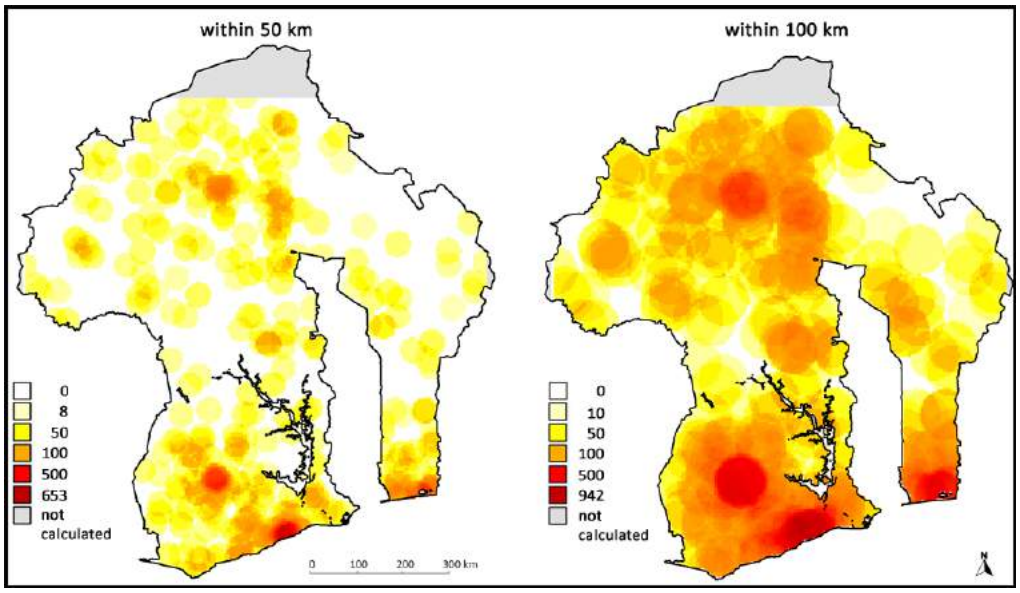
22 where di represents the geometric distance between observation i and point x,y and w a weighting function diminishing for distant observations. Most of the weighting functions used retain the observed values, in other words, the estimated value at point k remains equal to its observed value: $\hat{s}(xk,yk) = s(xk,yk)$.

23 To use these techniques it is necessary to have two items of information for each point observed: its position and the prevalence value at that point. However, the observed prevalence for each survey cluster, the prevalence calculated from the people tested in the cluster, presents high variance and margin of error. This is because the number of persons tested in each cluster is small, from 10 to 40 on average (see Table 1). In fact, the observed prevalence is more a reflection of random sampling variations than of the extent of the epidemic.

24 We have already described (Larmarange, 2007; Larmarange *et al.,* 2006) a methodological approach smoothing the prevalence for each cluster in advance, before executing standard spatial interpolation.

25 Taking inspiration from moving average smoothing techniques based on circles of equal radius (Griffin, 1949; Krumbein, 1956; Nettleton, 1954), a circle is drawn around each cluster and the prevalence of the central cluster is then recalculated from all the people tested within that circle[15]. Spatial interpolation can then be applied to these smoothed prevalence values.

Figure 4: Number of persons surveyed within a radius of 50 and 100 kilometres



N.B. For each point on the map is shown the number of persons surveyed within 50 and 100 kilometres of that point, for the DHS simulation in Figure 3.

26 However, the use of circles of equal radius is inappropriate, since the clusters are very unevenly distributed (Figure 4). A sufficiently large radius needs to be determined for the smoothed prevalence values to be calculated from a sufficient number of individuals, especially in those areas where the clusters are widely dispersed. At the same time, in densely populated survey areas, smaller circles could be used, since the numbers are amply sufficient. The accuracy of an estimated proportion is related to the number of observations. It is consequently better to use circles not of equal radius but of equal number of persons surveyed.

27 So once a number of persons surveyed N is fixed, the smoothed prevalence for each cluster is calculated from observations located within a circle such that the number of persons surveyed within it is at least N. If we denote as $cumi(r)$ the function that gives the cumulative total of observations located within a radius r of the observed cluster i, then the radius ri of the smoothing circle for that cluster, with a minimum number of individuals N, is $min[r \mid cumi(r) \geq N]$.

28 To produce a prevalence surface, the smoothed prevalences are then spatially interpolated by ordinary kriging. Kriging (Krige, 1951; Matheron, 1963) has the advantage of taking into account the spatial dependence structure of the data (Baillargeon, 2005). Variance as a function of the distance between points is measured empirically as a semivariogram, which is then modelled. The unknown values are then estimated from adjacent known values, weighted by this semivariogram in order to obtain an unbiased forecast with minimum variance.

Kernel estimator approaches

29 Another field of geostatistical analysis is based on estimating density surfaces from kernel estimators (Silverman, 1986; Wand and Jones, 1994). These techniques are designed to construct a surface from a scatter plot, where each point represents an observed case. The surface obtained may be expressed as the number of cases per surface unit (intensity surface) or by reducing the integral to one (density surface).

30 A density surface is constructed around each observed case in such a way that the density is highest at that point and diminishes with distance. The estimated intensity surface corresponds, therefore, to the sum of these density surfaces (see Figure 5 for a one-dimensional example). In mathematical terms, the intensity surface $\hat{s}$ at point (x,y) is estimated by the following expression:

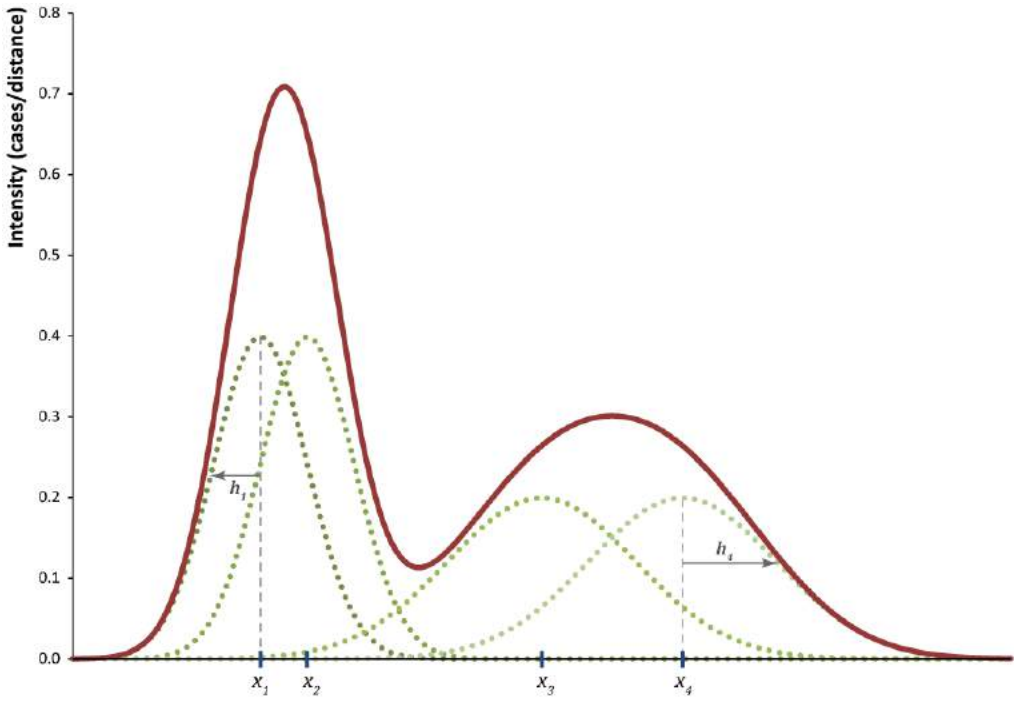
31 Equation 2

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

32 where n is the number of observed cases, di the geometrical distance between case i and point (x,y), K a density function (called the kernel) within an integral equal to 1, and hi the bandwidth used for case i. The density surface is obtained by dividing the intensity surface by the number of observed cases (n).

33 The bandwidth makes it possible to apply greater or lesser smoothing to the data. The estimate will be a fixed-bandwidth estimate if it uses a constant ($\forall i$, $hi = h$) or an adaptive bandwidth estimate if hi varies according to the observed case ($\exists\, a,b \mid ha \neq hb$).

Figure 5: Example calculation of an intensity function with a Gaussian adaptive bandwidth kernel estimator (one dimension)



N.B. Estimation from 4 observed points at *x1*, *x2*, *x3* and *x4*. The estimated intensity function (solid curve) is the sum of the 4 density functions (dotted curves) calculated with a Gaussian kernel for each observed point and centred on that point. *h1* and *h4* represent the width of bandwidth used to calculate the density functions.

34 Various researchers have used kernel estimators for spatial epidemiology (Gatrell, Bailey, Diggle and Rowlingson, 1996), such as estimating a surface of relative risks (Bithell, 1990; Davies and Hazelton, 2010; Kelsall and Diggle, 1995). The surface of relative risks corresponds to the ratio between the density surface of positive cases and the density surface of control cases (the population exposed to the risk). The two density surfaces are estimated separately from two independent scatter plots. The scatter plot of positive cases is usually taken from epidemiological monitoring. The control cases may be determined in various ways:

random selection from telephone directories in Gatrell, sample of postcodes in Davies and Hazelton, etc.

35 With fixed bandwidths, some researchers (Bithell, 1990; Kelsall and Diggle, 1995) suggest using the same constant h to estimate the two density surfaces (positive cases and control cases). However, some research (Bithell, 1990, Carlos, Shi, Sargent, Tanski and Berke, 2010; Davies and Hazelton, 2010) suggests that use of an adaptive bandwidth is more appropriate for health-related matters in order to correspond more closely to the spatial distribution of population and thus reduce the smoothing of information.

36 The main difficulty with kernel estimators is choosing the right value for the smoothing bandwidth. Kelsall and Diggle (1995) explore various approaches for automatically determining the value of the bandwidth from data for fixed bandwidths. Other research covers this question for adaptive bandwidths (Sain, 1994; 2002) to estimate a single surface but without investigating the question of the ratio of two surfaces estimated simultaneously.

37 In a recent article, Davies and Hazelton (2010) propose an approach for estimating a surface of relative risks using adaptive bandwidths whose values are determined from observed data. First, the authors determine a pilot value for the bandwidth, the same for both positive and control cases, using the Maximal Smoothing Principle[16] proposed by Terrell (1990). Then they determine the local values of the smoothing bandwidth from this pilot value separately for positive and control cases. Their approach has been tested on six different theoretical situations and compared, using the ISE criterion (Integrated Squared Error), with estimators using a fixed bandwidth determined by the maximal smoothing principle: their adaptive bandwidth approach turns out generally to be more efficient, particularly with large samples. This approach is used in a package called sparr for the statistics software R (Davies, Hazelton and Marshall, 2010).

38 In our case we are seeking not to produce a surface of relative risks (ratio of two density surfaces) but rather a surface of prevalence (ratio of two intensity surfaces). The functions in the sparr package can be used to calculate both density surfaces and intensity surfaces. So we tested Davies and Hazelton's approach adapted to calculate the ratio of two intensity surfaces instead of two density surfaces.

39 Davies and Hazelton address two independent scatter plots of positive and control cases, each based on simple random sampling: the location of the positive cases differs, therefore, from that of the control cases. In the case of DHSs, the data come from two-level sampling and the scatter plots correspond to level-one sampling. The positive cases (people tested HIV-positive) and the control cases (people tested irrespective of result) in a given cluster have the same spatial location. Consequently it is more appropriate to use a single bandwidth hi for cases in a single cluster.

40 We therefore tested another approach using adaptive bandwidth kernel estimators so that the bandwidth used for cases in a single cluster would depend solely on their location and specifically the number of observations in the vicinity of that cluster. For the estimation of the intensity surface of observed cases, the principle is similar to the nearest neighbour technique described by Silverman (1986) and Altman (1992) among others, and tested by Bithell (1990). A minimum number of observations N is set and the radius hi of the smoothing bandwidth is therefore proportional to the radius of the circle to be drawn around the cluster in order to capture this minimum number. These are in fact the radii ri of the smoothing circles of equal number described above in the spatial interpolation approach. In the special case of DHS data, the control cases in a particular cluster are given the same bandwidth: if points i and j belong to the same cluster k, then $hi = hj = \lambda rk$ (where λ is a scale factor). For the positive cases we apply the same bandwidth as that calculated for the control cases in the same cluster, namely λrk.

41 A number of density functions may be used for kernel K. It is generally agreed that the choice of function is less important than the size of the bandwidth. Davies and Hazelton (2010) report that Gaussian kernels (using the normal distribution) are often used to estimate two-dimensional surfaces, although the use of finite extent kernels[17] (such as the biweight function) is also common. Although finite extent kernels have a theoretical advantage for adaptive

bandwidths, in practice the Gaussian kernel is more suitable when the distribution of points is highly uneven, particularly in regions where the number of observations is small.

42 We consequently opted for the Gaussian kernel, using a scale factor λ of 0.5: the bandwidth *hi* of a case located in cluster *k* is thus *hi* = *rk*/2. This means that 86% of the kernel's intensity lies within a circle of radius *rk*.

Choice of a comparison indicator

43 In order to compare estimated prevalence surfaces with the prevalence surfaces in the model, we used the MISD indicator (Mean Integrated Squared Difference) (Anderson and Titterington, 1997), an indicator similar to the MISE (Mean Integrated Squared Error) (Wand and Jones, 1994). The MISD corresponds to the expected value of the squared difference at each point on the surface. Take two spatial surfaces $\hat{s}1(x,y)$ and $\hat{s}2(x,y)$. The MISD between $\hat{s}_1$ and $\hat{s}_2$ is calculated as follows[18]:

44 Equation 3

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

45 The MISD may be approximated from a fine matrix of *p* regularly spaced points:

46 Equation 4

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

47 In this way the MISD quantifies the deviation at each point between the estimated surface and the model surface. Consequently the best estimate will be the one that minimises the MISD.

Software used

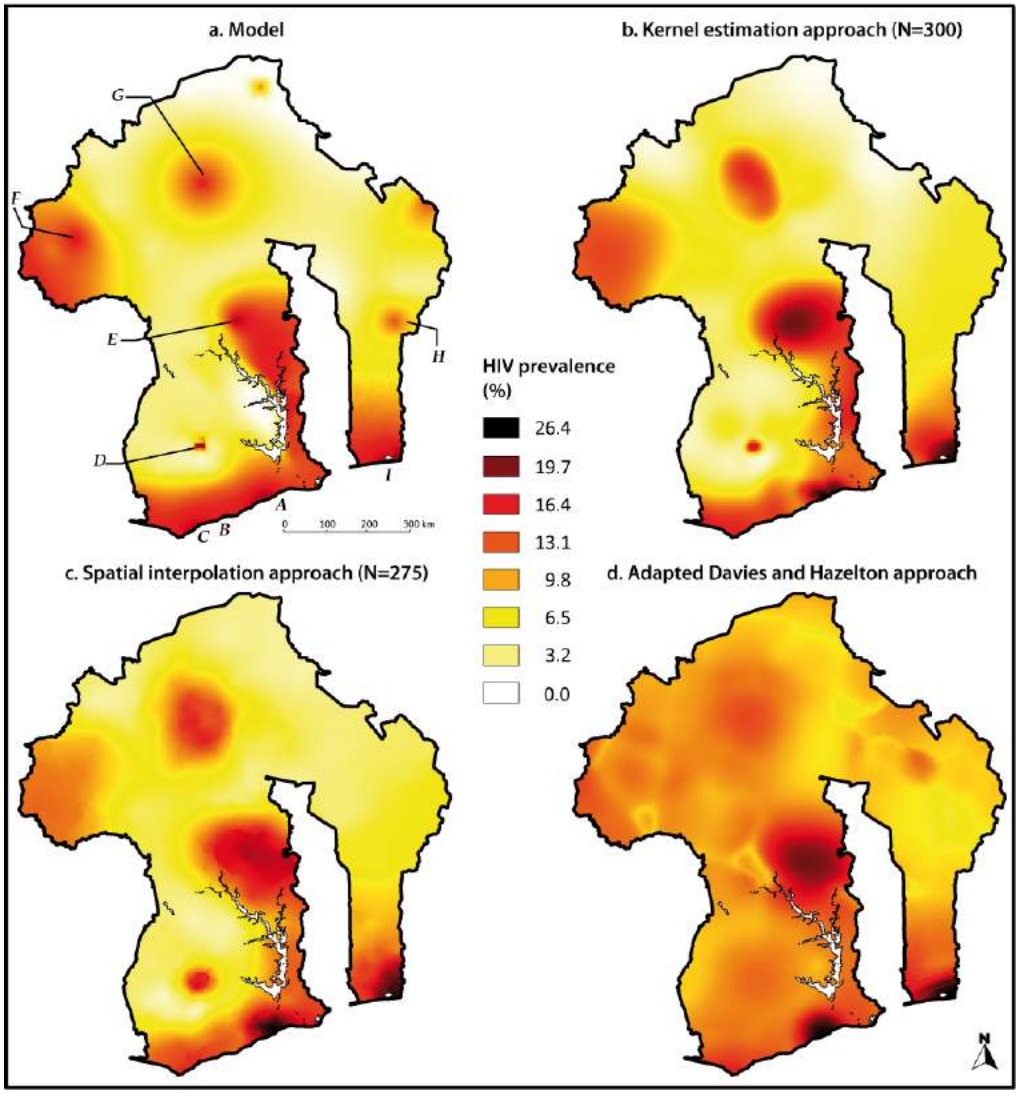
48 The calculations were done with R statistics software (R Development Core Team, 2007). We used the kriging function in the gstat package (Pebesma, 2004) for spatial interpolation by ordinary kriging. The approach adapted from Davies and Hazelton was implemented with the sparr package (Davies, Hazelton and Marshall, 2010) developed by the authors. For the kernel estimator approach we used the KernSur function in the GenKern package (Lucy and Aykroyd, 2010). We also developed our own functions, available in a package called *prevR*[19].

49 The surfaces in this article were drawn with *Quantum GIS*[20].

Results

Comparison of approaches

Figure 6: Model prevalence surface and estimated prevalence surfaces using three different approaches based on the same DHS simulation.

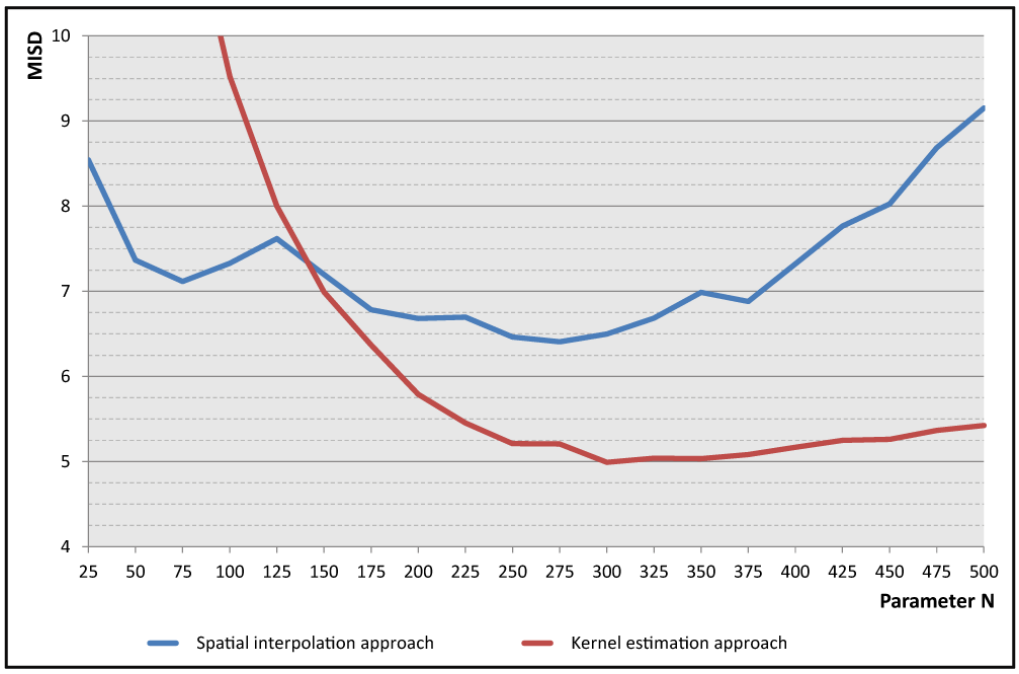


N.B. The colour scale is identical to that in Figure 2.

50 Figures 6.b, 6.c and 6.d represent prevalence surfaces estimated by the three approaches. Figures 8.b, 8.c and 8.d display the deviations between the estimated prevalence surfaces and the model prevalence surface: each point on the surface corresponds to the mathematical difference between estimated and model prevalence, namely $\hat{se}(x,y) - sm(x,y)$. The MISD is the mean of these deviations squared.

51 The adapted Davies and Hazelton approach (Figure 6.d) gives an MISD of 28.1, considerably greater than the MISDs obtained with approaches using circles of equal number of persons surveyed (spatial interpolation and kernel estimators, see Figure 7). Prevalence values are overestimated by at least five points over most of the surface (Figure 8.d). The smoothing produces “stripes” on the prevalence surface in sparsely surveyed areas (west of E, south of F, between G and H, etc.). The estimated prevalence values are heavily smoothed around D and between G and F, and do not reproduce the variations on the model surface.

Figure 7: MISD for various values of *N*



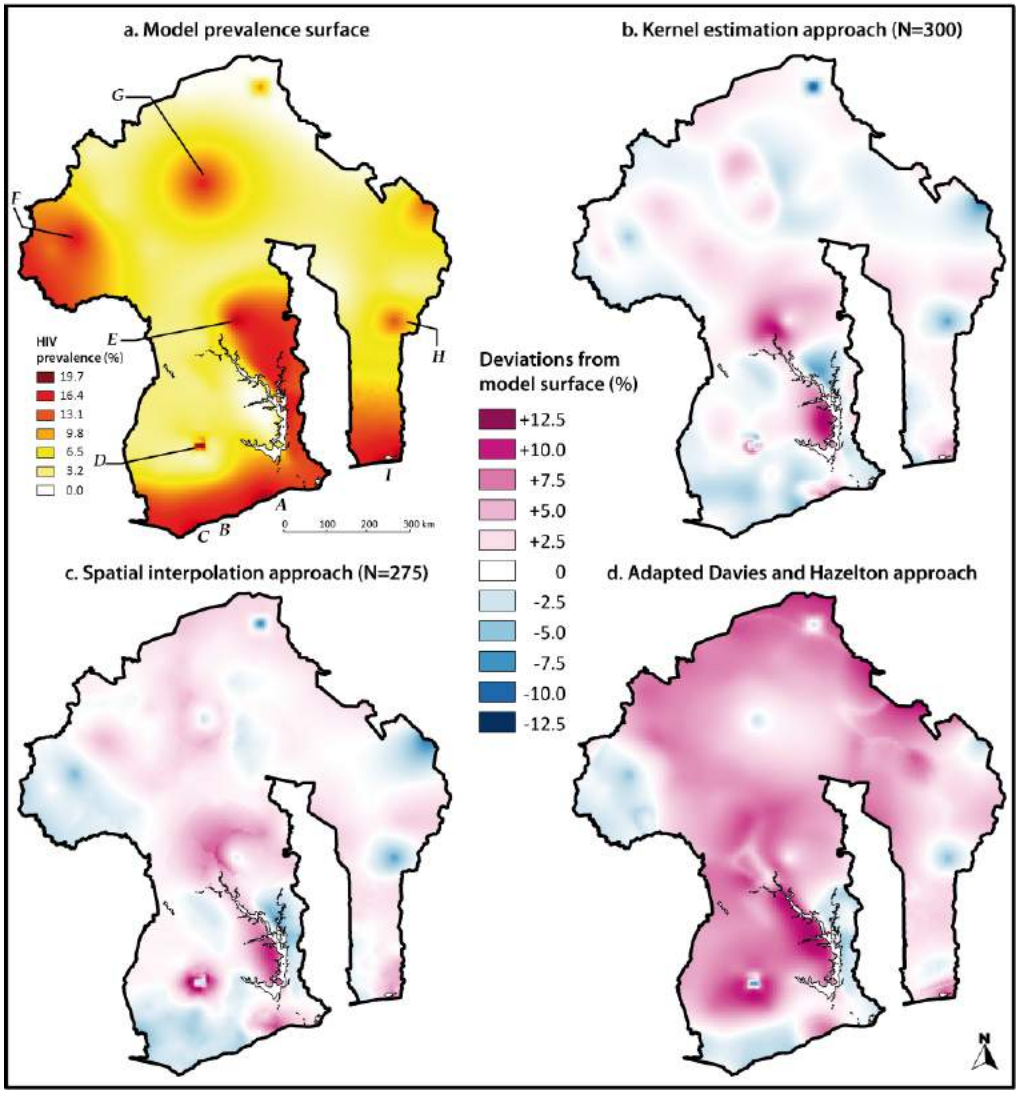
N.B. See Appendix Table 1 for value details.

52 The two approaches based on circles of equal number of persons surveyed behave similarly: the MISD indicator diminishes as parameter *N* increases, down to a minimum before the smoothing becomes excessive and the MISD rises (see Figure 7). At low values of *N*, the kernel estimator approach yields an MISD higher than the spatial interpolation approach, but its MISD diminishes faster. The kernel estimator approach's MISD bottoms out at 5.0 at *N*=300, whereas for the spatial interpolation approach the minimal MISD is 6.4 at *N*=275.

53 Figures 6.b and 6.c show estimated prevalence surfaces with these optimum values of *N*. Overall, the main variations in the model prevalence surface are reproduced. The gradient from the south coast northwards is there, with sharper contrast due to overestimation (Figures 8.b and 8.c) for conurbations A and I. Conurbation D still shows a more concentrated prevalence than its vicinity, a "concentration" better revealed by kernel estimators. Similarly the kernel estimators more closely reproduce the gradient on the western border and the diffusion of prevalence around conurbation G. On either side of the major lake, where a clear break was introduced into the model, prevalence is overestimated to the west and underestimated to the east, since neither approach allows for natural borders.

54 Finally, the variations in sparsely surveyed areas are not reproduced: the epidemic peak in the north of the country, the diffusion around conurbation H and starting at the eastern border.

Figure 8: Model prevalence surface and deviations (mathematical differences) between the surfaces estimated by the three approaches and the model surface

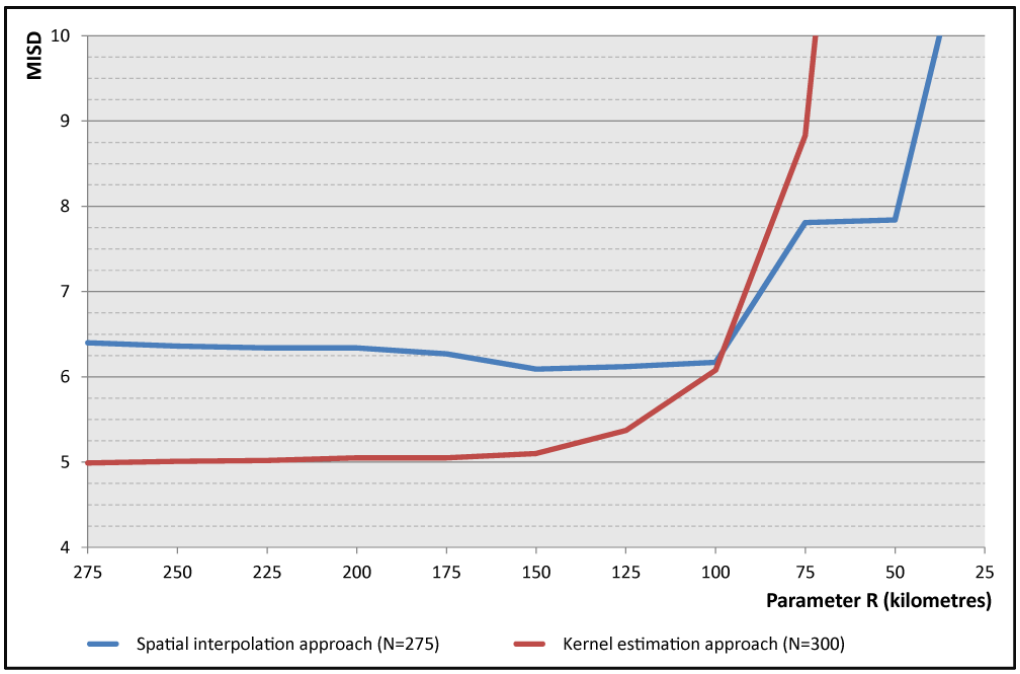


Restricting the size of the smoothing circles: addition of parameter R

55 Smoothing with circles of equal number of persons surveyed can be used to estimate prevalence from a sufficient number of observations and produces spatial smoothing that varies by region. In densely observed areas the radius of the smoothing circles is relatively small. Conversely, in sparsely surveyed areas, particularly near the borders, smoothing covers more widely dispersed clusters and the radius of the smoothing circles increases considerably (see Appendix Table 2).

56 In previous publications (Larmarange, 2007; Larmarange *et al.,* 2006), we suggested the possibility of adding a second parameter *R* corresponding to a maximum radius for the smoothing circles. This second parameter in practice only affects the clusters located in sparsely surveyed areas: if the number of persons surveyed *N* is not achieved within a radius of less than *R*, the radius of the smoothing circle is set at *R*.

Figure 9: MISD for various values of parameter *R* (fixed *N*)



N.B. Since the maximum radius of smoothing circles is 269 kilometres without parameter *R* (see table in Appendix 2), the prevalence surfaces obtained with a value of *R* of 275 kilometres are identical to those obtained without parameter *R*.

57 Figure 9 shows the MISDs obtained for various values of parameter *R* (where parameter *N* is fixed at the minimum MISD obtained without parameter *R*, namely 275 for the spatial interpolation approach and 300 for the kernel estimator approach.).

58 Although adding parameter *R* to the spatial interpolation approach slightly reduces the MISD (minimum obtained for *R* at 150 kilometres), there is no gain for the kernel estimator approach. In both cases a low value for *R* considerably increases the MISD.

59 The most effective approach for this simulation is consequently the kernel estimator approach with adaptive bandwidths of equal number of persons surveyed. Here the addition of a maximum radius for the smoothing circles does not improve the estimated prevalence surface.

Discussion

Choice of parameter N

60 The kernel estimator approach with adaptive bandwidths of equal number of persons surveyed reproduces the main variations in the model prevalence surface. The main difficulty in applying this method to real data comes from determining the right value of parameter *N* to use. For DHS simulations an MISD could be calculated because the prevalence surface to be estimated was known. Where real data are used, this prevalence surface is unknown and an MISD cannot be calculated.

61 Altman (1992) suggests that one way of determining the value for a smoothing parameter is to make a number of estimates with different values and subjectively select the one that best corresponds to the expected result. If the aim is to reveal the general outlines of the phenomenon, a high value for the smoothing parameter will be appropriate. Conversely, in order to examine local extreme values, a low value for that parameter is to be preferred. Subjective choice of the smoothing parameter is highly flexible and gives a general overview of the data. However, Altman points out that an objective method for selecting the smoothing parameter may be preferable for producing an automatic smoothing technique or for greater consistency of results among a number of investigators.

62 The adapted Davies and Hazelton approach has the advantage of proposing an automatic selection of bandwidth size on the basis of the available data. However, the prevalence surface produced is unsatisfactory in the present case, with a relatively high MISD. This is largely due to the fact that the Davies and Hazelton approach was developed for situations in which the

distribution of the two scatter plots (positive and control cases) was independent, which does not hold for the DHS data.

63 For approaches with circles of equal number of persons surveyed, the MISD varies relatively little about its minimum (see Figure 7). It is therefore reasonable to consider that the surfaces obtained, with values of *N* within plus or minus 50 of the value minimising the MISD, are acceptable.

64 In previous publications (Larmarange, 2007; Larmarange *et al.,* 2006), we attempted to model the optimal value of *N* (denoted *No*) as a function of the observed national prevalence (*p*), the number of persons tested (*n*) and the number of clusters surveys (*g*), namely the three parameters used to simulate a DHS. To that end we simulated 22,000 DHSs with various values for the three parameters and calculated the optimal value of *N* for each simulation. For reasons of time and computer power, the criterion used for determining the optimal value of *N* for a given simulation did not consist of minimising the MISD calculated over the entire prevalence surface. The indicator used was the minimisation of the average absolute deviation between estimated prevalence and model prevalence, calculated solely from the survey clusters. Compared with the MISD calculated from the entire prevalence surface, this indicator gives more weight to the most densely populated areas (since they contain more survey clusters). Use of absolute deviation increases the importance of clusters where deviations are small compared with square deviation, which increases the importance of large deviations. In practice, the optimal values for *N* calculated with this indicator are lower than the optimal values calculated by minimising MISD. Modelling, by regression, the results obtained produces the following formula:

65 Equation 5

66 $No = 2{,}688 \bullet n^{0.419} \bullet p^{-0.361} \bullet g^{0.037} - 91.011$

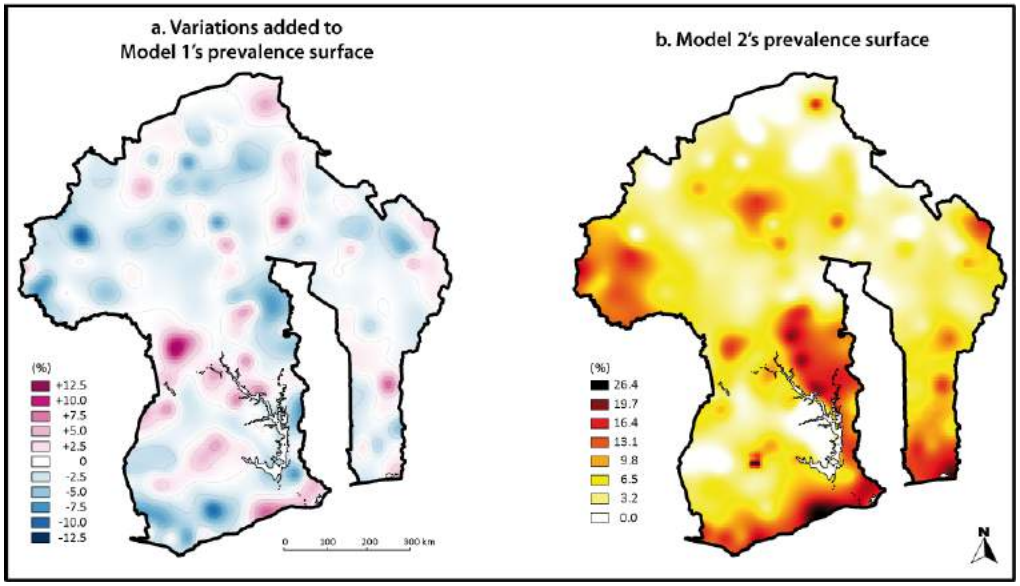
67 This result is purely illustrative. It depends on the prevalence surface imposed on the model. Other prevalence surfaces would produce other optimal values. However, this equation can be used to guide the choice of parameter *N* when applied to real data, by providing an order of magnitude.

Estimated prevalence surfaces and trend surfaces

68 The model prevalence surface is highly organised. It was constructed *ad hoc* to be multi-polarised and have continuous, regular gradients. It is only logical that this structure will be reproduced in general lines from a sufficient sample.

69 In order to test the results obtained from a less regular initial prevalence surface, we devised a second model adding localised random variations to the first model. A variation surface (Figure 10.a) was generated from 800 randomly selected points, each with a randomly defined bandwidth of action and a randomly determined positive or negative contribution. A correction was made to ensure that Model 2's prevalence surface[21] (Figure 10.b) had no negative prevalence values and national prevalence (allowing for population density) was always 10%. This new surface in fact contains many irregularities while having an underlying spatial structure (that of Model 1).

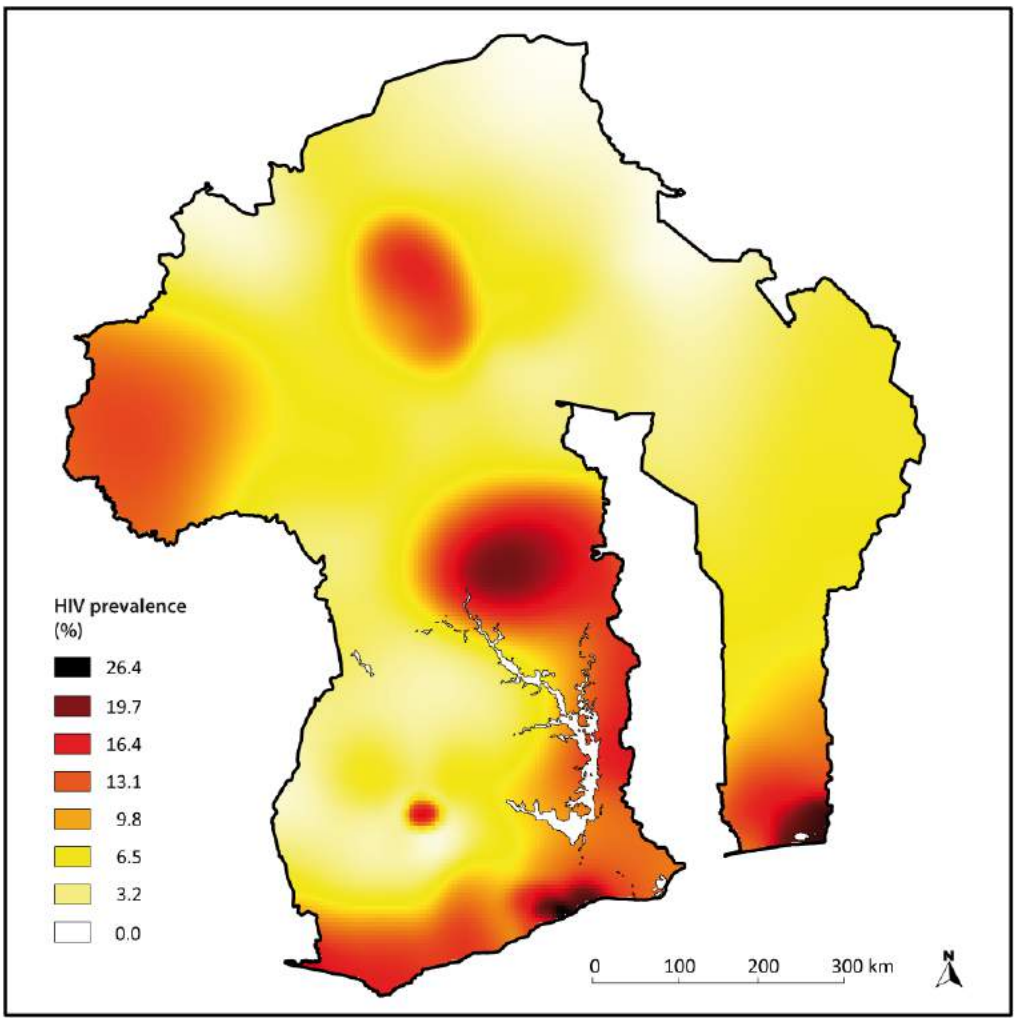
Figure 10: Random variations added to Model 1 and prevalence surface of Model 2



70 A further DHS was simulated from Model 2[22]. Figure 11 shows the prevalence surface estimated by the kernel estimator approach with N=300.

71 This surface is relatively similar to the one obtained by applying the same method to a DHS simulation based on Model 1 (Figure 6.b). This smoothing method filters out local variations to reveal underlying regional variations.

Figure 11: Prevalence surface estimated by kernel estimator approach (N=300) from DHS simulation from Model 2



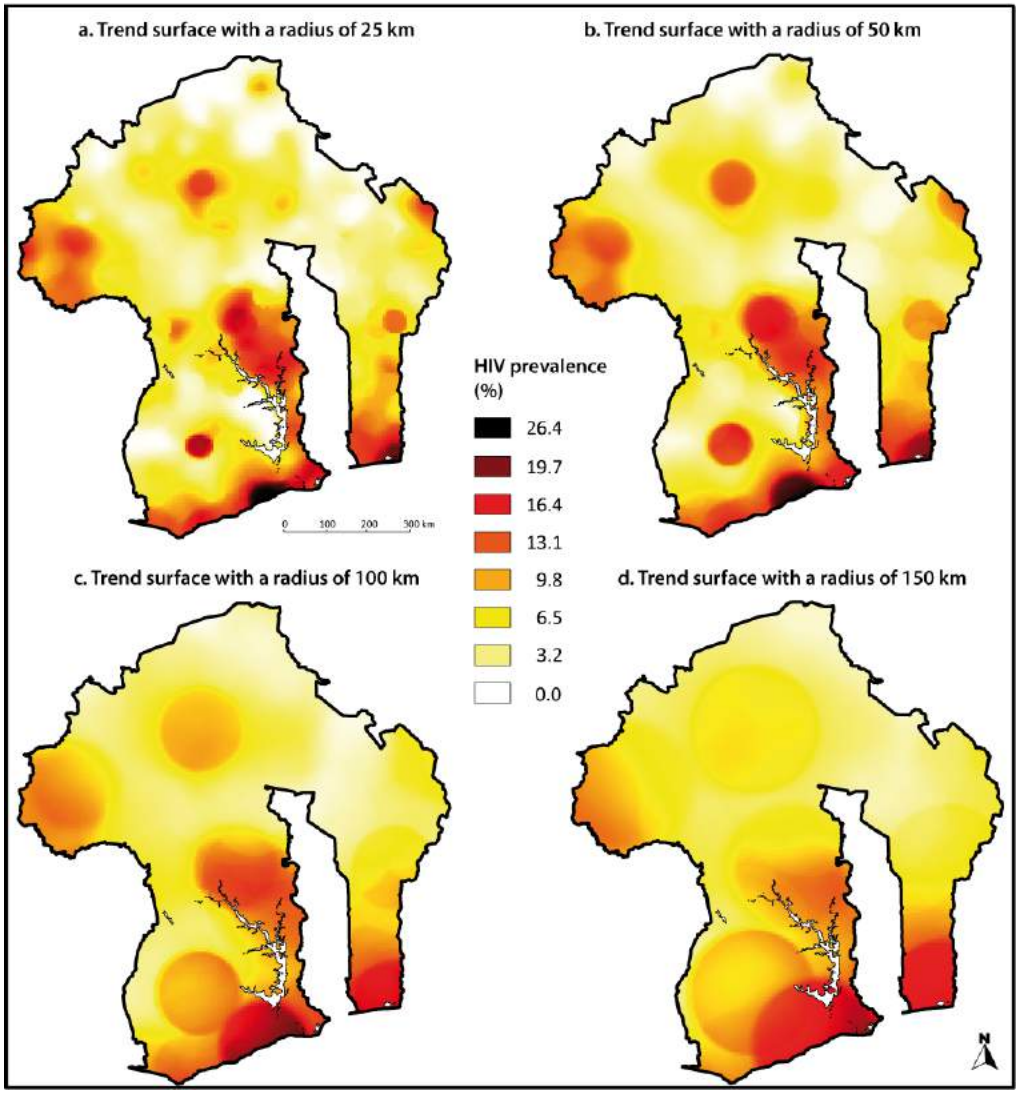
72 This result evokes the trend surface analysis (Chorley and Haggett, 1965; Griffin, 1949; Haggett, 1973; Krumbein, 1956) developed in the mid-20th century. These cartographic filtering techniques are applied to known surfaces and are designed to decompose the surface of the studied phenomenon into the sum of a surface of regional trends and a surface of local residuals. The regional trend surface is calculated so as to filter out the local details and reveal the main variations in the phenomenon. The local residual surface is simply the difference between the raw data surface and the regional trend surface.

73 One cartographic filtering method for calculating regional trends is known as the "ring" method (Griffin, 1949; Krumbein, 1956; Nettleton, 1954) similar to a moving spatial mean. For each data point a circle of set radius is drawn around that point and the indicator is calculated for the area within that circle.

74 Figure 12 shows various trend surfaces calculated from the Model 2 prevalence surface by the ring method with radii from 25 to 150 kilometres. They gradually reveal the model's underlying spatial structure.

75 The prevalence surface estimated from a DHS simulation (Figure 11) is closer to these trend surfaces than to the Model 1 prevalence surface. The MISD between the estimated surface and the Model 2 surface is 10.3, but 9.1 (other values 7.5, 6.9 and 10.6) between the estimated surface and trend surfaces using a radius of 25 kilometres (50, 100 and 150 kilometres).

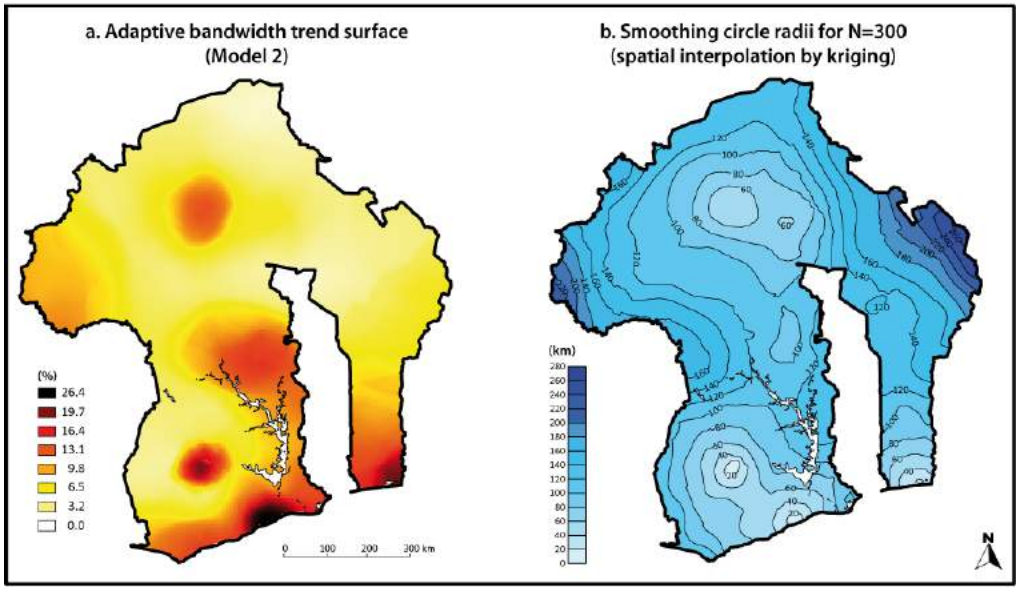
Figure 12: Trend surfaces calculated from Model 2 with rings of radius 25, 50, 100 and 150 kilometres



76 The smoothing circles used to define the value of the bandwidths in the kern estimator approach are similar to those in the "ring" method, but use adaptive bandwidths of equal number of persons surveyed and not set radii. This adaptive bandwidth concept can be

applied to calculating a trend surface. For that the radius of the rings to be used must be determined for each data point. Using the survey scatter plots in the DHS simulation for which a smoothing radius has been determined (for a given value of *N*), a radius surface is generated by spatial interpolation (Figure 13.b). An adaptive bandwidth trend surface (Figure 13.a) is then calculated using for each data point a bandwidth defined by the radius surface. The result obtained looks rather like the prevalence surface estimated by the kernel estimator approach (Figure 11). The MISD between the two surfaces is only 6.0.

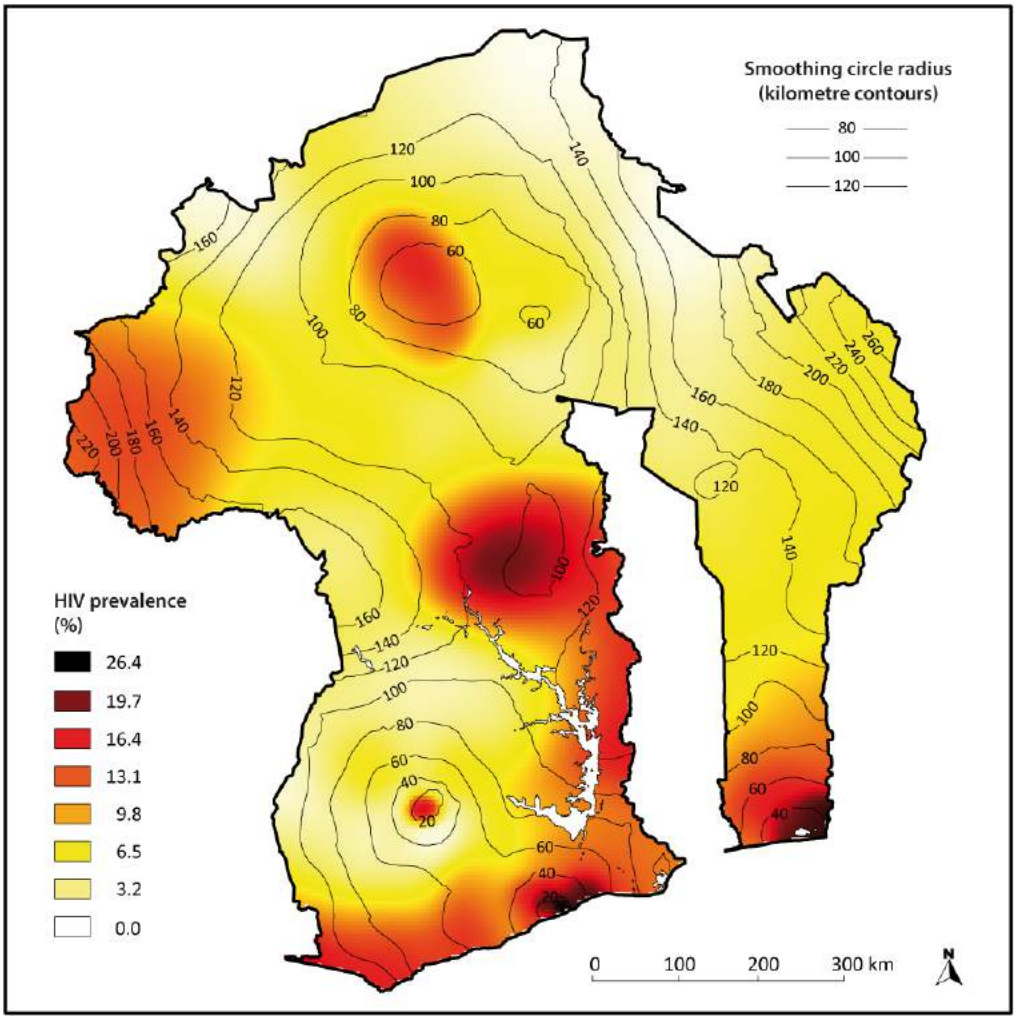
Figure 13: Trend surface using adaptive bandwidths and spatial interpolation of the smoothing circle radii (N=300)



77 The kernel estimator approach with adaptive bandwidths of equal number of persons surveyed cannot reproduce local variations on the model's prevalence surface. Some loss of information is inevitable because of the DHS sampling design. This approach cannot reproduce the actual prevalence surface.

78 However, the estimated prevalence surface does reflect certain real features of epidemics and is similar to an adaptive bandwidth trend surface in revealing underlying regional variations in the phenomenon. It is therefore important to take into account the variations in smoothing circle radii when interpreting the estimated prevalence surface. In order to facilitate interpretation, it is possible to display both datasets by overlaying the contours[23] for smoothing circle radius values on the prevalence surface (Figure 14).

Figure 14: Prevalence surface estimated by the kernel estimator approach (N=300) from a DHS simulation on Model 2 with contours of smoothing circle radius values

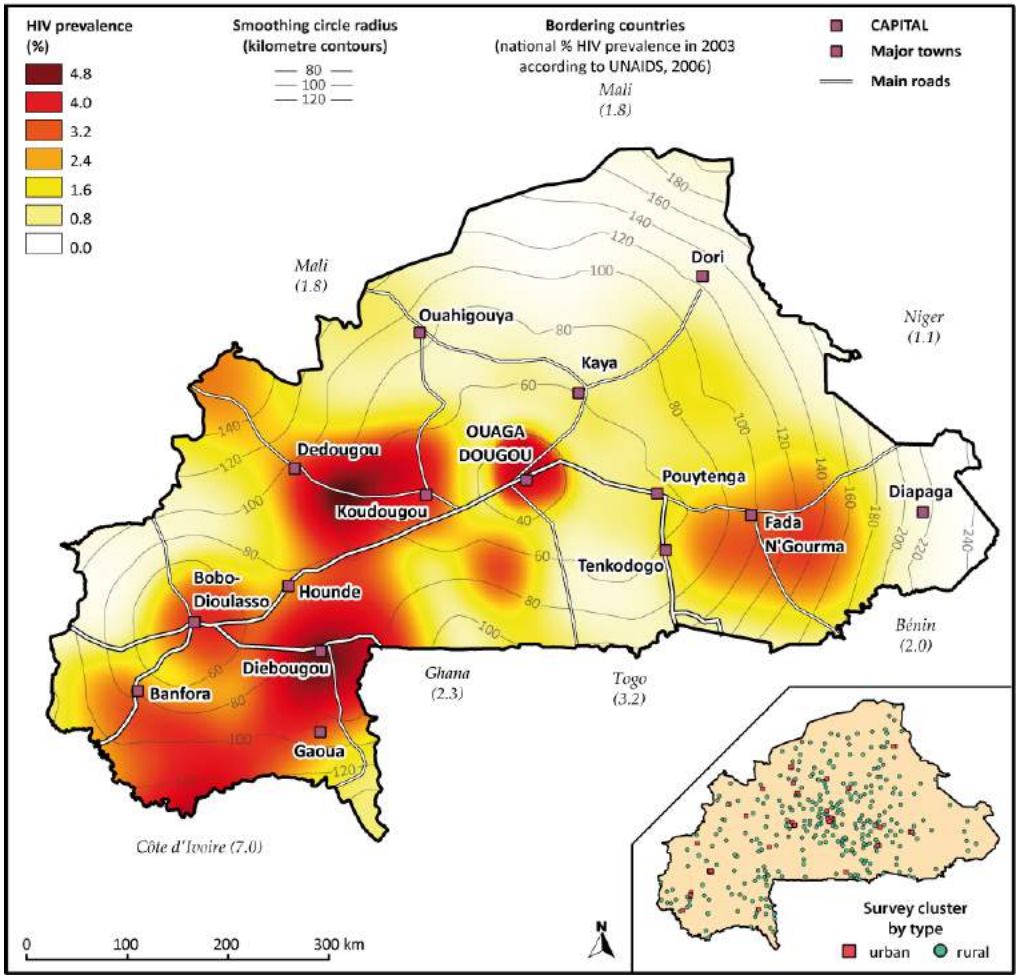


Application to real data: the Burkina Faso 2003 DHS

79 We applied the kernel estimator approach with adaptive bandwidths of equal number of persons surveyed to the Burkina Faso 2003 DHS data. The survey tested 7,244 people (aged 15-49) in 400 clusters. The national HIV prevalence measured in the survey was 1.8% (see Table 1).

80 The distribution of survey clusters is shown in an insert in Figure 15. The main surface is an estimated prevalence surface with parameter N=500, a value chosen using Equation 5[24]. National HIV prevalence values at end 2003 are also given for Burkina Faso's bordering countries, according to UNAIDS estimates (2006). The smoothing circle radius contours and Burkina Faso's main roads and conurbations have also been added.

Figure 15: Regional HIV prevalence (age 15-49) estimated by kernel estimator approach (N=500) applied to the Burkina Faso 2003 DHS



81 The prevalence surface estimated from the DHS is relatively consistent. First, the epidemic is worse in the south-western part of the country, near high-prevalence countries (Côte d'Ivoire, Ghana) than in the Sahelian north-east, sparsely populated and bordering low-prevalence countries (Mali, Niger). The worst affected regions are mostly around the main conurbations (Ouagadougou, Bobo-Dioulasso) and along the main roads to Côte d'Ivoire and Mali (via Dedougou). The region between Diebougou and Gaoua, where the epidemic has a local peak, is known for its gold mining, involving a relatively high number of seasonally migrant males and a non-negligible number of sex workers. The regions in the south-west are also those that received most returnees from Côte d'Ivoire at the end of 2002 and beginning of 2003 (SP/CONASUR, 2004).

82 This association of high-prevalence areas, areas of migration, gold mining and main urban centres cannot be used to establish a link between the phenomena, since mere geographical proximity does not count as evidence. However, these results are consistent with Georges Rémy's research (1999) in sub-Saharan Africa showing that "*cities, particularly the largest ones, are especially exposed to infection at all stages of the disease… But rural sites are also vulnerable. They are marked by various cash-earning businesses: mining centres, road stops, agri-food areas, markets.*"[25]

Conclusion

83 The samples of Demographic and Health Surveys are designed to achieve a certain accuracy in calculating HIV prevalence nationally and regionally. The numbers of persons surveyed, however, are too small to enable accurate estimations of prevalence in each survey cluster or at a finely grained local level.

84 The use of adaptive bandwidths of equal number of persons surveyed makes it possible to achieve a smoothing effect that adapts to the high irregularity of spatial distribution among the survey clusters, selecting the clusters according to population distribution. The surfaces thus generated are relatively accurate for densely populated areas and strongly smoothed in sparsely surveyed areas.

85 The DHS simulation of a fictitious country revealed that the kernel estimator approach with adaptive bandwidths of equal number of persons surveyed was more effective for this purpose than a spatial interpolation of prevalence values previously smoothed with the same smoothing circles. Similarly, there is no significant gain from adding a maximum radius as a second smoothing criterion.

86 Model 2 showed that, although local variations in the epidemic were filtered out by this type of technique, the regional component in the spatial variation of prevalence was generally reproduced, and the estimated prevalence surfaces could be interpreted as regional trend surfaces with adaptive bandwidths. A surface of this sort, by construction, is necessarily spatially continuous and self-correlated and in no way implies any potential discontinuities and local variations in the real surface of the epidemic, which remains inaccessible in the DHS data.

87 The disadvantage of this approach is that it does not provide an automated technique for selecting an optimal value for the smoothing parameter. Although some specialist research is being done in this field, relatively little of it concerns the ratio between two density surfaces. We tested one of these approaches, adapted from Davies and Hazelton (2010). It was ineffective for our purposes. DHS data have the particularity that the location of observed data is based on a two-level cluster sampling and not a one-level random sampling, since the scatter plot of positive cases is not in practice independent of the scatter plot of observed cases.

88 However, although Equation 5 is determined from simulations of the fictitious country and cannot strictly be generalised to other situations, it does provide an order of magnitude for the parameter to be used in practice. Application to Burkina Faso's 2003 DHS data produced a plausible map of regional prevalence values.

89 Although a map of this sort must be interpreted with caution, it does provide a descriptive indication of the state of the epidemic in a country independent of administrative divisions. It is a useful tool for displaying the main spatial variations of the phenomenon and identifying the worst affected regions. Although DHSs are insufficient for analysing the spatial determinants of HIV epidemics, they do make it possible to sketch out a preliminary picture in the absence of more specific surveys with better geographical coverage.

Interpretation: for 60% of clusters, the radius of the smoothing circles is equal to or less than 87 kilometres.

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. 1/2, 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 geóstatistique 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

Appendix Table 1: MISD for various values of *N* for the spatial interpolation approach and kernel estimator approach

N	Spatial interpolation	Kernel estimators
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

Appendix Table 2: Quantiles of smoothing circle radii with N=300

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

Notes

1 In epidemiology, the prevalence of a disease may be expressed in either absolute terms (number of cases) or relative terms (proportion of people infected in the study population). Throughout this paper we shall systematically use "prevalence" in its relative sense.

2 Joint United Nations Programme on HIV/AIDS.

3 In the sense of major administrative divisions within a country.

4 The sentinel surveillance of pregnant women consists of selecting antenatal clinics throughout a country and taking and testing blood samples from each woman attending her first antenatal appointment. These relatively cheap and easily implemented surveys have been held each year in a number of countries since the late 1980s and early 1990s.

5 “Ordinary” households exclude people living in institutions (prison, hospital, barracks, boarding school, convent, etc.). The definition of an ordinary household may vary from one census to another.

6 http://www.measuredhs.com/aboutsurveys/gis/methodology.cfm, web page consulted 15 September 2010.

7 http://www.measuredhs.com/aboutsurveys/ais/start.cfm, web page consulted 15 September 2010.

8 http://www.measuredhs.com/pubs/articles/start.cfm?selected=2, web page consulted 15 September 2010.

9 http://www.hivmapper.com/

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

11 This technique is explained below.

12 More precisely, the interpolated prevalence at the centroid of each primary unit was applied to the entire primary unit.

13 Because of rounding in the calculation of the number of clusters to be drawn from a stratum, the total may differ slightly (± 1) from the target figure.

14 The number of persons tested per cluster is not spatially self-correlated (Larmarange, 2007).

15 Taking into account the relative weighting of each individual.

16 Terrell’s Maximal Smoothing Principle for fixed-bandwidth kernels consists of using the largest bandwidth from a set of optimal bandwidths estimated from the observed variance of the sample.

17 The Gaussian kernel produces a density surface covering the entire surface ($\forall$ (x,y), K(di/hi) > 0), whereas finite extent kernels have a nil density outside the bandwidth (di > hi $\Rightarrow$ K(di/hi) = 0).

18 This equation differs from the MISE one only in comparing two known surfaces rather than an estimated surface and an unknown density surface.

19 On line at http://www.ceped.org/prevR.

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

21 Obtained by adding together Model 1’s prevalence surface and the variation surface.

22 With the same simulation parameters: national prevalence 10%, 8,000 people surveyed in 400 clusters.

23 Isolines calculated from the radius surface obtained by spatial interpolation of the radii of the smoothing circles for the survey clusters.

24 Equation 5 applied to the 2003 DHS data provides an optimal value for N of 502, which we rounded to 500.

25 Original text in French: “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.”

Pour citer cet article

Référence électronique

Joseph Larmarange, Roselyne Vallo, Seydou Yaro, Philippe Msellati et Nicolas Méda, « Methods for mapping regional trends of HIV prevalence from Demographic and Health Surveys (DHS) », *Cybergeo : European Journal of Geography* [En ligne], Systèmes, Modélisation, Géostatistiques, article 558, mis en ligne le 26 octobre 2011, consulté le 12 octobre 2012. URL : http://cybergeo.revues.org/24606 ; DOI : 10.4000/cybergeo.24606

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

In many countries, particularly in sub-Saharan Africa, Demographic and Health Surveys (DHSs) are the main way of estimating HIV prevalence nationally in the general population. Some DHSs record the longitude and latitude of the survey clusters.
We present three methodological approaches for mapping spatial variations in HIV prevalence using the DHSs. These approaches are applied to simulated DHS samplings from a model country. The estimated surfaces are then compared with the model's initial surface.
We demonstrate that a method using kernel estimators with adaptive bandwidths size of equal number of persons observed can be used to estimate the main regional trends in epidemics. Application to Burkina Faso's 2003 DHS data provides a plausible image of that country's epidemiological situation.

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.

Entrées d'index

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

Notes de l'auteur

This research received financial support from the French National Agency for Research on AIDS and Viral Hepatitis (Project ANRS 12114).

Cybergeo : European Journal of Geography

Systèmes, Modélisation, Géostatistiques

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

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

Référence électronique
Joseph Larmarange, Roselyne Vallo, Seydou Yaro, Philippe Msellati et Nicolas Méda, « Methods for mapping regional trends of HIV prevalence from Demographic and Health Surveys (DHS) », *Cybergeo : European Journal of Geography* [En ligne], Systèmes, Modélisation, Géostatistiques, article 558, mis en ligne le 26 octobre 2011, consulté le 12 octobre 2012. URL : http://cybergeo.revues.org/24606 ; DOI : 10.4000/cybergeo.24606

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

Document accessible en ligne sur :
http://cybergeo.revues.org/24606
Document généré automatiquement le 12 octobre 2012.