

Spatial Analysis of Cancer Incidences to Identify Risk Areas and Hot Spots: a Case Study in the Western Regions of Tamil Nadu, India



Biomedical

KEYWORDS : Cancer, spatial point pattern analysis, GIS

Dr. P. B. Harathi	Assistant Professor, Department of Zoology, PSGR Krishnammal College for Women, Coimbatore.
Janani Selvaraj	Research Scholar, Department of Environmental Management, Bharathidasan University, Trichy.
Dr. M. Prashanthi Devi	Assistant Professor, Department of Environmental Management, Bharathidasan University, Trichy.
Dr. S. Valarmathi	Assistant Professor, Department of Biochemistry, Sri Ramakrishna College of Arts and Science, Coimbatore
Dr. S. Balasubramanian	Director (Research), JSS University, Mysore

ABSTRACT

Background: The burden of cancer is growing globally and is one of the top leading causes of death. Information on cancer patterns is essential for effective planning of cancer control interventions. In specific the geographical study of cancer will help in identifying the high risk communities for further etiological studies. **Objective:** The present study aims to investigate the application of various spatial statistical tools to identify the high cancer risk zones in the western regions of Tamil Nadu, India. **Methodology:** Spatial point pattern analysis was performed to assess the area based risk factor for cancer in the study area. The cancer incidences recorded in each address were geo-coded to build point features. Dual kernel estimation method was used to simplify the complex point patterns without diminishing the significance of the incidence level data. The incident hot spots were verified and tested for their statistical significance against a random distribution by means of Nearest Neighborhood Index, Ripley's K, Geary's C and Moran's I test. CrimeStat software (CrimeStat III, 2004) and ArcGIS 9.1 were used to obtain these results. **Results and Conclusion:** The smoothed map produced through the Kernel estimation method showed high clustering in the Coimbatore North, Coimbatore South and Erode taluks and was confirmed statistically by the Nearest Neighbourhood Index and Ripley's K test. Further, from the values obtained by the Moran's I and Geary's C test it is observed that there exists positive partial autocorrelation in the point data. Hence the spatial analytical methods will be useful tools in conducting further etiological studies in the high risk regions. In addition, it will be also helpful for the health professionals to organize early cancer screening programs and better prevention strategies for the society.

Introduction

Cancer is one of the main causes of morbidity and mortality in the world which makes it an important health problem. Although most cancer incidence rates are found in developed countries, more than 60% of cancer deaths occur in low or medium income countries due to limited resources for prevention, diagnosis and treatment. The International Agency for Research on cancer (IARC,WHO) in its latest report has estimated 14.1 million new cancer cases and 8.2 million cancer-related deaths occurred in 2012, compared to 12.7 million and 7.6 million, respectively, in 2008(GLOBOCAN, 2012). In India, the IARC estimated indirectly that about 635000 people died from cancer in 2008 representing about 8% of all estimated global cancer deaths and about 6% of all deaths in India (Ferlay *et al*, 2010). The absolute number of cancer deaths in India is projected to increase because of population growth and increasing life expectancy (Rajesh *et al*, 2012). Rates of cancer deaths are expected to rise, particularly from increases in the age-specific cancer risks of tobacco smoking which increase the incidence of several types of cancer (Jha, 2009). India is a culturally diverse country, with huge regional and rural-to-urban variation in lifestyles and in age-specific adults' death-rates (SRS, 2009). Thus understanding the geographical and social distribution of specific cancers is essential to target cancer control programs and spur further research into the causes of cancer.

Geographical Study for Cancer

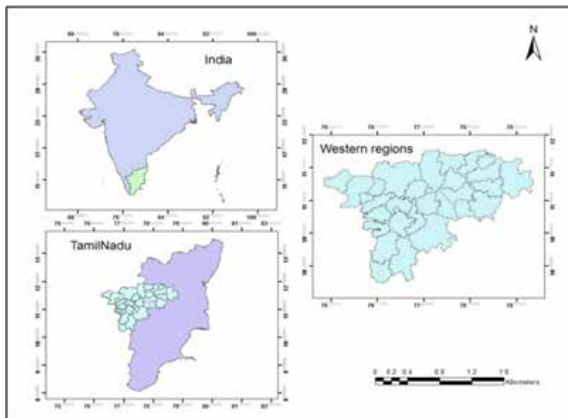
Fighting against cancer requires knowledge about the occurrence of the disease and its variations in the different regions. Lee and Irving (1999) state that the notion of epidemiology is to understand the spatial patterns of diseases in a population that will provide insights to their causes and controls. Disease mapping is an important tool for medical geographers that will help to identify associations between

disease and related factors such as environmental pollution. Inevitably, disease maps stimulate the formation of causal hypothesis by enabling the simultaneous examination of multiple factors associated with disease linked by location. Rushton, (1999); Elliot *et al*. (2000) and Pickle (2002) have applied medical geography as a tool for disease mapping and geographical correlation studies to health-related issues. Geographic studies of cancer incidence can provide important guidance for disease control and prevention practices by highlighting high risk communities in need of enhanced interventions. The potential application of Geographical Information Systems (GIS) will help the cancer researchers to gain deep insights into the etiologies of the disease. It has long been recognized that cancer rates vary by region (Dorn and Cutler, 1955; Mason *et al*, 1975), but only recently it has become apparent that local neighborhoods can also have an influence on cancer outcome (Roux, 2001), perhaps through shared environmental exposures, cultural and behavioral factors. GIS and spatial analysis provide such a solution and thus can play a major role in cancer control.

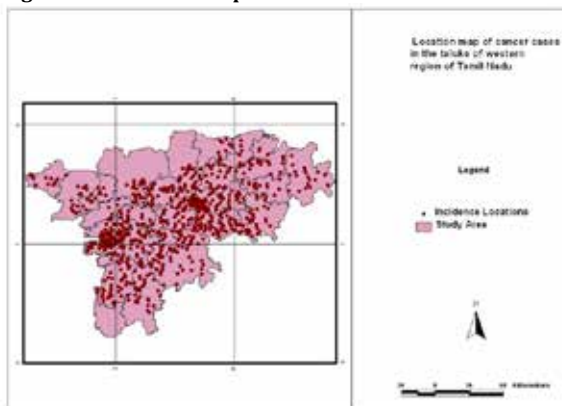
The present study aims to investigate the application of spatial statistical tools to identify the high cancer risk zones in the western regions of Tamil Nadu, India.

Study Area Description

The data for the present cross-sectional study was collected from the records of NCRP, as well as recognized cancer hospitals from the western region of Tamil Nadu for six years from 2001 to 2006. The western region of Tamil Nadu includes the districts Old Coimbatore, Erode, Namakkal, Salem and Nilgiris (Figure 1).

Figure1: Study area location

There are twenty-nine taluks (sub-unit of a district) from the five districts in the western region of Tamil Nadu. The necessary attributes from the oncology case sheets were entered into a database. The topo-sheets 58A, 58B, 58E, 58F and 58I from SOI (Survey of India), Government of India, covering the western region of Tamil Nadu were used in the preparation of base map with the scale of 1:250,000. The individual cancer cases were geo-coded using Google Maps, Google Earth and a GPS (Global Positioning System).

Figure 2: Geo-coded map of cancer incidences

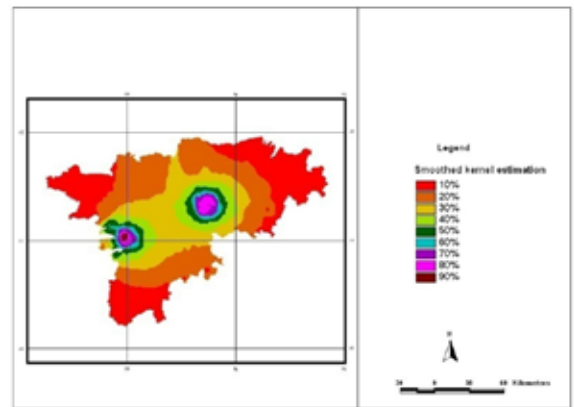
Statistical methods

Spatial point pattern analysis was performed to assess the area based risk factor for cancer in the study area. The cancer incidences recorded in each address were geo-coded to build point features. Spatial point pattern involves the use of dual kernel estimation method to simplify the complex point patterns without diminishing the significance of the incidence level data. This approach interpolates and smoothens point locations to an entire area. A moving three dimensional function is placed on a given radius or bandwidth that visits each event in turn and the area around each point is weighted according to the distance from the centre of the space. The smoothed surface is created based on the sum of these individual kernels. A ratio of the density estimate of the disease to the density estimate of the total population was obtained and mapped through interpolation method. The incident hot spots can then be verified and tested for their statistical significance against a random distribution by means of Nearest Neighborhood Index, Geary's *C* and Moran's *I* test. CrimeStat software (CrimeStat III, 2004) and ArcGIS 9.1 were used to obtain these results. Various researchers have utilized the spatial autocorrelation statistics to find the cancer hot spots in the defined study region. (Shamsul Azhar Shah *et al*, 2014; Khalid Al-Ahmadi and Ali Al-Zahrani, 2013)

Methodology and Results

Smoothing of the entire area is performed using kernel estimation as an interpolator. A ratio of the density estimate of cancer incidences to the density estimate of total population was

obtained and mapped through interpolation method. The dual kernel technique involves using the density of the individual cases over the population at risk and the ratio is applied to obtain a spatial smoothed map of risk. The estimation was made using a grid cell size of 250 meters and a fixed bandwidth of 1500 meters (Ali *et al*, 2003). The smoothed surface is created based on the sum of these individual kernels.

Figure 3: Smoothed map of kernel estimates using point pattern analysis

From the resultant Figure 3 it is clearly observed that risk areas are clustered in Coimbatore North, Coimbatore South and Erode taluks. The intensity of distribution of cancer appears to vary in the study area and it exhibits a gradual pattern. This may be related to the distribution of the spatial and related environmental factors among the areas. Kernel density is applied to identify the hotspots and it is not an interpolation technique, but more precisely the estimation of a probability surface. To test the statistical significance Nearest Neighbor Index and Ripley's *K* is applied to the data.

Nearest Neighbor Index

Nearest Neighbor Index (NNI) is applied to statistically analyze the intensity of the disease spread. The NNI is obtained by taking the ratio of the observed mean nearest neighbor distance to the mean random distance. NNI values indicate whether the disease is clustered or dispersed. The values of NNI range between 0 and 2.1491. If all the points in a pattern fall at the same location the nearest neighbor distance is 0. The more closely the points are clustered together then the value of NNI will be closer to 0 because the average nearest neighbor distance decreases. The points are said to be randomly spaced when NNI gets closer to 1 and points are more uniformly spaced when NNI value is closer to 2.1491. Table 1 indicate that the cancer data exhibits a clustered pattern, with Z test statistics of -19.7303 significant at $p=0.001$ for both one-tailed and two-tailed tests. NNI value is 0.6295 which lies between 0 and 1, more or less as a centroid value indicating partial clusters and dispersion. The observed clusters had high peaked values ranging from 0.5 to 22.

Ripley's *K*

Ripley's *K* function describes the spatial dependence between events of the same type. The *K* function is the most commonly used method and identifies the distance at which clustering occurs

$$K(s) = \frac{1}{\lambda^2 R} \sum_{i \neq j} \sum_{s \leq d_{ij}} I_s(d_{ij}) \quad \text{.....(1)}$$

R = equals the area of a region of interest

d_{ij} = is the distance between i th and j th events in R

I_s = is an indicator function which equals 1 if $d_{ij} \leq s$ or 0 otherwise.

When spatial correlation is present, each event is likely to be in close proximity to other members of the same type and for small values of s , $K(s)$ will be large. The dotted lines represent the calculated maximum and minimum values of Ripley's *K* for that lag distance in the plot and the dark line represents the ac-

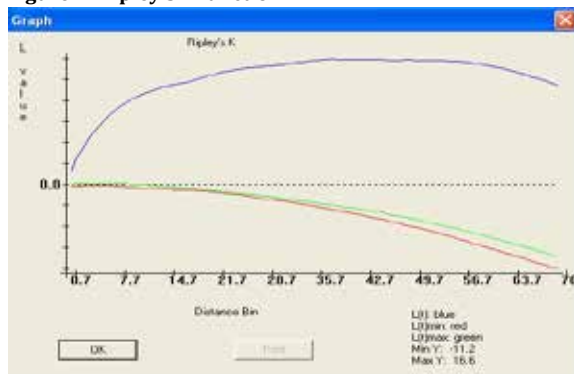
tual K value in the plot. The presence of the dark line outside the maximum or minimum limits indicates significant clumping (if above the maximum value) or significant hyper-dispersion (if below the minimum value).

Table 1: Nearest Neighbor Analysis of cancer incidences in the western part of Tamil Nadu using CrimeStat

Sample size	775
Mean nearest neighbor distance	23.76 km
Standard dev of nearest Neighbor Distance	17.77km
Minimum Distance	1.43 km
Maximum Distance	2733.49 km
Nearest Neighbor Index	0.6295
Test statistic	-19.7303
p-value (one tail, two tail)	0.001

This hypothesis was applied for the cancer data of the western Tamil Nadu using CrimeStat III as mentioned earlier and the calculated values of K function was used to prepare a graph and presented as Figure 4. From the graph, it is observed that the L values are away from the boundary (envelope) which clearly indicates that clustering of the cancer cases was observed. The Ripley's K function confirms the earlier NNI as presented in Table 1.

Figure 4: Ripley's K function



Moran's I

Moran's *I* coefficient of autocorrelation quantifies the similarity of an outcome variable among areas that are defined as spatially related (Moran, 1950). Moran's *I* statistic is given by

$$I = \frac{n \sum_i \sum_j W_{ij} (Z_i - \bar{Z})(Z_j - \bar{Z})}{(\sum_i \sum_j W_{ij}) \sum_k (Z_k - \bar{Z})^2} \quad \text{.....(2)}$$

Z_i - residuals of $O_i - E_i$

W_{ij} - Measure of the closeness of areas *i* and *j*

A weighted matrix is used to define the spatial relationships so that regions close in space are given greater weight when calculating the statistic than those that are distant (Moran 1950). Moran's *I* value lies between +1 to -1. Moran's *I* of 0 indicates the null hypothesis of 'no' clustering and positive Moran's *I* indicates positive autocorrelation (clustering of areas of similar attribute values) and negative coefficient indicates negative autocorrelation (that neighboring areas tend to have dissimilar attribute value).

In the present study, the value for Moran's *I* is observed to be 0.092 which is a positive value oscillating near 0. Therefore it is inferred that there exists a partial positive spatial autocorrelation in the point data (Table 2).

Table2: Spatial autocorrelation for point data

S.No	Spatial Autocorrelation	Value
1.	Moran's I	0.092272
2.	Geary's C	1.179323

Geary's C

Moran's *I* consider similarity between neighboring regions where as Geary's *C* considers similarity between pairs of regions. Geary's *C* or Geary's contiguity ratio is another weighted estimate of spatial autocorrelation.

$$C = \frac{(N-1) \sum_i \sum_j w_{ij} (X_i - X_j)^2}{2W \sum_i (X_i - \bar{X})^2} \quad \text{.....(3)}$$

where N is the number of spatial units indexed by *i* and *j*; *X* is the variable of interest; $\bar{X}$ is the mean of *X*; *w_{ij}* is a matrix of spatial weights; *W* is the sum of all *w_{ij}*.

Geary's *C* value ranges from zero to two, with zero indicating perfect positive spatial autocorrelation and two indicating perfect negative spatial autocorrelation for any pair of regions. In our study the value is 1.179 which is oscillating between 0 and 2 that indicates partial positive spatial autocorrelation which is confirmed by Moran's *I* (Table 2).

Discussion

Interpretation of cancer data requires additional knowledge and further analytical approach for cluster detection and visualization of hot spots. Point pattern analysis is performed (stochastic statistical method) to group variables or observations into strongly interacted subgroups. Kernel density estimation is applied to transform the point events into a more or less smoothed continuous surface. With an underlying population at risk, kernel density estimation represents the relative risk of disease occurrences. This approach attempts to estimate even low event frequencies across the study area based on the point patterns. The recorded cancer data exhibits a clustered pattern which is confirmed by NNI and Ripley's *K* test. To the same cancer data, spatial autocorrelation statistics is applied to determine the degree of similarity observed among neighboring values over a study area. Moran's *I* and Geary's *C* test indicated that there exists partial positive autocorrelation in the point data. Point pattern analysis techniques provide powerful hypothesis-testing capabilities but their complexity means that they are not intuitive to non-specialists. Integrating these techniques into GIS provides useful forms of presentation and the methods employed are relatively intuitive in the field of environmental health. Once the spatial patterns of cancer are identified, public health researchers and other researchers should be able to conduct case-control, retrospective cohort and observational studies to follow up.

REFERENCE

1. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin D. Estimates of worldwide burden of cancer in 2008: GLOBOCAN2008. *Int J Cancer* 2010; 127: 2893–917 | 2. Jha P. Avoidable global cancer deaths and total deaths from smoking. *Nat Rev Cancer* 2009; 9: 655–64. | 3. Registrar General of India and Centre for Global Health Research. Causes of death in India, 2001–2003: Sample Registration System. New Delhi: Government of India, 2009. | 4. Rajesh Dikshit, Prakash C Gupta, Chinthanie Ramasundarahettige, Vendhan Gajalakshmi, Lukasz Aleksandrowicz, Rajendra Badwe, et al : Cancer mortality in India: a nationally representative survey, *Lancet* 2012; 379: 1807–16 | 5. Elliott P, Wakefield J C, Best N G and Briggs D (2000) "Spatial Epidemiology: Methods and Application". Oxford University Press. | 6. Lee C V and Irving J L (1999) "Sources of spatial data for community health planning". *Journal of Public Health Management and Practice*, 5: 7-22 | 7. Pickle L W (2002) "Spatial analysis of disease". In: Beam C, ed. *Biostatistical Applications in Cancer Research*. Boston: Kluwer Academic Publishers; pp. 113-150. | 8. Rushton G (1999) "Methods to evaluate geographic access to health services". *J Public Health Manag Pract*, 5(2): 93-100. | 9. Roux D A V (2001) "Investigating neighborhood and area effects on health". *Am J Public Health*, 91:1783-1789. | 10. Mason T J, McKay F W, Hoover R N, Blot W J and Fraumeni J F Jr. (1975) "Atlas of cancer mortality for U.S. counties: 1950-1969". Bethesda, MD, U. S. Department of Health, Education, and Welfare, USGPO (DHEW pub no (NIH) 75-780).. | 11. Dorn H F and Cutler S J (1955) "Morbidity from cancer in the United States. I. Variation in incidence by age, sex, race, marital status, and geographic region". *Public Health Service*; pp.1-121. [Public Health Monograph] | 12. Ali M, Wagatsuma V, Emch M and Breiman R F (2003) "Use of geographic information system for defining spatial risk for Dengue transmission in Bangladesh: Role for Aedes Albopictus in an urban outbreak" *Am.J. Trop. Med Hyg.* 69(6): 634-640 | 13. Moran P A P (1950) "Notes on continuous stochastic phenomena". *Biometrika* 37, 17-23. | 14. Shamsul Azhar Shah, Hui-min Neoh, Syed Sharizman Syed Abdul Rahim, Zahir Izuan Azhar1,3, Mohd Rohaizat Hassan, Nazarudin Safian, Rahman Jamal, Spatial Analysis of Colorectal Cancer Cases in Kuala Lumpur, *Asian Pacific Journal of Cancer Prevention*, Vol 15, 2014, 1149-1154, | 15. Khalid Al-Ahmadi and Ali Al-Zahrani, Spatial Autocorrelation of Cancer Incidence in Saudi Arabia, *Int. J. Environ. Res. Public Health* 2013, 10, 7207-7228 |