DOI: 10.7860/JCDR/2020/43191.13452

Review Article

Statistics Section

Spatial Cluster Models: Model to Predict Disease Casual Association with Physical, Social and Environmental Risk Factors in Public Health Research

GLADIUS JENNIFER HIRUDAYARAJ[1], M BAGAVANDAS[2]

ABSTRACT

Spatial clustering will help us to identify spatial pattern and also predict geographical factors associated with disease. Spatial cluster models are classified as Global, Local and Focused clusters. This article aims to discuss various types of spatial cluster models such as Moran I, Geary C, Tango EET, CUSUM, GAM, K function, Scan statistic and other with suitable examples which will sensitise the medical researchers about this technique.

Keywords: Focused clusters, Global clusters, Local clusters

INTRODUCTION

Epidemiology defined by John Last, "the study of the distribution and determinants of health related states or events in specified populations, and the application of this study to the control of health problems" [1]. When neighbourhood factors plays significant association with disease then spatial studies will be appropriate. Spatial epidemiological analysis are use to describe spatial patterns, identification of disease clusters, and explanation or prediction of disease risk. The data is said to be spatial data if it contains geo codes such as zip/village/town codes, with their positions latitude and longitude of the observations along with the attributes of the study. The spatial statistical methods in public health data will help us to evaluate differences in rates observed from different geographic areas, identify disease "clusters," and assess the significance of potential exposures. The objective of spatial cluster is to identify the locations, magnitudes and shapes of statistical significance pattern description. This will be explained either by using maps or by spatial models. In other words, it describes the spatial aggregation of disease events but as the observed spatial pattern can merely be a purpose of the distribution of the population at risk/various features. In statistical way, it quantifies a relevant aspect of spatial pattern. The present review paper discusses about the different spatial cluster models depending upon the type of data and its hypothesis.

DISCUSSION

There are numerous approaches for evaluating clusters as either specific or non-specific; In order to assess whether clustering is apparent throughout the study region or not Global (non-specific) clustering methods are used but it does not identify the position of clusters [2]. They provide a single statistic that measures the degree of spatial clustering, the statistical significance of which can then be assessed [2]. Local (specific) methods of cluster detection define the positions and range of clusters, and can be further divided into focused and non-focused tests [2]. Non-focused tests recognise the location of all possible clusters in the study region, while focused tests examine whether there is an increased risk of disease around a pre-determined point *clustering* applies to global methods while *cluster detection* refers to local methods of analysis [2].

These cluster association will be calculated either by boundary over lap (map), cluster over lap and cluster morphology (heterogeneity, multiple testing) [2]. Basically, spatial clustering algorithms identify homogeneous groups of objects based on the values of either attributes: Partition, Hierarchal, Divisive, Grid based and Locality based algorithms.

Global Clusters

Global clusters are used to measure whether clustering is apparent all through the study section but do not identify the local clusters [2]. It measures the degree of spatial clustering significance of which can be assessed. Usually the null hypothesis will be, no cluster exists in given area [2]. The choice global cluster models may be differing from type data set. For aggregated data Moran I, Geary C, EET, MEET will be used; for point data k nearest neighbourhood, Riply k function and cusum; regarding space-time clustering knox, k function, EMM are used to find global spatial clusters.

I. Aggregated Data

1. Moran I

Global Moran I quantify the similarity of an outcome variable among areas that are defined as spatially related. It is similar like Pearson correlation ranged -1 to +1, 0 indicates no cluster, +1 indicates positive spatial auto correlation, -1 indicates negative spatial auto correlation. This method is mainly used for continuous data, it can also be useful for count data set, if any observed autocorrelation may be the result of variation in regional population size rather than any genuine spatial pattern in the disease counts [2]. This method will not adjust for heterogeneous population density; adjusted for population size. It is mainly used to find similarity between the neighbour regions. For example, Kadian S et al., used this method to understand burden of anaemia among children in north eastern states India [3].

2. Geary C (contiguity ratio)

It is weighted estimates of spatial auto correlation; it considers similarity between pairs of region and ranged from 0 to 2. 0 indicate perfect positive spatial auto correlation and 2 indicate negative spatial auto correlation [2]. Kadian S et al., used this method to understand burden of anaemia among children in north eastern states of India [3].

3. Tango Excess events test (EET/MEET)

It measures the closeness among regions based on distant matrix [2]. It is weight sum of excess events; the test statistics considers the difference between observed and expected rate of cases and then weight of these differences by a measure of the distance

between the regions [2]. Higher the weight closer the location. It has parameter λ, choosing larger λ makes test sensitive to large scale clustering and smaller for small scale clusters [2]. Maximum excess event test searches for λ which gives smallest p-value of observed value of the test statistic. It is based on exponential weight functions of two distances [2]. EET method was used to identify significant clustering of brain cancer mortality rates among adults, whereas previous studies used MEET to identify disease cluster of prostate cancer incidence [2].

II. Point Data

1. Cuzick and Edwards k nearest neighbour test

This test is used to study the spatial clustering that takes into account the potentiality heterogeneous distribution of population at risk [2]. It is based on the location of cases and it randomly select controls from specified region. It has parameter k, which is spatial scale based on number of nearest neighbourhoods not geographical distance. Perez A et al., used this method to identify spatial clustering of bovine tuberculosis in Argentina [4].

2. Ripley k function

This method describes the spatial analysis between events of same type and it identifies the distribution at which clustering occurs [2]. It is second order analysis, variance of k increases with increasing distance; it estimates general tendency towards clustering over distances that are small compared with size of the region [2]. It does not depend on the shape of the study region and precise spatial locations of events are used in its estimation [2]. Vadrevu KP and Badarinath KVS, used this method to study the synoptic behaviour of fire events and spatial pattern using nine years integrated fire count dataset [5].

3. Rogerson's Cumulative Sum (CUSUM)

It is a cumulative sum statistic for distinguishing changes in spatial pattern using modified version of Tango [6]. It helps us to detect emerging clusters shortly after they occur. Menotti J et al., used this test to detect early evidence of an outbreak in monitoring of nosocomial invasive aspergillosis [7].

III. Space time clustering

Space time interactions are important to determine whether the disease is infectious or not; it will evaluate whether cases are close in space are also close in time [2]. Under this classification knox, k function, Mantel and Barton tests are there.

1. Knox test

This method will identify spatio temporal clustering of disease events. In this method pair of cases separated by less than a user defined critical space distance are considered to be near in space and also near in time [2]. This will yield 2×2 contingency table and test statistic compared with simulated results under poison model [2]. Marshall JB et al., used this method in detection of clusters of events occurring close together both temporally and spatially in disease outbreaks within a geographical region [8].

2. Space time k function

It is a secondary data analysis to study the space time interactions in point process data [2]. This density function estimates cumulative number of cases expected within distance and time interval which attribute to interaction between space and time [2]. In other words, it detect the space time distribution pattern of incidents occurring in network. Xu X and Peng Z, used this analysis to study space time interaction on road network data [9].

3. Ederer Myers Mantel (EMM) test

EMM is used to detect time clustering in several times series, this test is insensitive to differences in population size over the areas from which the time series originate [2]. Case counts in each time interval are required. This method is cell occupancy approach for exploring space time clustering whereby study region is divided into series of space time sub regions [2]. Stark CR and Mantel N, studied spatial pattern of number of measles cases in various countries in period of five years by using this technique [10].

4. Mantels test

This method is use to compare inter event distances in space and time; it is calculated by sum of all pairs of cases of the spatial distances multiplied by the time distances [2]. Transformations will be used to reduce the effects of large space and time distances which would not be expected to be correlated for infectious diseases [2].

5. Barton's test

This test is use to detect changes in spatial pattern associated with the passage of time, based on ANOVA. It is considered that the spatial patterns will change with time [2]. This test is applied to study space time clustering of cases of childhood leukaemia in England [2].

6. Jacquez's k nearest neighbour test

This method is used to study space time interaction; help us to find association between time and space adjacencies. This method seems to be better than mantel and knox when considering power against cluster size [2].

The Choice of spatial models regarding types of data and hypothesis is described in [Table/Fig-1].

Research condition	Global cluster models	Local cluster models	Focused cluster models
Aggregated data	1. Global Moran I 2. Geary C 3. Excess event/Tango MEET	1. Local Moran I 2. Gi(d) statistics	1. Stone's test 2. Lawson-Waller score test 3. Bithell's linear risk score test 4. Diggle's test
Point data	1. Cuzick and 2. Edwards k nearest neighbour test 3. Ripley k function 4. Rogerson's cumulative sum	1. Open shaw's GAM 2. Turnbull's CEPP 3. Kulldorff's spatial scan statistics	
Space time data	1. Knox test 2. Space time k function 3. Ederer Myers Mantel (EMM) test 4. Mantels test 5. Barton's test 6. Jacquez's k nearest neighbour test		Spatial scan statistics

[Table/Fig-1]: Choice of spatial models regarding types of data and hypothesis.

Local Clusters

The locations and extent of clusters and can be further divided into focused and non-focused. The Focused clustering consider whether there is an increased risk of disease around a pre-determined point, wherein non-focused clustering the location of all likely clusters in the study region [2]. These are also classified further for aggregated, point and space time in local spatial cluster analysis.

I. Local Aggregated data

1. Getis and Ordis local Gi(d) statistics

It is a sign of local clustering that measures the concentration of a spatially distributed characteristic variable [2]. Gi identifies hotspots in spatial data. It helps us to quantify the pattern and intensity of spread of disease away from the core of a hotspot by estimating a series of local statistics at different time periods [2]. Dangisso MH et al., studied Spatio-temporal analysis of smear positive tuberculosis in the sidama zone, southern Ethiopia [11].

2. Local Moran I (LISA)

It is also known as Local Indicators of Spatial Association. It detect local spatial auto correlation in aggregated data by decomposing global Moran I into contribution of each area within study region; also use to test outliers in global spatial pattern in the form of Moran scatter plot [2]. When plot vector of observed values with weighted average of neighbourhood values it yield four types of associations

as quadrants High-High, High-Low, Low-High, Low-Low. It indicates the stability of spatial association throughout the data [2]. This method identifies more localised clusters compare to spatial scan statistics [2]. Example: Bhunia GS et al., used to study spatio-temporal variation and hotspot detection of kalaazar disease [12]

II. Local Point Data

1. Open shaw's geographical analysis (GAM)

This method help us to explore disease data for evidence of spatial pattern; with the application of fine grid across a study area and making a sequences of circles of variable radii with their centres based at each intersection of the grid [2]. Besag J and Nevell J developed a new method to rectify the problems in GAM by defining k expected cluster size where k ranged 2 to 10 for rare diseases. Each area with non-zero cases is considered in turn as the center of a possible cluster [13].

2. Turnbull's cluster evaluation permutation procedure (CEPP)

The first test was able to locate and test the significance of disease clusters; it creates a circular window for each area that contains a pre-determined number of individuals at risk [2]. A series of overlapping windows is created with a population at risk of constant size. The cluster size has a defined priori for the procedure to be valid [2]. This method provides a quantitative assessment of the statistical significance of identified clusters.

3. Kulldorff's spatial scan statistics

Spatial scan method brings advantages of previous methods together in spatial scan series of circles of varying radii for specified location [2]. Each circle absorbs nearest neighbouring locations that fall inside it and the radius of each circle is set to increase continuously from zero until some fixed percentage of the total population is included [2]. Tiwari N et al., used scan statistics to investigate geo spatial hotspots for the occurrence of tuberculosis [14].

Focused Clusters

Focused cluster method has assumption of location of cluster centre, is specified priori and likelihood of that location truly being a cluster centre is then determined [2]. Stones test, Lawson Waller score test, Bithells linear risk score and Diggles test are types of focused local cluster methods.

1. Stone's test (NCD)

Stone test developed a class of tests for trend, the MLR and Poisson maximum tests, both of which use the first isotonic regression estimator, considering there will be a monotonic delay of risk with increasing distribution from any point source of disease [2]. This test is mainly used in non-communicable disease clustering like cancer incidence.

2. Lawson-Waller score test

This also known as uniformly most powerful test, it detect decreasing trend in disease frequency associated with declining exposure to a point focus [2]. Mostly used in surveillance data. Waller LA used this, to detect leukaemia incidence and Trichloroethylene (TCE) contaminated dumpsites [15].

3. Bithell's linear risk score test

This is also known as linear risk score test, it calculates the disease incidence is weighted by some distance function from a point score [2]. Mainly used to detect environmental hazard or distribution from Focui for given disease by using reciprocal of distance [2].

4. Diggle's test

Non parametric kernel smoothing to describe natural variation of disease and then MLR to evaluate the possibility of raised incidence around a pre specified point source [2]. Diggle JP also developed a conditional approach which converts the original point process into non-linear binary regression model for the spatial variation in risk. Diggle JP explored this test to find association of disused industrial incinerator with incidence of laryngeal cancer and compared it with other lung cancers [16].

CONCLUSION(S)

This article explained various types of cluster models, the research investigator have to choose appropriate model in order to obtain a précised significance. This paper aims to sensitise the research scholars about these very useful analytical research techniques and to empower them to carry out research in a better manner. It helps us to carry out spatial cluster analysis with proper steps and graphical presentation techniques.

REFERENCES

[1] Park K, Chapter 3, Textbook of social and preventive medicine, 24th edition.
[2] Dirk UP, Timothy PR, Mark S, Kim BS, David JR, Archie CAC. Spatial analysis in epidemiology. Oxford University Press, 2008;32-64.
[3] Kadian S, Kaur A, Singh KJ. Understanding the burden of anaemia among children in north eastern states india: Evidence from NFHS. Demography India. 2017;138-42.
[4] Perez A, Ward MP, Torres P, Viviana R. Use of spatial statistics and monitoring data to identify clustering of bovine tuberculosis in Argentina. Preventive Veterinary Medicine. 2002;56(1):63-74.
[5] Vadrevu KP, Badarinath KVS. Spatial pattern analysis of fire events in central India- A case study. Geocarto International. 2009;24(2):115-31.
[6] Tango T. A class of test for detecting general and focused clustering of rare diseases. Statistics in Medicine. 1995;14(21-22): 2323-34.
[7] Menotti J, Porcher R, Ribaud P, Lacroix C, Jplivet V, Hamane S, et al. Monitoring of nosocomial invasive aspergillosis and early evidence of an outbreak using CUSUM. Clin Microbial Infect. 2010;16:1368-74.
[8] Marshall JB, Spitzner DJ, Woodall WH. Use of the local knox statistic for the prospective monitoring of disease occurrences in space and time. Stat Med. 2007;26(7):1579-93.
[9] Xu X, Peng Z. The K function analysis of space time point pattern on road network, China. 19th International Conference on Geoinformatics, Geoinformatics. 2011. 1-5. Available at: https://www.cpgis.org/userfiles/file/finalProgram(3).pdf. Doi:10.1109/GeoInformatics.2011.5981103.
[10] Stark CR, Mantel N. Temporal spatial distribution of birth dates for Michigan children with leukemia. Cancer Research. 1967;27(1):1749-75.
[11] Dangisso MH, Datiko DG, Lindtjorn B. Spatio- temporal analysis of smear positive tuberculosis in the sidama zone, southern Ethiopia. Plos One. 2015;10(6):e0126369.
[12] Bhunia GS, Kesari S, Chatterjee N, Kumar V, das P. Spatial and temporal variation and hotspot detection of kalaazar disease in vaishali district India. BMC Infectious Diseases. 2013;13:64.
[13] Besag J, Nevell J. The detection of Clusters in rare diseases. Journal of Royal Statistical Society. 1991;154(1):143-55.
[14] Tiwari N, Adhikari CMS, Tewari A, Kandpal V. Investigation of geo spatial hotspots for the occurrence of tuberculosis in Almora district, India using GIS and spatial scan statistic. International Journal of Health Geographics. 2006;5:33.
[15] Waller LA. Chronic disease surveillance and testing of clustering of disease and exposure: Application to leukemia incidence and TCE contaminated dumpsites in upstate Newyork. Environmetrics. 1992;3(3):281-300.
[16] Diggle JP. A Point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. Journal of the Royal Statistical Society. Series A (Statistics in Society). 1990;153(3): 349-62.

PARTICULARS OF CONTRIBUTORS:
1. Assistant Professor, Department of Community Medicine, Karpaga Vinayaga Institute of Medical Sciences and Research Centre, Madhuranthakam, Tamil Nadu, India.
2. Professor, School of Public Health, SRM University, Chennai, Tamil Nadu, India.

NAME, ADDRESS, E-MAIL ID OF THE CORRESPONDING AUTHOR:
Dr. Gladius Jennifer Hirudayaraj,
4/44, Varadaraja Nagar, Anumanthaputherri, Chengalpattu-603001, Tamil Nadu, India.
E-mail: gladiusjennifer@gmail.com

PLAGIARISM CHECKING METHODS: [Jain H et al.]
- Plagiarism X-checker: Nov 09, 2019
- Manual Googling: Dec 11, 2019
- iThenticate Software: Dec 28, 2019 (23%)

ETYMOLOGY: Author Origin

AUTHOR DECLARATION:
- Financial or Other Competing Interests: No
- Was Ethics Committee Approval obtained for this study? No
- Was informed consent obtained from the subjects involved in the study? NA
- For any images presented appropriate consent has been obtained from the subjects. NA

Date of Submission: **Nov 09, 2019**
Date of Peer Review: **Dec 02, 2019**
Date of Acceptance: **Dec 11, 2019**
Date of Publishing: **Jan 01, 2020**