# Estimating spatial sub-national variations from Demographic and Health Surveys data

Joseph Larmarange[a], Roselyne Vallo[b], Seydou Yaro[c], Philippe Mselatti[d], Nicolas Méda[c], Benoît Ferry[a]

**Session:** Spatial demography (2101)

a. IRD - Centre Population et Développement (CEPED) • UMR 196 Paris Descartes INED IRD
Mails : joseph.larmarange@ceped.org
CEPED 221 boulevard Davout 75020 Paris France
Téléphone : +33 1 78 94 98 70 • Fax : +33 1 78 94 98 79
b. Université de Montpellier I •EA 4205
c. Centre Muraz • Bobo-Dioulasso – Burkina Faso
d. IRD • CrecSS (Centre de Recherche Cultures Santé Sociétés)/IFEHA • Université Paul Cézanne

## Short abstract

For many countries, in particular in sub-Saharan Africa, Demographic and Health Surveys (DHS) are the only national source of data (depending of the subject). Several DHS collect latitude and longitude of surveyed clusters but the sampling method is not appropriate to derive local estimates: sample size is not large enough for a direct spatial interpolation.

We develop in this paper a new approach for estimating a proportion for each sample cluster by aggregating data from neighbouring clusters. This estimated proportion can then be interpolated by kriging method. Estimation parameters were computed from 24500 survey simulations on a model country.

This approach allows estimating regional trends of a phenomenon under the assumption that it is spatially continuous. The method was developed to map HIV prevalence in Burkina Faso and Cameroon at sub-national level but it can be applied to any other proportion. Our results will be compared with maps by DHS regions.

## Extended abstract

We developed a methodological approach inspired by techniques used for calculating regional trends [1-3] and based on rings of the same number of observed persons. The main idea consists to decompose spatial variations as the sum of regional trends and local residuals, taking in count a random error. By aggregating neighbouring clusters, regional trends will be estimated.

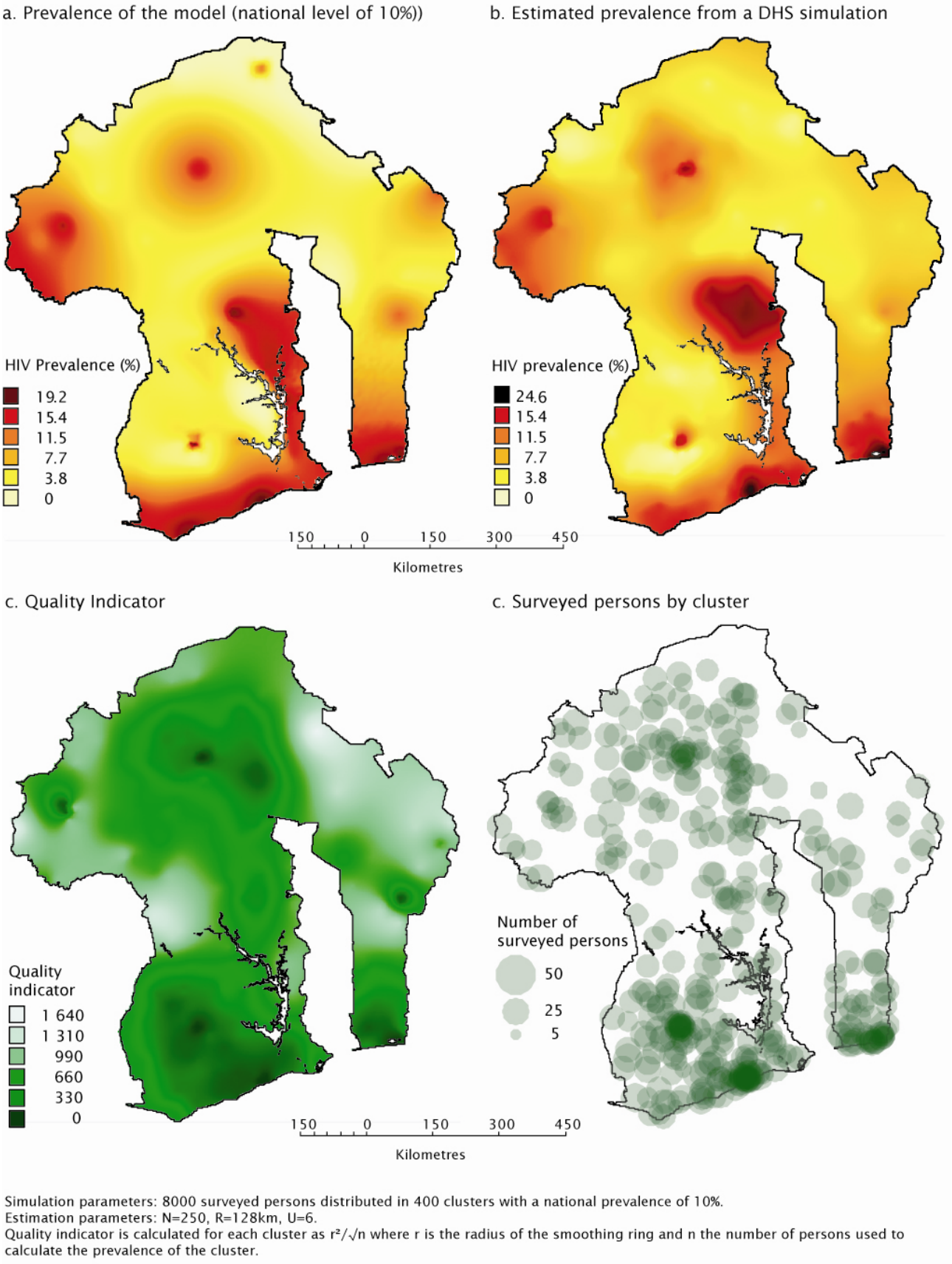
We conduct all these analyses using the free and open-source statistical software R [4]. A specific package called prevR was written and can be downloaded at http://www.ceped.org/prevR/.

Boundary files from the Digital Chart of the World (DCW)[5] and geo-localisation of main cities from the Global Rural-Urban Mapping Project (GRUMP) [6] were used.

### *DHS simulation and model country*

To test our methodology, we elaborated a model country where it was possible to simulate DHS and to compare epidemic of the model and reconstituted epidemic. This country was built by aggregating Benin, Burkina Faso and Ghana. Population density in 2000 and urban extents from the GRUMP [7, 8] were used and the country was divided into 11 regions and 9137 clusters. A fictive epidemic was applied to the model and was elaborated in order to present different models of diffusion: cities with concentrated or diffuse source of infection; local rural source of infection; discontinuity from one side of a lake to the other; gradient from coast to earth or from a border country (see figure 1.a).

**Figure 1.** HIV prevalence of the model, estimated prevalence from a DHS simulation and complementary maps




Simulation parameters: 8000 surveyed persons distributed in 400 clusters with a national prevalence of 10%.
Estimation parameters: N=250, R=128km, U=6.
Quality indicator is calculated for each cluster as $r^2/\sqrt{n}$ where r is the radius of the smoothing ring and n the number of persons used to calculate the prevalence of the cluster.

DHS were simulated with three parameters: total number of tested persons, number of clusters and level of national prevalence. The simulations reproduce a stratified two-stage sampling. Clusters are selected randomly by stratum with a probability proportional to the cluster's population. The number of tested persons is randomly calculated to reproduce variations in household size observed in DHS. Lastly the

number of positive persons by cluster is determined by a binomial rule. Sampling weights are then calculated for each surveyed persons.

### *Estimating the prevalence of each cluster*

First, a ring was drawn around each cluster and then the prevalence of the central cluster was estimated from all the clusters inside the ring. Insofar as the sub-sample size is meaningful for prevalence calculation, rings with the same number of observations were used. Once a number *N* is selected, the ring radius of a cluster is determined so that the number of tested persons located inside that ring is at least equal to *N*. The prevalence of the central cluster is then calculated from all those observations.

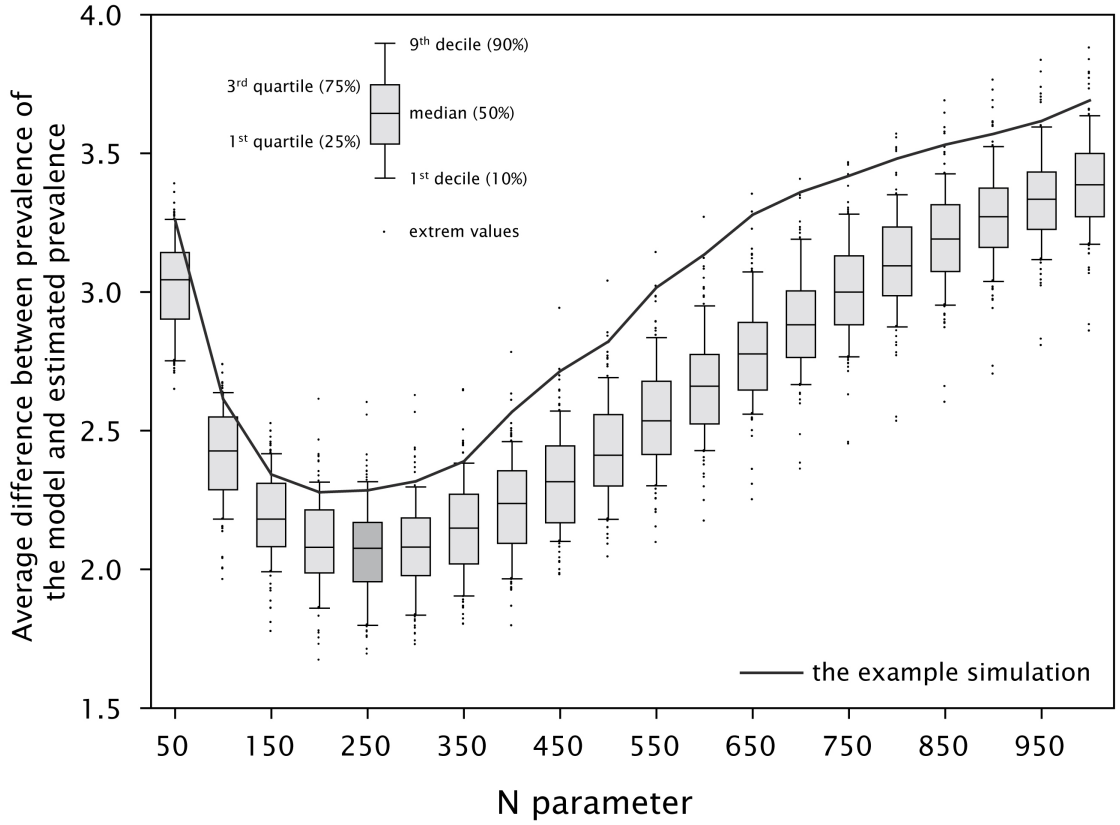
Using a minimum number of observations as a parameter to determine rings size simultaneously led to an estimation of prevalence from enough observations and application of a different smoothing level (corresponding to the ring area) depending on the density of tested persons.

As the *N* parameter increased gradually, the estimated prevalence was calculated on more observations and smoothed, attenuating random sample errors. At the same time, they tended progressively towards a single value (standardization effect). It was then necessary to find a compromise that sufficiently minimized random variations and preserved local precision of regional trends.

In the model, it is possible to calculate, for one simulation and one value of *N*, difference for each cluster between the estimated prevalence and the prevalence of the model and then the average difference. For one simulation, when *N* increases, the average difference decreases to a minimum and then increases (see the curve on figure 2).

In proportion to little variations from a simulation to another, an optimal value for *N* is defined as the median, on 100 simulations with the same parameters, of the value of *N* for which the average difference is minimal (see box plots on figure 2).

**Figure 2.** Average deviation by N parameter for 100 DHS simulations and the example simulation




Calculating $N_{optimal}$ for a set of simulation parameters requires a long calculation time (around 12 hours). 24500 DHS have been simulated with a national prevalence varying from 1% to 45% and a number of tested persons from 5000 to 14930. The number of clusters has a little effect on $N_{optimal}$. If the number of tested

persons is higher than 5000 persons (higher than 7000 if national prevalence less than 2% and higher than 8500 if national prevalence less than 1%), the value of $N_{optimal}$ can be modelled as a simple equation:

$$N_{optimal} = 14.172 \cdot sample\_size^{0.419} \cdot national\_prevalence^{-0.361} \cdot clusters\_number^{0.037} - 91{,}011$$

These simulations show also that when national prevalence increases, $N_{optimal}$ decreases, prevalence estimations needing less observations to compensate sampling variations and being more accurate. When sample size increases, $N_{optimal}$ increases too but, in the same time, the average radius of the smoothing ring decreases, estimations being more localised.

In lightly surveyed areas, particularly along national borders with a small population, prevalence estimates are based on clusters that are very distant from each other. So it is better to limit smoothing to smaller rings, even though prevalence estimates will be calculated on fewer observations. The maximum radius imposed on rings is called *R*. To modify estimations only for eccentric clusters, we choose to use the 9th decile of the ring radius when only the *N* parameter is applied.

Lastly, many studies have shown differentials between urban and rural areas.[9, 10] The gradual spread of an epidemic can be observed around a city or, conversely, its concentration in an agglomeration. It is therefore appropriate to consider whether a cluster is located inside an urban agglomeration or not. Being inside an urban agglomeration is not documented in DHS, so we used information provided by the GRUMP [6].

The prevalence for clusters outside urban agglomerations will be therefore only calculated on clusters of the same type. On the other hand, those clusters inside an agglomeration will only be estimated on clusters in the same urban agglomeration. It is therefore essential for selected urban agglomerations to have been sufficiently surveyed in the DHS. So only urban agglomerations with enough observations and distributed on several clusters should be selected. However, if another source of data is available (like sentinel surveillance of pregnant women by example), it is possible to select agglomerations which, although their number of observations is lower, provide in DHS a prevalence similar to the prevalence observed in the other dataset. A discussion about the selection of urban agglomerations with be presented with the example of Burkina Faso and Cameroon. The number of selected urban agglomerations is called *U*.

### *Spatial interpolation*

After estimating HIV prevalence for each cluster, maps of spatial variations in HIV prevalence were obtained by spatial interpolation (see figure 1.b). We use ordinary kriging [11], a sophisticated method that interpolates across space according to a spatial lag relationship that has both systematic and random components. That method takes into account the spatial dependence structure of the data and is indicated when the spatial distribution of observed points is irregular.

### *Complementary maps*

The quality of prevalence estimates was not constant throughout the country. In populated areas, information was accurate because of good cluster density. Intraregional variations in the epidemic were visible, but in areas where the number of people tested was smaller, only regional trends could be seen. Lastly, for non-surveyed areas, the results needed to be interpreted with caution, as the estimated variations resulted from interpolation based on clusters from neighbouring areas. The irregularity of estimate quality came from DHS sampling, designed to be representative of populations and not of territories, and from our methodological approach, which consisted of adaptive smoothing according to observation density.

Two complementary maps were produced to facilitate results interpretation. The first represented the number of people tested per cluster with transparent circles. Pale areas correspond to areas that were briefly documented or not at all (see figure 1.d). The second was obtained by spatial interpolation of a quality indicator (see figure 1.d). The value of this indicator is calculated for each cluster as $R_c^2 / \sqrt{N_c}$ where $R_c$ is the radius of the smoothing ring used to estimate the prevalence of the cluster and $N_c$ the number of observations included in this ring. A low value of this quality indicator indicates a good estimation. No confidence intervals were calculated according to the fact that confidence intervals depend in the same time

of the number of observations and of the level of prevalence. So, variations of magnitude of confidence intervals represent more variations of prevalence than variations of estimation quality. Furthermore, confidence intervals do not take into account the radius of smoothing circle.

## Results

Figures 1.a and 1.b are presented with the same colorimetric scale. However, the maximum of the estimated prevalence (figure 1.b) is higher than the one of the model (figure 1.a). Variations of estimated prevalence are more contrasted.

Globally, main patterns of the model epidemic have been reconstituted. The gradients from the south coast to the north and from the east frontier to the west are visible. The two main urban agglomerations at the east on the south coast, included for $U$ parameter, present a higher contrast, but the small cities at the west of the coast, that were not included in the $U$ parameter, cannot be distinguished from their neighbourhood. Concentrated or diffuse epidemics around cities included for $U$ parameter are correctly reconstituted although differences are slightly smoothed and more irregular. So, taking into account main urban agglomerations with $U$ parameter allows the highlighting of spatial variations of epidemic in their neighbourhood.

From one side of the lake to the other side, where a discontinuity was introduced in the model, prevalence has been overestimated in the west and underestimated in the east. The estimation procedure not taking into account natural frontiers and considering geographical surface as a spatial continuum, a making uniform effect is observed. However, main differences are still visible.

High prevalence area at the north of the lake is reconstituted but contrasts are higher than those of the model. The small city in this area, not surveyed enough to be included for $U$ parameter, cannot be distinguished, only sub-regional prevalence being visible.

Several elements have not been reproduced on figure 1.b. First, the local source of infection in rural area in the north of the country is not visible, according to the fact that this zone is not enough peopled to have been surveyed. Gradients from the east frontier and around the small city at hundred kilometres in the south are not clearly visible. The number of clusters in this region is low and the value of the quality indicator high. The number of observations is not enough to reconstitute with accuracy the local variations and only a regional trend can be estimated (making uniform effect).

Figure 1.c and 1.d are similar. It is possible to distinct areas not surveyed where estimated prevalence are continuity of regional trends; areas few surveyed where only sub-regional trends are reconstituted, local variations becoming uniform; and areas with enough observations to highlight local differentials. Low values of the quality indicator correspond effectively to the areas where the reconstitution was good.

## Discussion

Globally, this methodology estimates national, regional and sub-regional variations, with a more or less important smoothing according to the accuracy of the data of each zone, although a precise estimation of levels of each point is not possible. Analysing differentials is thus valid at regional or even sub-regional level, but is not meaningful at a very local level. Estimations are more accurate when the total number of tested persons and the national prevalence are higher. At contrary, estimations become uniform in not enough populated areas.

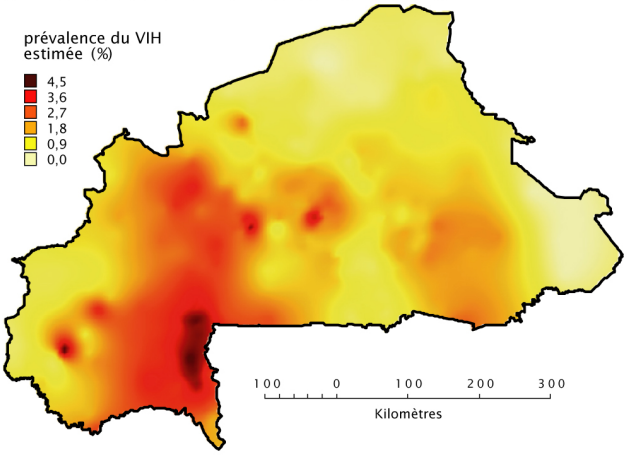
DHS are designed to be representative of populations and not of territories. Spatial repartition of surveyed persons is not random but reflects population density. The methodology developed here rests on two elements. First, estimation of prevalence of each cluster with rings of same number of observations compensate ales of the second-stage sampling (selection of eligible households). Secondly, spatial interpolation by kriging compensates the first-stage sample (selection of clusters).

Rings of the same number of observations make possible to realise a spatial interpolation, take in count the spatial diversity of the accuracy of the data and highlight a maximum of the spatial variations which can be reproduced with incomplete data. However, this methodology is limited by DHS design, making impossible to estimate very local variations.

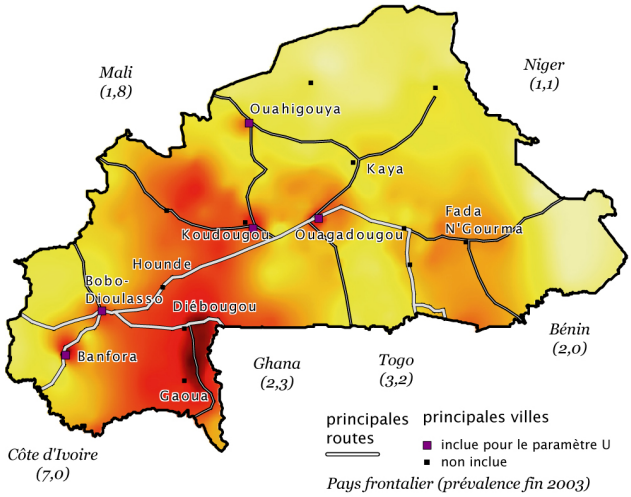
Application of this methodology to data from 2003 Burkina Faso DHS and 2004 Cameroon DHS have produced coherent maps with the other data sources as migrations for example (detailed results will be presented – see figure 3 for application to Burkina Faso and figure 4 for comparison with a map by DHS region).

**Figure 3.** Application to HIV prevalence in Burkina Faso (DHS 2003)

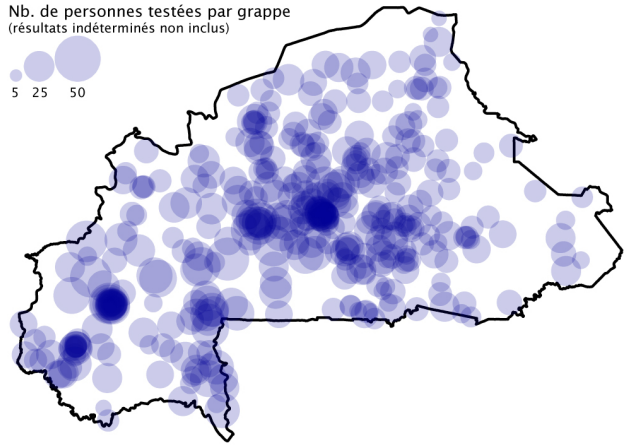
a. Prévalence estimée du VIH (15-49 ans)



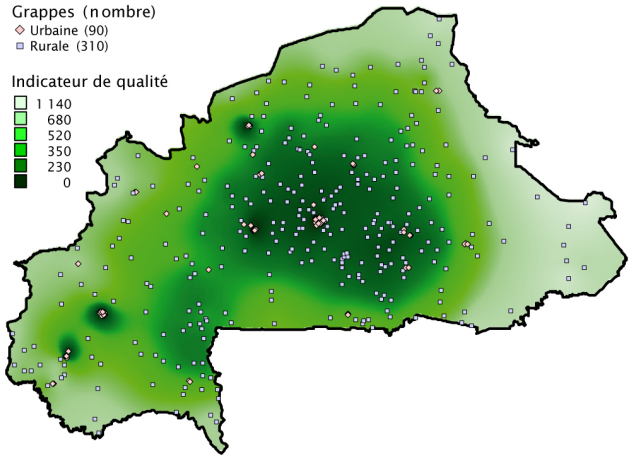

b. Prévalence estimée du VIH, villes et routes




c. Nombre de personnes testées par grappe



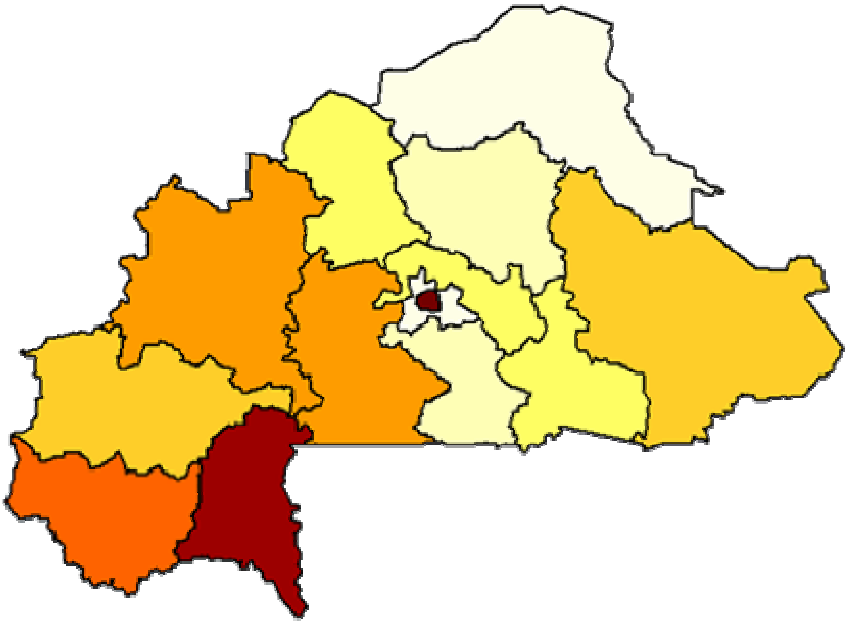

d. Indicateur de qualité et grappes enquêtées




Note 1 : La prévalence nationale du VIH (15-49 ans) est de 1,8% dans l'EDS 2003 du Burkina Faso. 7.244 personnes ont été testées (résultats indéterminés exclus), réparties en 400 grappes.
Note 2 : Les paramètres utilisés pour l'estimation de la prévalence du VIH sont : N=500, R=117 km et U=5 (voir carte b pour les villes retenues).

Note 3 : L'indicateur de qualité est calculé pour chaque grappe selon l'expression $r^2/\sqrt{n}$ où r est le rayon du cercle de lissage et n le nombre de personnes testées dans ce cercle.
Sources : EDS 2003 du Burkina Faso pour les données VIH, DCW pour les frontières nationales, GRUMP pour les principales villes, ArcAtlas (ESRI) pour les principales routes, rapport ONUSIDA 2006 pour les prévalences fin 2003 des pays frontaliers.

**Figure 4.** HIV prevalence in Burkina Faso (DHS 2003) by DHS region



## *References*


1. Chorley RJ, Haggett P. **Trend-Surface Mapping in Geographical Research**. *Transactions of the Institute of British Geographers* 1965:47-67.
2. Griffin WR. **Residual Gravity in Theory and Practice**. *Geophysics* 1949,14:39-56.
3. Krumbein WC. **Regional and local components in facies maps**. *AAPG Bulletin* 1956,40:2163-2194.
4. R Development Core Team. **R: A Language and Environment for Statistical Computing**. In. Vienne (AT): R Foundation for Statistical Computing; 2006.
5. Environmental Systems Research Institute Inc. (ESRI). **Digital Chart of the World (DCW) data set**. In: Pennsylvania State University; 1996.
6. Center for International Earth Science Information Network (CIESIN) of Columbia University, International Food Policy Research Institute (IFPRI), The World Bank, Centro Internacional de Agricultura Tropical (CIAT). **Global Rural-Urban Mapping Project (GRUMP), Alpha Version: Settlement Points**. In. Palisades, New York (US): Socioeconomic Data and Applications Center (SEDAC) of Columbia University; 2004.
7. Center for International Earth Science Information Network (CIESIN) of Columbia University, International Food Policy Research Institute (IFPRI), The World Bank, Centro Internacional de Agricultura Tropical (CIAT). **Global Rural-Urban Mapping Project (GRUMP), Alpha Version: Population Density Grids**. In. Palisades, New York (US): Socioeconomic Data and Applications Center (SEDAC) of Columbia University; 2004.
8. Center for International Earth Science Information Network (CIESIN) of Columbia University, International Food Policy Research Institute (IFPRI), The World Bank, Centro Internacional de Agricultura Tropical (CIAT). **Global Rural-Urban Mapping Project (GRUMP), Alpha Version: Urban Extents**. In. Palisades, New York (US): Socioeconomic Data and Applications Center (SEDAC) of Columbia University; 2004.
9. Mishra V, Vaessen M, Boerma JT, Arnold F, Way A, Barrere B, *et al.* **HIV testing in national population-based surveys: experience from the Demographic and Health Surveys**. *Bull World Health Organ* 2006,84:537-545.
10. Asamoah-Odei E, Garcia Calleja JM, Boerma JT. **HIV prevalence and trends in sub-Saharan Africa: no decline and large subregional differences**. *Lancet* 2004,364:35-40.
11. Krige D. **A statiscal approach to some basic mine valuation problemens on the witwatersrand**. *Journal of the Chemical, Metallurgical and Mining Society* 1951,52:119-139.