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

Joseph Larmarange Centre Population et Développement (CEPED UMR 196 Paris Descartes Ined IRD)

Context

For many countries, in particular in sub-Saharan Africa, Demographic and Health Surveys (DHS) are the main national source of data (depending on 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.

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

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. DHS simulations were carried out to reproduce data comparable with actual surveys.

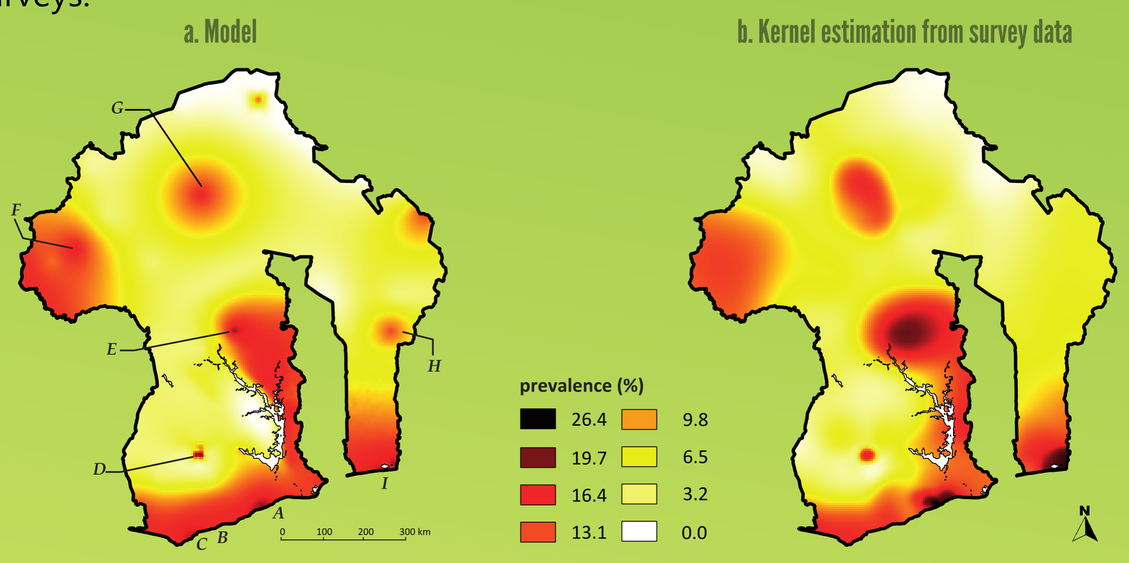


Figure a represents the prevalence surface created for the model (national prevalence of 10%). A DHS was simulated with a sample size of 8,000 people distributed in 400 clusters. Figure b represents the result of applying our method on the data obtained from this simulation.
Overall, the main variations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

Conclusion

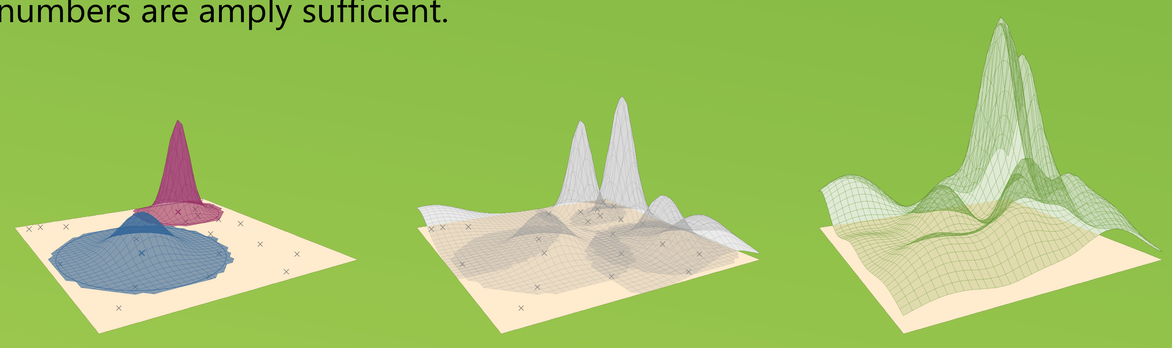
The use of adaptive bandwidths of equal number of persons surveyed makes it possible to achieve a smoothing effect that adapts to the high irregularity of spatial distribution among the survey clusters. The surfaces thus generated are relatively accurate for densely populated areas and strongly smoothed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

Methods

Kernel estimators techniques are designed to construct a surface from a scatter plot, where each point represents an observed case.
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.



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 neighbourhood of that cluster. For the estimation of the intensity surface of observed cases, the principle is similar to the nearest neighbour technique. 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.

The prevR package

This approach has been implemented in a free package for R.
Its functionalities are:

- bilingual (English and French)
- importation of files provided on measuredhs.com
- estimation of a prevalence and/or a risk ratio surface
- kernel density estimation and kriging interpolation
- results exportation in Shape files and ASCII Grids
- open Source Licence (CeCILL-C)

More Information

prevR on CRAN

http://cran.r-project.org/web/packages/prevR/

Source code on GitHub

https://github.com/larmarange/prevR

Reference

http://cybergeo.revues.org/24606

J. Larmarange, R. Vallo, S. Yaro, P. Msellati, N. Méda (2011), « Methods for mapping regional trends of HIV prevalence from Demographic and Health Surveys (DHS) », *Cybergeo : European Journal of Geography*
DOI : 10.4000/cybergeo.24606

CePeD Centre Population & Développement
IRD Institut de recherche pour le développement
anRS French National Agency for Research on AIDS and Viral Hepatitis | An autonomous agency at Inserm |