RESEARCH PAPER | Medical Science | Volume : 3 | Issue : 7 | July 2013 | ISSN - 2249-555X

Bayesian Kriging of Tuberculosis in Chennai : A Small Scale Analysis

KEYWORDS

Srinivasan R.
National Institute for Research in Tuberculosis, ICMR, Chennai – 600031.

Venkatesan P.
National Institute for Research in Tuberculosis, ICMR, Chennai – 600031.

ABSTRACT *Bayesian Kriging would be more appropriate in the case of tuberculosis risk where we know that other factors are strong predictors. Kriging the residuals and adding the predicted residuals to the model predictions produce predictions that are closer to the observed SMR in each ward. Kriging method used to extrapolate the location from unmeasured locations. The aim is to study the spatial pattern of tuberculosis within a Chennai ward population to gain insight into the disease spread and also, to extrapolate the location from the unmeasured locations. SAS software was used for spatial analysis of tuberculosis spread. Data was obtained from National Institute for Research in Tuberculosis for Chennai district. The location of the each case was geographically marked through their co-ordinates in the Chennai map. Kriging were used to extrapolate the location from the unmeasured locations. The results of the spread of tuberculosis in chennai ward has been diverse, with many wards having a low rate of infection and the epidemic being most extreme in slum areas. Spatial analysis is proved to be more useful for studying spread of Tuberculosis analysis and modeling of disease prediction.*

Introduction

Kriging is a technique used in the analysis of spatial data. The data from the measured location can be used to estimate the variable at the location where it had not been measured. This extrapolation from measured location to unmeasured location is called kriging. Measurements of variable at a set of points in a region is used to extrapolate points in the region where the variable was not measured outside the region that we believe will behave similarly. In both cases, we will need to first fit a variogram model to our data. The three major functions used in spatial statistics for describing the spatial correlation of observations are the correlogram, the covariance, and the semi-variogram. The last is also more simply called the variogram. The variogram is the key function in spatial statistics as it is used to fit a model of the spatial correlation of the data.

Observations made at different locations may not be independent. For example, measurements made at nearby locations may be closer in value than measurements made at locations farther apart (Tobler law). This phenomenon is called spatial autocorrelation which measures the correlation of a variable with itself through space. This spatial autocorrelation value can be positive or negative. Positive spatial autocorrelation occurs when similar values occur near one another. Negative spatial autocorrelation occurs when dissimilar values occur near one another. The two classic works reviewing and extending spatial statistical theory are given by Cliff and Ord (1981), whom have motivated research involving spatial autoregression, and Cressie (1993) has summarized research involving geostatistics. Geostatistics uses variance-covariance matrix while spatial autocorrelation uses inverse of this matrix. Griffith and Layne (1999) among others, show links between geostatistics and spatial auto regression.

Moran's I is one method used to see the autocorrelation of disease based on the location which is one of the oldest indicators of spatial autocorrelation (Moran, 1950). It measures the strength of spatial autocorrelation in a map. For calculating spatial autocorrelation, neighboring values can be identified by an n x n binary geographic weights matrix, say, C; if two locations are neighbors, then $c_{ij} = 1$, and if not, then $c_{ij} = 0$, in which two areal units are deemed neighbors if they share a common non-zero length boundary. Test of significance can be used for testing independence. There are many ways to approach the analysis of the spatial pattern of tuberculosis and HIV. If there is any systematic pattern in the spatial distribution, it is said to be spatially autocorrelated. One approach is to define disease in a ward to be close to one another, and then determine, whether pattern may have similar characteristics. Once the spatial correlation structure of a variable has been identified, the data from the measured locations can be used to estimate the spatial dependence based on location is significant or not (Banerjee et al., 2004).

Moran's I

Moran's I (Moran, 1950) used to tests spatial autocorrelation for continuous data. It uses cross-products of the deviations from the mean and is calculated for n observations on a variable x at locations i, j as:

$$I = \frac{n}{S_0} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}, \quad (1)$$

where $\bar{x}$ is the mean of the x variable, w_{ij} are the elements of the weight matrix, and S_0 is the sum of the elements of the weight matrix:

$$S_0 = \sum_i \sum_j w_{ij} \quad (2)$$

Moran's I is similar to a correlation coefficient. It varies from -1 to +1. For a row-standardized spatial weight matrix, the normalizing factor S_0 equals n and the statistic simplifies to a ratio of a spatial cross product to a variance.

The variance is:

$$\mathrm{Var(I)} = \frac{n\{n^2 - 3n + 3)S_1 - nS_2 + 3S_0^2\} - k\{n(n-1)S_1 - 2nS_2 + 6S_0^2\}}{(n-1)(n-2(n-3)S_0^2} - \frac{1}{(n-1)^2} \quad (3)$$

where

$$S_1 = \frac{1}{2}\sum_i \sum_j (W_{ij} + W_{ji})^2 = 2S_0 \text{ for symmetric W containing 0's and 1's}$$
$$S_2 = \sum_i (W_{i0} + W_{0i})^2 \text{ where } W_{i0} = \sum_j W_{ij} \text{ and } W_{0i} = \sum_j W_{ji} \quad (4)$$

Testing of Significance: Empirical distributions can be compared to the theoretical distribution by dividing by an estimate of the theoretical standard deviation

$$Z(I) = \frac{I - E(I)}{S_{E(I)}} \quad (5)$$

Volume : 3 | Issue : 7 | July 2013 | ISSN - 2249-555X

$$S_{E(I)} = SQRT\left[\frac{N^2\sum_{ij} w_{ij}^{\ 2} + 3(\sum_{ij} w_{ij})^2 - N\sum_i (\sum_j w_{ij})^2}{(N^2-1)(\sum_{ij} w_{ij})^2}\right] \quad (6)$$

Geary's C:
Geary's C statistic (Geary 1954) is based on the deviations in responses of each observation with one another:

$$C = \frac{n-1}{2S_0}\frac{\sum_i\sum_j w_{ij}(x_i - x_j)^2}{\sum_i (x_i - \bar{x})^2} \quad (7)$$

Geary's C ranges from 0 to a positive value for high negative autocorrelation. If the value of Geary's C is less than 1, it indicates positive spatial autocorrelation.

$$\mathrm{Var}(c) = \frac{1}{n(n-2)(n-3)S_0^2}\{S_0^2[(n^2-3) - k(n-1)^2] +$$

$$S_1(n-1)[n^2 - 3n + 3 - k(n-1)] + \frac{1}{4}S_2(n-1)[k(n^2 - n + 2) - n^2 + 3n - 6]\} \quad (8)$$

where S_0, S_1, and S_2 are the same as in Moran's I. and C is never negative and has mean 1 for the null models; low values indicates positive spatial association. Also C is a ratio of quadratic forms in Y like I is asymptotically normal if the Y_i are i.i.d.

Test of significance is given by

$$Z(C) = \frac{C - E(C)}{S_{E(C)}} \quad (9)$$

Moran's I is a global measurement and sensitive to extreme values of x, whereas Geary's C is sensitive to differences in small neighborhoods. Moran's I is preferred in most cases, Cliff and Ord (1981) have shown that Moran's I is consistently more powerful than Geary's C.

GEO-STATISTICAL PREDICTION
Kriging is a technique used in the analysis of spatial data. The data from the measured location can be used to estimate the variable at the location where it had not been measured. This extrapolation from measured location to unmeasured location is called kriging. This method of Prediction introduced by Kriging (Krige, 1966; Oliver et al., 1992; Carrot and Valleron, 1992; Diggle et al., 1998) is based on the assumption that covariance between points is entirely a function of distance between them as modeled by means of the variogram. Further it is assumed that the underlying mean of the quantity being predicted is constant. Kriging is based on the idea that you can make inferences regarding a random function Y(S), given data points $Y(S_1)$, $Y(S_2)$,…,$Y(S_n)$.

$$Y(S) = m(S) + \gamma(h) + \varepsilon \quad (10)$$

where there are three components namely constant mean, random spatially correlated component and residual error used to predict the unmeasured value.

Ordinary Kriging
The model and notation is followed by Cressie (1993) for $Y(S)$ is as below;

$$Y(S) = \mu + \varepsilon(S) \quad (11)$$

here, μ is the fixed, unknown mean of the process, and $\varepsilon(S)$ is a zero mean, which represents the variation around the mean. In most practical applications, an additional assumption is required in order to estimate the covariance C_Z of the $Y(S)$ process. This assumption is second-order stationarity

$$C_Z(s_1, s_2) = E[\varepsilon(s_1)\varepsilon(s_2)] = C_z(s_1 - s_2) = C_z(h) \quad (12)$$

This requirement can be relaxed slightly when you are using the semi-variogram instead of the covariance. In this case, second-order stationarity is required of the differences $\varepsilon(s_1) - \varepsilon(s_2)$ rather than $\varepsilon(s)$

$$\gamma_z(s_1, s_2) = \frac{1}{2}E[\ \varepsilon(s_1) - \varepsilon(s_2)\ ^2] = \gamma_z(s_1 - s_2) = \gamma_z(h) \quad (13)$$

By performing local kriging, the spatial processes represented by the previous equation for $Y(s)$ are more general than they appear.

Bayesian kriging
A Bayesian approach brings additional flexibility to the classical prediction framework outlined above. First of all, the issue of incorporating uncertainty in covariance parameters follows directly from posterior inference. More specifically, Bayesian prediction derives from the posterior predictive distribution which integrates over the posterior distribution of all model parameters,

i.e.,

$$p(Z(S_0) \mid Z, X, x(S_0)) = \int p(Z(S_0)\ \beta, \alpha, \eta, \theta \mid Z, X, x(S_0))\ d\beta dx d\eta d\theta,$$

$$= \int p(Z(S_0) \mid Z, X, x(S_0)\ \beta, \alpha, \eta, \theta) p(\beta, \alpha, \eta, \theta \mid Z, X) d\beta d\alpha d\eta d\theta \quad (14)$$

In many cases, this integration may be numerically or theoretically intractable leading to the use of MCMC methods for evaluation. In addition to providing a mechanism for accounting for uncertainty in variance-covariance modeling, Bayesian thinking enters the spatial prediction process in additional ways as well. One simple advantage is the conceptual ease of accommodating functions of model parameters. In the Bayesian context, transformations are less problematic, as predictions derive from the posterior predictive distribution rather than a necessarily linear combination of observations. This is particularly evident in MCMC implementation where a sample from the posterior distribution of model parameters may be transformed to a sample from the posterior distribution of the transformed parameters with little effort.

Material and methods:
The datasets is taken from National Institute for Research in Tuberculosis (NIRT), formerly TRC, which is conducting clinical trials for Tuberculosis and HIV in Chennai and its suburbs since 1956. The patients registered during 2004 to 2006 for an ongoing trial were considered for this study. Chennai had 155 wards and for each ward the total number of TB cases recorded between 2004 and 2006 were identified. For variogram analysis, 28 wards of Chennai district were selected for our study for which the locations of 72 cases were geographically marked through their co-ordinates in the Chennai map.

Kriging analysis was carried out using SAS software by dividing the whole area into some 100 by 100 grid matrix and prediction is calculated using ordinary kriging method. For Bayesian kriging, WinBUGS software was used for prediction of certain locations based on available information about other location. The spatial prediction permits spatial interpolation and prediction in WinBUGS. The data for this work consist of values of SMR, and coordinates of x and y of the each wards. The spatial.exp function allows the fitting of a fully parameterized covariance function within a multivariate normal distributional model. Spatial.unipred provides a method of predicting values of the fitted surface at unsampled locations.

Results:
Table 2 shows the observed and predicted values of the SMR in Chennai wards.

Table 2 Kriging (prediction) Estimates

Wards	SMR	Ordinary Kriging		Bayesian Kriging	
		Prediction	Std Error	Prediction	Std Error
Ward1	0.469	0.359	0.11	0.431	0.038
Ward 2	1.611	1.181	0.43	1.591	0.02
Ward 3	0.628	0.428	0.2	0.614	0.014

Volume : 3 | Issue : 7 | July 2013 | ISSN - 2249-555X

Ward 4	0.646	0.87	0.224	0.635	0.011
Ward 5	0.352	0.232	0.12	0.362	0.01
Ward 6	0.609	0.94	0.331	0.611	0.002
Ward 7	2.492	0.942	1.55	2.123	0.369
Ward 8	0.858	0.488	0.37	0.812	0.046
Ward 9	0.701	0.211	0.49	0.712	0.011
Ward 10	1.01	1.31	0.3	1.11	0.1

The Bayesian Kriging prediction gave results close to the observed SMR. Also the Bayesian approaches resulted in lower SE

Spatial Autocorrelation

In our study, nearby areas are more alike, and it indicates positive spatial autocorrelation and there is no negative autocorrelation or Random patterns exhibit in this area. The results are presented in Fig. 2.6.

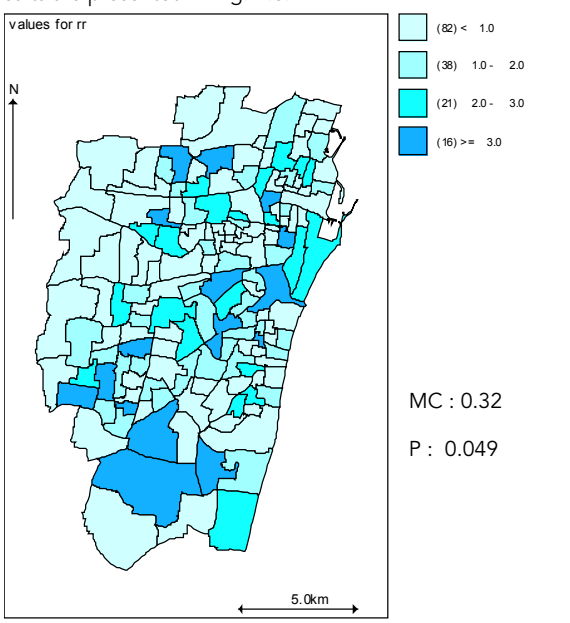

Figure 2.6 Measure of Spatial Autocorrelation for Chennai wards

The Moran's spatial autocorrelation of Chennai city is MC: 0.32 and statistically significant also. The result reveals that the high risk area surrounded by high risk and moderately high risk area and low risk area surrounded by low risk and risk free area and some wards show no autocorrelation between disease and spatial pattern. It is reflected in map also.

Conclusion

The GIS system proves to be a friendly interface for spatial and a spatial information retrieval, which supports users with all type of statistical analysis GIS model. Spatial dependence exists between small distances of tuberculosis cases found in Chennai wards. Kriging has significantly improved the prediction of tuberculosis risk in parts of the Chennai city, however, given that the data used for obtaining the model are not a random sample of the population or a spatially well distributed set of sampling points, and extrapolating the predicted risk to points outside the data set closely matched with our observed SMR. A concern with spatial data is the potential for spatial correlation in the observations, which could lead to incorrect estimates. An infectious disease that is heavily associated with other variables is likely to be spatially clustered. The model derived here explains some of the spatial dependence of tuberculosis risk, with significant spatial correlation, particularly over short distances.

Bayesian Kriging would be more appropriate in the case of tuberculosis risk where we know that other factors are strong predictors. The Moran's spatial autocorrelation of Chennai city is 0.32 and statistically significant also. The result reveals that the high risk area surrounded by high and moderately high risk areas and low risk areas surrounded by low risk areas and some wards shows no autocorrelation between disease patterns.

REFERENCE

Banerjee S, Carlin B and Gelfand A E (2004): Hierarchical Modeling and Analysis for Spatial Data. Boca Raton: Chapman & Hall. | Berger J O (1985): Statistical Decision Theory and Bayesian Analysis, 2nd ed., New York: Springer-Verlag. | Best N, Ickstadt K and Wolpert R (2000): Spatial Poisson regression for health and exposure data measured at disparate resolutions. Journal of American Statistical Association, 95: 1076-1088. | Carrat F and Valleron A J (1992): Epidemiologic mapping using the kriging method: application to an influenza-like illness epidemic in France, American Journal of Epidemiology, 135: 1293-1300. | Cliff A and Ord J (1981): Spatial Processes. Pion, London. | Cressie N (1985): Fitting variogram models by weighted least squares, Mathematical Geology, 5: 563-586. | Cressie N (1986): Kriging Non-stationary Data. Journal of the American Statistical Association, 81: 625–634. | Cressie N (1993): Statistics for spatial data. New York: John Wiley & Sons. | Diggle P (1993): Statistical Analysis of Spatial Point Patterns. New York: Academic Press. | Diggle P, Morris S and Wakefield J (2000): Point-source modelling using matched case-control data. Biostatistics, 1: 1-17. | Diggle P, Tawn J A and Moyeed R A (1998): Model-based geostatistics. Journal of Royal Statistical Society, Series-C (Applied Statistics), 47: 299 | Gelfand A Diggle P and Guttorp P (2010): Handbook of spatial statistics, Chapman & Hall / CRC. | Isaaks E and Srivastava R (1989): An Introduction to Applied Geostatistics. Oxford: Oxford University Press. | Lawson A (1993): On the analysis of mortality events around a prespecified fixed point. Journal of the Royal Statistical Society, Series-A, 156: 363-377. | Matheron G (1963): Principles of geostatistics, Economic Geology, 58: 1246-1266. | Pickle L, Mungiole M, Jones G and White A (1999): Exploring spatial patterns of mortality: the new atlas of United States mortality. Statistics in Medicine 18: 3211- 3220. | Ripley B (1981): Spatial Statistics. New York: Wiley. | Shapiro A and Botha J (1991): Variogram fitting with a general class of conditionally nonnegative definite functions. Computational Statistics and Data Analysis, 11: 87-96. | Stein A and Corsten L (1991): Universal kriging and cokriging as a regression procedure. Biometrics, 47: 575-587. | Stein A, Van Eijnbergen, A C and Barendregt, L G (1991): Cokriging nonstationary data. Mathematical Geology, 23: 703-719. | Stein M (1999a): Interpolation of Spatial Data: Some Theory for Kriging. New York: Springer-Verlag. | Venkatesan P and Srinivasan R (2010): Modeling the variogram of tuberculosis in Chennai ward, Indian Journal of Science and Technology. 3 : 167-69. | Wakefield J and Morris S (2001): The Bayesian modeling of disease risk in relation to a point source. Journal of American Statistical Association, 96: 77-91. | Wakefield J and Salway R (2001): A statistical framework for ecological and aggregate studies. Journal of Royal Statistical Society, Series-A, 164: 119-137.