Indian Journal of Science and Technology Vol. 3 No. 2 (Feb 2010) ISSN: 0974- 6846

Modeling the spatial variogram of tuberculosis for Chennai ward in India.

P. Venkatesan and R. Srinivasan
Tuberculosis Research Centre, ICMR, Chennai - 600 031, India
venkaticmr@gmail.com, srinivast_r@yahoo.com

Abstract

In this paper, we have used statistical measures and spatial deviational ellipse to determine the spatial pattern of tuberculosis within a Chennai ward population to gain insight into the disease spread. Variogram is used to describe the spatial dependence of tuberculosis in Chennai wards and it is compared with theoretical variogram model of spherical, Gaussian and exponential fitted to tuberculosis data. Arc View GIS 9.2 and SAS software were used for spatial analysis of tuberculosis spread. Data were obtained from District Hospital records for Chennai wards. The results of the spatial pattern revealed that the spread of tuberculosis in Chennai wards have been diverse, with many wards having a low rate of infection and the epidemic being most extreme in slum areas. Variogram increases with distance at small distances and then level off which implies spatial dependence exists between small distance of tuberculosis cases. Spherical model fits data better. Spatial analysis is proved to be more useful for studying spread of tuberculosis analysis and modeling of disease analysis.

Keywords: Bayesian, disease mapping, variogram, spatial correlation and deviational ellipse.

Introduction

Geographical data are correlated in space. Data in close geographical proximity is more likely to be influenced by similar factors and thus affected in a similar way. In the case of tuberculosis, spatial correlations are present at both short and large scales, reflecting the transmission of tuberculosis infection and the effects of environmental factors (Venkatesan and Srinivasan, 2008).

The three major functions used in spatial statistics for describing the spatial correlation of observations are the correlogram, the covariance and the semi-variogram or 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. Measurements of variable at a set of points in a region are used to extrapolate points in the region where the variable was not measured outside the region that we believe will behave similarly. We can use predictions on our measured values by kriging or we can incorporate some factors and make predictions using a regression model. In both cases, we need to first fit a variogram model.

In the variogram analysis, more robust estimators were suggested by Cressie and Hawkins (1980) and Genton (1998). Two kernel-type estimators are introduced by Garcia (2003, 2004), which result from adapting the local linear estimators to the context of spatial data. However, the parametric model of variogram pioneered by Menezes *et al*. (2005) for different spatial dependence situations. A goodness of fit test has been suggested in Maglione and Diblasi (2004), where the random process is assumed to be Gaussian and isotropic.

The variogram analysis describes the spatial continuity or roughness of a data. It consists of the experimental variogram calculation and the variogram model fitted to the data. The experimental variogram is calculated by averaging one half of the differences squared of the z-values over all pairs of observations with the specified separation distance and direction. It is plotted as a two-dimensional graph. It is used to evaluate whether disease characteristics of the cases are clustered or random. Clustered disease cases would be reflected in positive values in the variogram at the distances corresponding to the spatial scale of clustering. If a case is clustered, neighbors are more likely to share the same disease status than are tuberculosis separated by larger distances.

The mathematical form of the variogram is as follows; the variogram value $\hat{\gamma}(h)$, or estimated semi variance, for lag distance h is defined as:

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [z(x_i) - z(x_i + h)]^2 \quad \text{.....(1)}$$

where, N(h) is the number of pairs of points separated by h, z (x_i) is the data value for the point x_i, and z (x_i+h) is the data value at cells separated from x_i by the lag distance h in the chosen direction. The variogram is characterized in Fig.1.

In this study, we have used three models namely spherical, exponential and Gaussian models as stated below: *Spherical model*. Most commonly used in variogram model is the Spherical model and its equation is,

$$\gamma(h) = \begin{cases} 1.5\left(\frac{h}{a}\right) - \left(\frac{h}{a}\right)^3 & \text{if } h \le a \\ 1, & \text{otherwise} \end{cases} \quad \text{.....(2)}$$

Indian Journal of Science and Technology Vol. 3 No. 2 (Feb 2010) ISSN: 0974- 6846

where 'a' is the range. It has a linear behavior at small distances near the origin but flattens out at larger distances, and reaches the sill at a.

The Exponential model: The exponential model is given by,

$$\gamma(h) = 1 - \exp\left(-3\frac{h}{a}\right) \quad \text{.....(3)}$$

This model reaches its sill asymptotically, with the practical range 'a' defined as that distance at which the variogram value is 95% of the sill. It rises more steeply and then flattens out more gradually.

The Gaussian model: This is another transition model that is often used to model extremely continuous phenomena. Its equation is given by,

$$\gamma(h) = 1 - \exp\left(-3\frac{h^2}{a^2}\right) \quad \text{.....(4)}$$

It is like an exponential model, but differs only in its parabolic behavior near the origin.

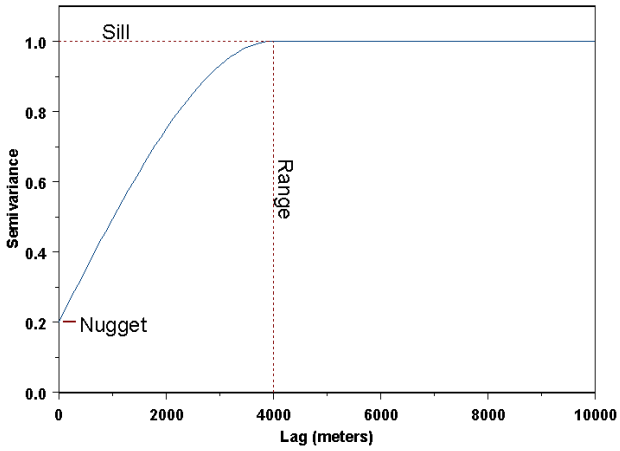

Fig.1. The characterize of the variogram

Materials and methods

This study focuses on the spatial pattern of tuberculosis within Chennai wards population. To gain insight into the disease spread, we have used spatial measure of statistics. This work used Arc View GIS 9.2 with inbuilt RDBMS for tuberculosis descriptive analysis and SAS software for variogram analysis. Data for this study was obtained from District Hospital records of Chennai district. Twenty eight wards of Chennai district were selected for this study. The locations of the cases were geographically marked through their co-ordinates in the Chennai map.

Variogram model is used to check the spatial dependence of tuberculosis in Chennai and it is also compared with theoretical variogram models of spherical, Gaussian and exponential fitted to tuberculosis cases. The location of 72 cases was geographically marked through their co-ordinates in the Chennai map. To describe the spatial pattern of tuberculosis cases, spatial mean, spatial standard deviation and spatial standard deviational ellipse were calculated. Spatial mean measured through the mean of the X and the Y coordinates for a set of points which is also called centroid. Standard distance deviation was measured, to see a dispersion of the incidents around the mean center, but it does not capture any directional bias. Standard deviational ellipse was used to study directional of the distribution. Variogram models were used to mathematically describe the shape of the disease. We plotted the variogram, and then assessed the shape of the variogram to determine which model is more appropriate for our dataset using proc variogram and proc mixed command in SAS software.

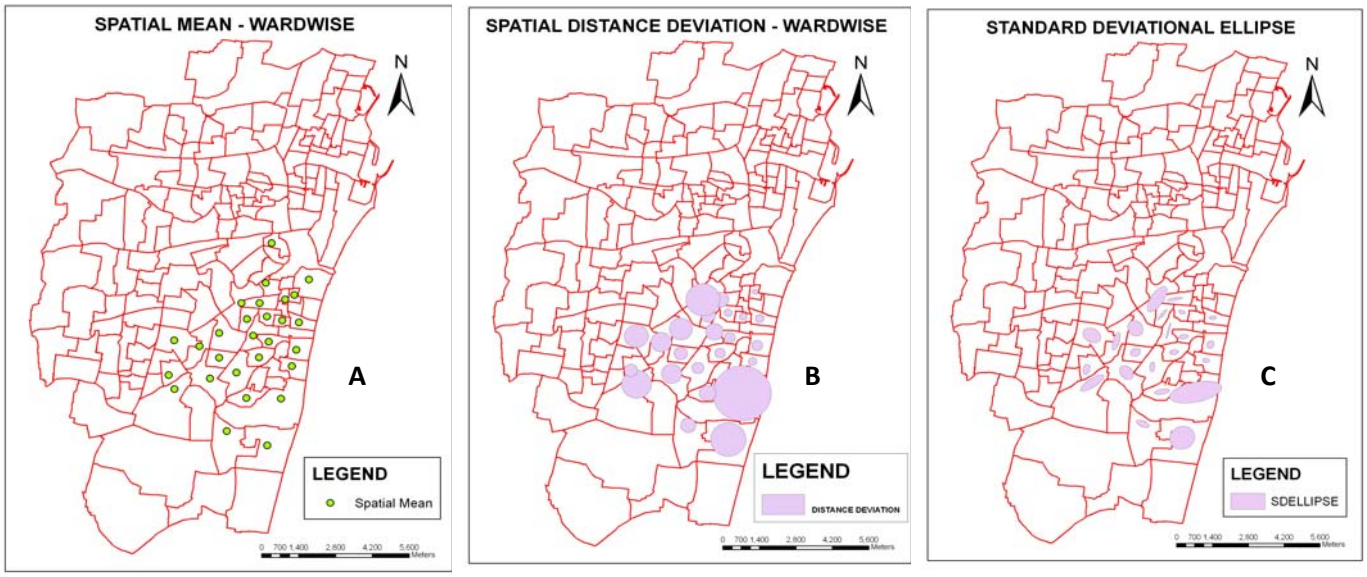

Fig. 2. A. Spatial Mean, B.Spatial Standard Deviation and C. Standard Deviational Ellipse

Indian Journal of Science and Technology Vol. 3 No. 2 (Feb 2010) ISSN: 0974- 6846

The Statistical assumptions involved in this analysis are i) stationary i.e., mean and variance are not a function of location, ii) second-order stationary which implies variance is a function of the separation distance, iii) isotropy- no directional trends occur in the data, and iv) lag distances.

Results

Fig.2A. shows the spatial mean of the X and the Y coordinates for a set of points for selected ward in Chennai district which is the centroid of the particular selected ward.

The standard distance deviation map (Fig.2B) gives dispersion of the cases around the mean center, but this does not capture any directional bias and standard deviational ellipse (Fig.2C) is used to study directional of the distribution of disease spread.

The diagonal covariance of the variogram follows the very general increasing then flattening shape of our data. Also, the variogram increases with distance at small distances and then levels off after certain point. This general shape is suggestive of a spatial correlation that is positive and strong at small distances and becomes less so as distances increase until reaching a certain distance.

We consider models with three different covariance structures and compare the likelihoods of these models. Fig. 3. includes the graph of three theoretical variograms and the variogram calculated from our data. Though all three theoretical variograms follow the very general increasing then flattening shape, the Gaussian variogram appears to be closely matching the data. In all three of these models, the variogram increases with distance at small distances and then levels off. This graph also reveals how these three theoretical variograms differ in shape: exponential increases gradually and is concave over the range; spherical features a sharp increase and a quick leveling off and Gaussian offers a compromise between the two.

Our results (Table 1) revealed that the spherical value has less Deviance, AIC, AICC and BIC, compared to Gaussian and exponential and diagonal covariance structure. Gaussian model is very close to spherical model, but spherical covariance structure best fits the data.

Conclusion

This study has shown that the GIS system proves to be a friendly interface for spatial information retrieval, which supports users to map particular disease of tuberculosis. The spatial dependence exists between small distances of tuberculosis cases found in Chennai wards. Spherical model is a better fit followed by Gaussian and exponential model. Spatial analysis is proved to be more useful for modeling of disease analysis.

Table 1. Fit statistics

	Diagonal	Spherical	Gaussian	Exponential
Deviance	298.1	199.9	204.2	255.2
AIC	300.1	203.9	208.2	259.2
AICC	300.1	204.1	208.4	259.4
BIC	302.2	208.2	212.5	263.4

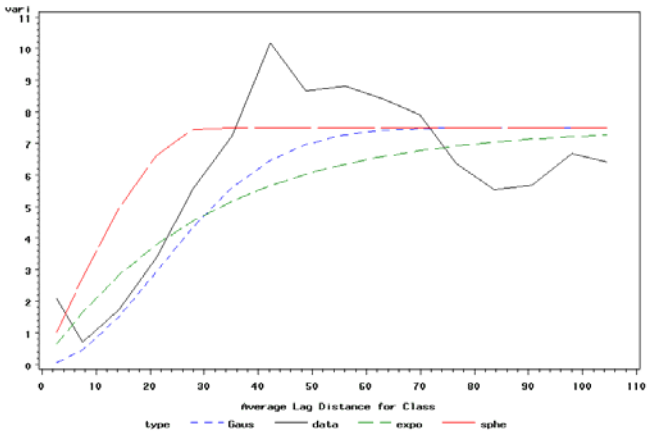

Fig. 3. Comparison of theoretical variograms with study data

References

1. Cressie N and Hawkins DM (1980) Robust estimation of the variogram. *J. Internet. Assoc. Math. Geol.* 12, 115-125.
2. Garcıa Soidan P (2003) Local linear regression estimation of the variogram. *Statist. Probab. Lett.* 64, 169-179.
3. Garcıa Soidan P (2004) Nonparametric kernel estimation of an isotropic semi variogram. *J. Statist. Plann. Inference.* 121, 65-92.
4. Genton M (1998) Highly robust variogram estimation. *Math. Geol.* 30, 213-221.
5. Isaaks EM and Srivastava RM (1989), In Introduction to Applied Geostatistics, Oxford University, NY.
6. Maglione DS and Diblasi AM (2004) Exploring a valid model for the variogram of an intrinsic spatial process. *Stoch. Envir. Res. Risk Ass.* 18, 366-376.
7. Menezes R, Garcia Soidán P and Febrero Bande M (2005), A comparison of approaches for valid variogram achievement. *J. Comput. Stat.* 20, 623 - 640.
8. Venkatesan P and Srinivasan R(2008), Applied Bayesian statistical Analysis. Proceeding of NSABSA 2008, 51-56.