Disease Mapping of Syphilis in Forsyth County, North Carolina with Enhanced Geoprivacy and Spatial Resolution

Lani Fox[1], William Miller[2,3], Dione Gesink[4], Irene Doherty[3], Peter Leone[2,3,5], Dell Williams[5], Marc Serre[1]

[1] Department of Environmental Sciences and Engineering, Gillings School of Global Public Health, University of North Carolina at Chapel Hill
[2] Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill
[3] School of Medicine, Division of Infectious Diseases University of North Carolina at Chapel Hill
[4] Department of Public Health Sciences, Dalla Lana School of Public Health, University of Toronto
[5] Division of Public Health, Communicable Disease Branch, HIV/STD Prevention and Care Unit North Carolina Department of Health and Human Services

Abstract

Using North Carolina state health department data this paper refines the spatial resolution of disease maps utilizing geomasked syphilis cases moved by a random displacement to preserve their anonymity. A moving window approach is introduced to create ubiquitous areas to control the modifiable areal unit problem and the edge effect present in conventional methods. Syphilis incidence rates are then processed using the Uniform Model Bayesian Maximum Entropy method to correct for the small number problem. Our hypothesis is this approach will better delineate the geographical extent of clusters, improving outbreak detection and reducing ambiguous and spatially incorrect results. A cross-validation analysis demonstrated this new methodology performed considerably better than a traditional approach.

Keywords: Space-Time, MAUP, edge effect, incidence rates, geomask, spatial smoothing

1. Introduction

In 2009, North Carolina experienced an 84% (n=937) increase in infectious syphilis cases from 2008 (n=509). This increase was clearly evident in Forsyth County, which encountered more than a fourfold rise in infectious syphilis cases from 46 in 2008 to 195 in 2009 (NC, 2010). Beyond the concern for syphilis morbidity, ulcerative sexually transmitted infections such as syphilis have a significantly higher likelihood to communicate HIV when one partner is HIV positive (NC, 2010; Sena, 2008). In the outbreak in Forsyth County syphilis cases increased significantly within the HIV positive community.

The goal of this paper is the Bayesian Maximum Entropy space/time analysis (Allhouse, 2009; Gesink, 2006; Hampton, 2011; Law, 2006; Serre, 2004) of the infectious syphilis incidence among the population tested in Forsyth County from 1999-2011. This paper advances the methodology for outbreak analysis by refining spatial resolution with the use of geomasked data (Allhouse, 2010; Hampton, 2010) and applying a global moving window approach to control the MAUP and the edge effect. The small number problem (Hampton, 2011; Goovaerts, 2005) is also removed by employing Uniform Model Bayesian Maximum Entropy (Hampton, 2011). Our hypothesis is this method will better delineate the geographic extent of clusters, reducing some of the ambiguous and spatially incorrect results in past methodologies.

2. Methods

A syphilis case is defined as a Forsyth County resident infected with syphilis, and diagnosed between January 1, 1999 and April 30, 2011. Syphilis patient residences were geocoded using ESRI's ArcGIS 9.3.1 and matched to three geographic locators used by the State of North Carolina. After geocoding, the data were geomasked using the Donut Method (Allshouse, 2010; Hampton, 2010).

An incidence area is the geographic foundation on which a rate is defined. Universally, incidence rates are created by aggregating individual-level data to pre-existing administrative areas, such as counties and assigning the data to the centroid in an effort to protect patient privacy. The centroids of these incidence areas are often arbitrarily located in space with varying distances between the centroids. This creates an uneven clustering of arbitrary group (AG) centroids and introduces spatial uncertainty also known as the Modifiable Areal Unit Problem (MAUP) (Gatrell 1996; Kamel, 2006; Ratcliffe, 1999). Results can be further misled when a hot spot is located on the boundary of two or more aggregation areas a phenomenon commonly referred to as the "edge effect". MAUP and the edge effect can be corrected by enriching an AG dataset with incidence area centroids based on a regular grid. A moveable sub-region of overlapping circles is then applied on the grid (Gattrell, 1996). Cases located within each of these grid-based incidence areas, which we will refer to as ubiquitous groups (UG) are identified and utilized to calculate a rate. A rate is then assigned to each centroid (grid point) of the incidence areas.

For this study AG cases were aggregated spatially by AG boundaries (in this case census block groups), and temporally with a rolling time period of 6 months to lessen the small number problem. The crude incidence rate at location s_i and incidence period of duration T expressed in years (i.e. T = 0.5yr for a 6 month incidence), centered at time t_j is denoted as R_{ij} and calculated as $R_{ij}=y_{ij}/(n_{ij}T)$. Where y_{ij} is the number of syphilis cases within the incidence area i, n_{ij} is the population at time t_j. The UG dataset consists of the AG data combined with grid-based ubiquitous incidence areas. Each grid point was used as the centroid of a UG.

The number of cases within each UG was calculated as a function of the probability each case is within a UG at a selected time period. The incidence rate, R_i over area UG_i at time j where i is the spatial location is calculated as: $R_i \approx \frac{\sum_{l=1}^{N} w_{li}}{n_i * T}$. Where w_{li} is the probability case l is in area UG_i, N is the total number of cases in UG_i and

$w_{li} = \frac{\|AR_l \cap UG_i\|}{\|AR_l\|}$. Furthermore, $AR_l = DR_l \cap AG_l$ where DR_l is the size of the geomasked donut before the restriction the area of the donut outside of the AG must be removed, AG is the area of the AG. AR_l is the area the geomasked point was located at geocoding and prior to geomasking.

This work uses a BME geostatistical analysis to estimate the syphilis incidence in Forsyth County, NC. The BME framework allows for the incorporation of soft data modeled by a distribution such as UMBME (Hampton, 2011), a method utilized to minimize the small number problem. The UMBME analysis was conducted using an open source add-in library, BMElib (available at: http://www.unc.edu/depts/case/BMELIB/) for MATLAB. A cross-validation of the AG and UG methods was conducted to identify the most accurate method to model the syphilis rates. The cross-validation errors for each method are summarized as a function of the mean square errors (MSE).

3. Results

The cross-validation demonstrated the UG model performed noticeably better in predicting the latent rate than the AG model. This is revealed in the percent change in MSE (Hampton, 2011) which decreases as the population percentile increases as shown in Figure 2.

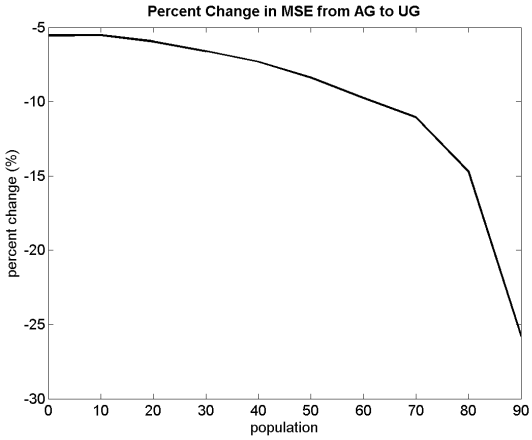

Figure 2: The percent change in MSE from AG to UG as a function of the population percentile

4. Conclusions

This study demonstrates BME mapping of sexually transmitted diseases is highly effective for quantifying and understanding the progression of health outcomes. The results of this study demonstrate the use of geomasked data and a moving window approach provide superior locational information to define core areas of infection and provide insight into outbreak patterns of transmission. This study reveals the appearance of new hotspots, increased connectivity in hotspots and places hot spots in their actual locations. Moreover, many of the new hotspots in the UG method are

located at the boundaries of the AGs demonstrating the AG maps are suffering from the edge effect.

5. References

Allshouse, W.B., Fitch, M.K., Hampton, K.H., Gesink, D.C., Doherty, I.A., Leone, P.A., Serre, M.L. & Miller, W.C. 2010, "Geomasking sensitive health data and privacy protection: an evaluation using an E911 database", *Geocarto International,* vol. 25, no. 6, pp. 443-452.

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,* , pp. 256-274.

Gesink Law, D.C., Bernstein, K.T., Serre, M.L., Schumacher, C.M., Leone, P.A., Zenilman, J.M., Miller, W.C. & Rompalo, A.M. 2006, "Modeling a syphilis outbreak through space and time using the Bayesian maximum entropy approach", *Annals of Epidemiology,* vol. 16, no. 11, pp. 797-804.

Goovaerts, P., Jacquez, G.M. & Greiling, D. 2005, "Exploring Scale-Dependent Correlations Between Cancer Mortality Rates Using Factorial Kriging and Population-Weighted Semivariograms", *Geographical Analysis,* vol. 37, no. 2, pp. 152-182.

Hampton, K.H., Fitch, M.K., Allshouse, W.B., Doherty, I.A., Gesink, D.C., Leone, P.A., Serre, M.L. & Miller, W.C. 2010, "Mapping health data: improved privacy protection with donut method geomasking", *American Journal of Epidemiology,* vol. 172, no. 9, pp. 1062-1069.

Hampton, K.H., Serre, M.L., Gesink, D.C., Pilcher, C.D. & Miller, W.C. 2011, "Adjusting for sampling variability in sparse data: geostatistical approaches to disease mapping", *International journal of health geographics,* vol. 10, pp. 54-072X-10-54.

Kamel Boulos, M.N., Cai, Q., Padget, J.A. & Rushton, G. 2006, "Using software agents to preserve individual health data confidentiality in micro-scale geographical analyses", *Journal of Biomedical Informatics,* vol. 39, no. 2, pp. 160-170.

Law, D.C., Serre, M.L., Christakos, G., Leone, P.A. & Miller, W.C. 2004, "Spatial analysis and mapping of sexually transmitted diseases to optimise intervention and prevention strategies", *Sexually transmitted infections,* vol. 80, no. 4, pp. 294-299.

Ratcliffe, J.H. & McCullagh, M.J. 1999, "Hotbeds of crime and the search for spatial accuracy", *Journal of Geographical Systems,* vol. 1, no. 4, pp. 385-398.

Seña, A.C., Torrone, E.A., Leone, P.A., Foust, E. & Hightow-Weidman, L. 2008, "Endemic early syphilis among young newly diagnosed HIV-positive men in a southeastern US state", *AIDS Patient Care and STDs,* vol. 22, no. 12, pp. 955-963.

Serre, M., Christakos, G. & Lee, S. 2004, "Soft data space/time mapping of coarse particulate matter annual arithmetic average over the US" in *geoENV IV—Geostatistics for Environmental Applications* Springer, pp. 115-126.

State of North Carolina (NC) 2010 HIV/STD Surveillance Report . 2011. http://epi.publichealth.nc.gov/cd/stds/figures/std10rpt.pdf