Mapping Demographic and Health Surveys (DHS): a method to estimate regional trends of a proportion

A previous version of this work was presented in Marrakech in 2009 at XXVI IUSSP International Conference (http://iussp2009.princeton.edu/abstracts/92702). The methodological approach presented here has been completely updated using now kernel density estimators instead of kriging interpolation.

Context and objective

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, 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.

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. Some DHSs also use GPS to collect the geographical coordinates of each survey cluster.

Although there are many publications concerning the DHSs, spatial analysis based on them are less numerous. Of more than 1154 scientific papers identified on the MeasureDHS website as relating to DHS data, only 15 are classified under "Spatial analysis"[1]. Most of them are atlases of choropleth maps (TACAIDS, 2006); national or regional maps, as with the online tool HIVmapper[2]; 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[3].

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.

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.

Methods

The 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).

[1] http://www.measuredhs.com/Publications/Journal-Articles-by-Country.cfm?selected=2, web page consulted 5 October 2012.

[2] http://www.hivmapper.com/

[3] HIV Spatial Data Repository: http://www.hivspatialdata.net/

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.

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.

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.

The use of fixed bandwidths is inappropriate for DHS, since the clusters are very unevenly distributed. A sufficiently large radius needs to be determined for estimating proportions 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 bandwidths not of equal radius but of equal number of persons surveyed.

We therefore developed an 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 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. For the positive cases we apply the same bandwidth as that calculated for the control cases in the same cluster.

A number of density functions may be used for kernel. 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[4] (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.

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

Testing the method

In order to test our approach, 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.

Benin, Burkina Faso and Ghana were combined to create a fictitious country to be used as a model. Data from the Global Rural-Urban Mapping Project (GRUMP) were used to distribute the population over the territory. The territory was then divided into 9,137 primary units (7,818 rural and 1,319 urban) and 11 administrative units. We created a prevalence surface by spatial interpolation from points chosen *ad hoc* for the surface to present various diffusion patterns.

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.

Figure 1

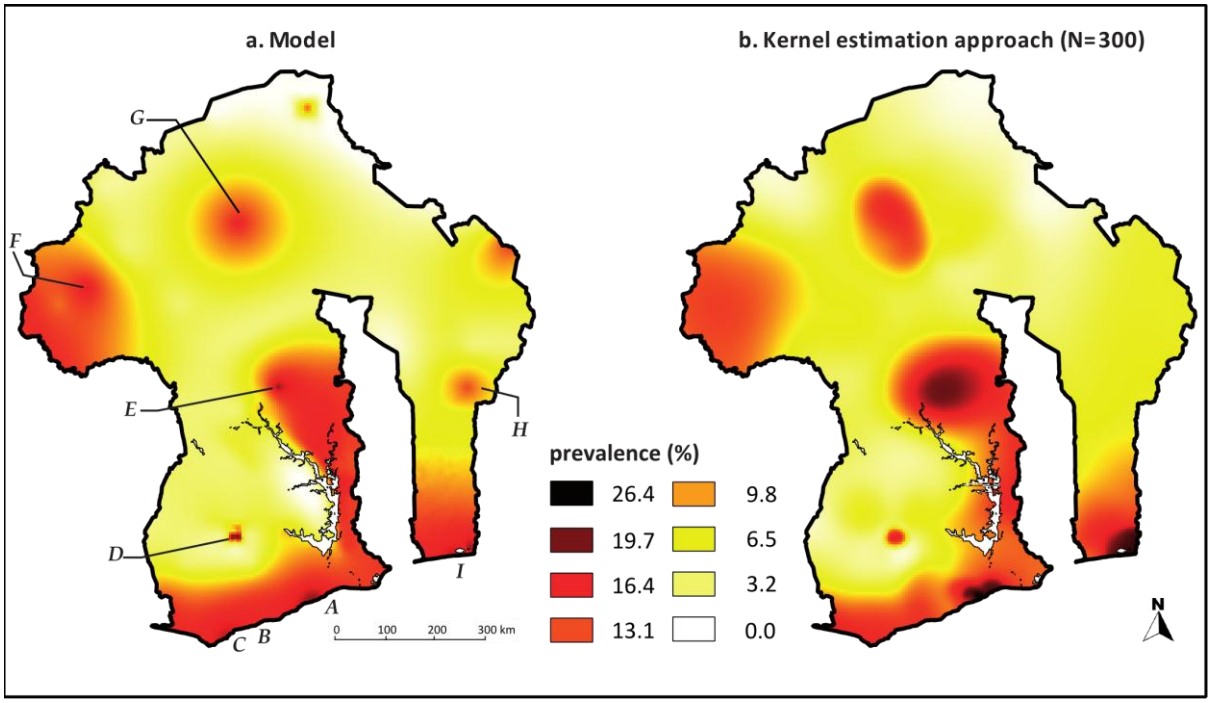


Figure 1.a represents the prevalence surface created for the model (national prevalence of 10%). A DHS was simulated with a sample size of 8,000 people distributed in 400 clusters. Figure 1.b represents the result of applying our method on the data obtained from this simulation.

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 for conurbations A and I. Conurbation D still shows a more concentrated prevalence than its vicinity. The kernel estimators 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. 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.

Implementation in R: prevR

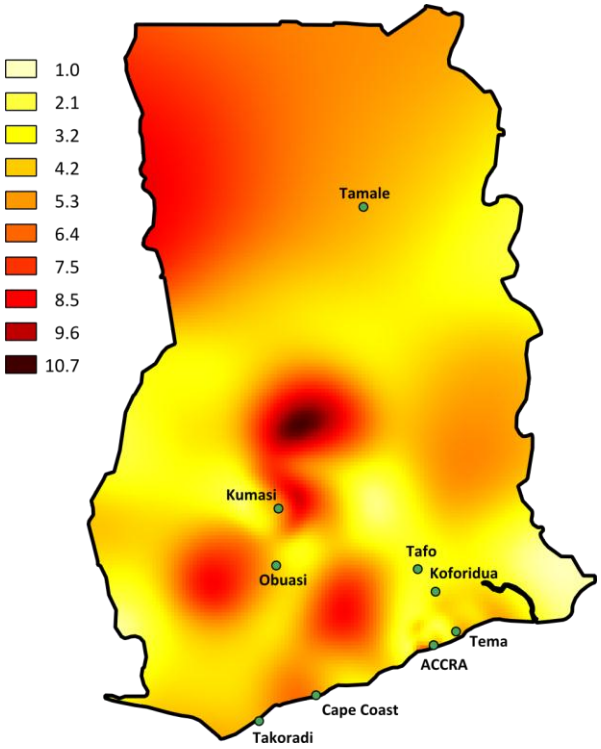
We conduct all these analyses using the free and open-source statistical software R. A specific package called prevR was written and can be downloaded free of charge on the *Comprehensive R Archive Network* (http://cran.r-project.org/web/packages/prevR/index.html). prevR is bilingual

(English and Fench) and allows to import data, to perform analysis and to export results to GIS software.

Application to real data

This approach could be applied to a wide range of indicators (mortality rate, diseases prevalence, education, household furnishings…) as long as they could be formulated as a proportion. Figure 2 presents an application to infant mortality rate (deaths bellow one year) in Ghana from the 2008 DHS. Other examples of application will be presented.

Figure 2. Infant mortality rate (in %) in Ghana 2003-2008



Source: DHS Ghana 2008. N parameter: 150.

Conclusion

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. The surfaces thus generated are relatively accurate for densely populated areas and strongly smoothed in sparsely surveyed areas.

Although local variations 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 (Chorley and Haggett, 1965; Nettleton, 1954) 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.

While produced maps should be interpreted with caution, they do provide a descriptive indication of the state of a phenomenon in a country independent of administrative divisions. It is a useful tool for displaying the main spatial variations and identifying potential hotspots. Although DHSs are insufficient for analysing spatial determinants, they do make it possible to sketch out a preliminary picture in the absence of more specific surveys with better geographical coverage. Furthermore, the method could be easily applied using prevR, the dedicated R package.

References

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

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

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

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

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.

Kelsall J.E., Diggle P.J., 1995, "Kernel Estimation of Relative Risk", *Bernoulli*, vol. 1, No. 1/2, 3-16.

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

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.

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.

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