Austin Biometrics and Biostatistics

Open Access

Austin Publishing Group

Review Article

Current Statistical Methods for Spatial Epidemiology: A Review

Osei FB*
Department of Mathematics and Statistics, University of Energy and Natural Resources, Ghana

***Corresponding author:** Osei FB, Department of Mathematics and Statistics, University of Energy and Natural Resources, Sunyani, Ghana

Received: October 13, 2014; **Accepted:** November 11, 2014; **Published:** November 25, 2014

Abstract

The current advances in technology and disease surveillance systems have often made available the spatial/geographical orientation of disease occurrences. Statistical analysis of such data is often complicated by the spatial structure of the data which manifest itself as spatial autocorrelation. Methods to account for spatial autocorrelation rarely found in the mainstream classical statistics literature. However, current practices in spatial epidemiology seek to unveil and understand the spatial distribution of diseases. Therefore any determination to model spatial autocorrelation is a non-trivial effort which complements the classical statistics approaches. The objective of this review is to discuss the current statistical methods in spatial epidemiology as well as their relative weaknesses. Much attention and focus is provided for methods which are relatively advantageous and widely used.

Keywords: Statistical methods; Spatial epidemiology; Cluster analysis; CAR; GIS methods

Introduction

Spatial epidemiology is the study of the spatial/geographical distribution of disease incidences and its relationship to potential risk factors. Knowledge of the spatial variations of diseases and characterization of its spatial structure is essential for the epidemiologist to better understand the population's interaction with its environment. The origin of spatial epidemiology dates back to 1855 with the classic epidemiologic studies of John Snow on cholera transmission. Snow's study of London's cholera epidemic provides one of the most famous examples of spatial epidemiology. Mapping the locations of cholera victims, Snow was able to trace the cause of the disease to a contaminated water source. Spatial analysis in the nineteenth and twentieth century mostly took the form of plotting the observed disease cases or rates [1]. Advances in technology now allow not only disease mapping but also the application of spatial statistical methods, such as cluster analysis [2,3] and ecological analysis [4-6] in epidemiological research. Geographic Information System (GIS) methods and modern statistical methods allow an integrated approach to address both tasks; i.e. inference on the geographical distribution of the disease and its prediction at new locations. Many diseases are influenced by environmental variables, and since these variables are spatially continuous in natures, the disease rates tend to exhibit spatial dependency, popularly known as spatial autocorrelation. Thus such patterns of spatial autocorrelation confirm the natural law of nature, popularized by Tobler [7] as the first law of geography: "Everything is related to everything else, but near things are more related than distant things". The use of standard/classical statistical techniques for modeling spatially distributed diseases either leads to over estimation or under estimation of parameters in question. The objective of this manuscript is to provide a review of the current statistical methods that are useful in analyzing and modeling spatially distributed diseases, their relative weaknesses and strength. Much attention and focus is provided for methods which are relatively advantageous and widely used.

Cluster Analysis

Fundamental to the spatial epidemiologist is the investigation of possible disease clustering. Cluster analysis provides opportunities for the epidemiologist to understand the spatial distribution of diseases and the possible association between demographic and environmental exposures [8-11]. Searching for disease clustering involves an assessment of local or global accumulation of the disease incidences [12,13]. The focus of global cluster analysis is to determine the presence or absence of clustering in the whole study region. There are numerous methods for testing global clustering, including those proposed by Alt and Vach [14], Besag and Newell [8], Cuzick and Edwards [15], Diggle and Chetwynd [16], Grimson [17], Moran [18], Tango [19-21], Walter [22-24] and Whittemore et al. [25]. The most widely used measure of global clustering in epidemiology is the method proposed by Moran [18]. Moran's Index is a weighted correlation coefficient that is used to measure deviation from spatial randomness. The Index I_M statistic is similar to the Pearson correlation coefficient [18,26,27] with the form:

$$I_M = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(r_i - \bar{r})(r_j - \bar{r})}{\sum_i (r_i - \bar{r})^2}, \qquad (1)$$

where N is the number of spatial objects, w_{ij} is the element in the spatial weights matrix corresponding to the spatial object pairs i, j; and r_i and r_j are the observed rates for objects i and j with mean rate $\bar{r}$. When the weights are not row-standardized, the scaling factor N/S_o is applied, such that $S_o = \Sigma_i \Sigma_j w_{ij}$. Values range from −1 (indicating perfect dispersion) to +1 (perfect clustering or deviation from randomness). Negative (positive) values indicate negative (positive) spatial autocorrelation.

Deviation from spatial randomness indicates specific spatial arrangements of geographic location information such as clusters [18]. Although Moran's Index was originally developed to analyze continuous data, its application to analyze count data of health events

Citation: Osei FB. Current Statistical Methods for Spatial Epidemiology: A Review. Austin Biom and Biostat. 2014;1(2): 7.

is enormous [28-31]. Other health applications of Moran's Index include studies of Kitron and Kazmierczak [32] of Lyme disease in the Wisconsin state, studies of Glick of cancer in Pennsylvania, the geographical distribution of human giardiasis in Ontario, Canada [33], Lyme disease in the New York state [28], and the geographical patterns of cholera in Mexico [34].

Global cluster analysis can obscure local effects since the assumption of stationary is rarely met. Local cluster analysis defines the characteristics of the clusters, such as size, location and intensity. Several formal methods and techniques for identifying local disease clusters have been developed for both point and areal level data [8,9]. Examples of local clustering methods include spatial correlograms [35-39] the Local Indicator of Spatial Association [40], the local G_i* statistics [41], Ripley's *K*-function [42-44], Cluster Evaluation Permutation Procedure (CEPP) [45], the Knox test [46,47], and Kulldorff's spatial scan statistic [2]. Other methods for space-time clustering include Mantel's test [48], Ederer-Meyer-Mantel test [49], Barton's test [50], Diggle et al. test [51], Jacquez's *k* nearest neighbors test, and Kulldorff's space-time scan statistic [2].

The spatial scan statistic developed by Kulldorff [10,11,52] offers several advantages over the others: (1) it corrects for multiple comparisons, (2) it adjusts for the heterogeneous population densities among the different areas in the study, (3) it detects and identifies the location of the clusters without prior specification of their suspected location or size thereby overcoming pre-selection bias, (4) and allows adjustment for covariates. Also Kulldorff's spatial scan statistic is both deterministic (i.e., it identifies the locations of clustering) and inferential (i.e., it allows for hypothesis testing and evaluation of significance). The spatial scan statistic has been used to detect and evaluate various disease clusters including leukemia [9,53], cancer [10,45,53-56], giardiasis [57], tuberculosis [58], diabetes [59], Creutzfeldt-Jacob disease [60], granulocytic ehrlichiosis [61], and amyotrophic lateral sclerosis [62].

The flexible spatial scan statistic is a recent cluster detection methodology developed by Takahashi and Tango. This approach is based on the original idea of Kulldorff. Unlike Kulldorff's approach, however, which imposes a circular window to define the potential cluster areas [9], Takahashi and Tango's flexible spatial scan statistic imposes an irregularly shaped window on each region connecting its adjacent regions.

For any given location *i*, a set of irregularly shaped windows consisting of *k* connected locations including *i* moves from 1 to a pre-set maximum window size *K* (which is proportional to the population at risk). To avoid detecting a cluster of an unlikely peculiar shape, the connected locations are restricted as the subsets of the set of location *i* and (*K* - *1*)-nearest neighbors to location *i*. In effect a very large number of different but overlapping arbitrarily shaped windows are created. For location *i*, the flexible scan statistic considers *K* concentric circles plus all the sets of connected locations, including location *i*, whose centroids are located within the K^{th} largest concentric circle. Let $W_{ik(j),}$ $j=1,\ldots, j_{ik}$ denote the j^{th} window which is a set of *k* regions connected starting from the region *i*, where j_{ik} is the number of *j* satisfying $W_{ij(k)} \subseteq w_{ik}$ for $k = 1,\ldots,K$. Then, all the windows to be scanned are included in the set:

$$W = \{w_{ik(j)} \backslash 1 \leq i \leq m,\ 1 \leq k \leq K,\ 1 \leq j \leq j_{ik}\}. \tag{2}$$

Under the alternate hypothesis, there is at least one window *W* for which the underlying risk is higher inside the window when compared with outside. For each window the likelihood of the observed number of occurrences within and outside the window under the Poisson assumption is computed as:

$$L(\mathrm{W}) = \sup_{W \in \mathbf{W}} \left(\frac{O(W)}{E(W)} \right)^{O(W)} \left(\frac{O(\hat{W})}{E(\hat{W})} \right)^{O(\hat{W})} I\left(\frac{O(W)}{E(W)} > \frac{O(\hat{W})}{E(\hat{W})} \right), \tag{3}$$

Where $\hat{W}$ indicates all the regions outside the window *W*, and *O ( )* and *E ( )* denote the observed and expected number of occurrences within the specified window, respectively. The indicator function *I ( )* is 1 when the number of occurrences within the window is more than the expected number and 0 otherwise. The window W^* that attains the maximum likelihood is defined as the Most Likely Cluster (MLC). This approach is able to detect arbitrarily shaped clusters, and this statistic is well suited for detecting and monitoring disease outbreaks in irregularly shaped areas.

Popular software packages for conducting cluster analysis includes Sat Scan for circular spatial scan statistics developed by Martin Kulldorff [11] and FleX Scan developed by Tango and Takahashi [63] for flexible shaped scan statistics. Sat Scan can implement both purely spatial and space-time cluster analysis; however, these are not yet implemented in FleX Scan. The scan statistics technique has also been implemented the SpatialEpi [64] package of the *R* software for statistical computing.

Ecological Analysis

A significant interest in spatial epidemiology lies in identifying associated risk factors which enhance the risk of infection, the so called *ecological analysis* [65,66] or *geographic correlations studies* [67]. The term ecological analysis is used loosely here to denote associating aggregated disease outcomes with related risk factors or covariates, where inference still remains at the aggregated level.

Classical linear methods

The most prominent method is the classical linear regression model, where the response variable *y* is assumed to be independent normal or Gaussian distributed and covariates, say $x_1,\ldots,x_p$ act linearly on the response. By assumption, the conditional expectation of y is:

$$\eta = E(y \backslash x_1,\ldots,x_p) = \beta_0 + x_1\beta_1 + \ldots + x_p\beta_p\eta\ , \tag{4}$$

where the regression coefficients $\beta_1,\ldots,\beta_p$ determine the strength of the influences of the covariates, and the linear predictor η is the sum of the covariate effects. Here, each observation has an underlying mean of $\Sigma i\ \chi_i\ \beta i$ and normally distributed random error term ε. Generally, the random error term $\varepsilon = (\varepsilon_1,\ldots,\varepsilon_p)$ has zero mean and uncorrelated variance-covariance matrix Σ_σ, i.e. $\varepsilon i \sim N(0,\Sigma\sigma)$, where $\Sigma_\sigma = Var(y) = Var(\varepsilon) = \sigma^2 I$, and *I* is $p \times p$ identity matrix. The assumption of independent observations also implies that $E(\varepsilon_i \varepsilon_j) = E(\varepsilon_i)\ E(\varepsilon_j) = 0$.

For disease counts of small areas with relatively small populations at risk and few observed cases, rates may not follow the assumptions of the linear model. In such cases, a direct connection between the expectation of y and the linear predictor η is not possible. Generalized Linear Models (GLMs) extend the classical linear model for Gaussian responses to more general situations such as binary or count data [68-71] to ensure the appropriate domain of $E(y/x_1,\ldots,x_p)$. By introducing a more general transformation or response function *h*, equation (1.1) can be rewritten as:

$$h(\eta) = E(y\backslash x_1,\ldots,x_p) = h(\beta_0+ \beta_1 x_1+\ldots+x_p\beta_p). \qquad (5)$$

Both the classical linear model and GLMs provide the means to quantify and describe only *first-order* effects or *large-scale* variation in the mean of the disease outcome. These methods ignore *second-order* spatial effects or *small-scale* variations that arise from interactions between neighbors, i.e. spatial autocorrelation. Both methods assume that any spatial pattern observed in the outcome *y* is entirely due to the spatial patterns in the covariates; therefore, no residual spatial variation is accounted for. If an important covariate is inadvertently omitted, however, estimates of β will be biased [72], and if this covariate varies spatially, residual spatial variation will often manifest itself as spatial autocorrelation in the residual process. Hence when these methods are used to analyze spatially correlated data, the standard error of the covariate parameters would be underestimated and thus the statistical significance would be overestimated [73].

Spatial methods

Spatial statistical methods, such as spatial regression, incorporate spatial autocorrelation according to the way spatial neighbors are defined. A spatial regression model may be parameterized as equation (4). A modification of the variance-covariance matrix Σ is then required to allow spatially correlated error terms. Common methods to incorporate spatially correlated error terms in the variance-covariance matrix Σ_σ is the Simultaneous spatial Autoregressive (SAR), Conditional spatial Autoregressive models (CAR), and Spatial Moving Average models (SMA). Both the SAR and CAR correspond to autoregressive procedures in time series analysis [43]. These models are well explained in Cliff and Ord [26], Haining [74], Ripley [43], and Cressie [73].

Under CAR model specification, the conditional expectation of the response variable *y* is specified as

$$\eta = E(y\backslash x_1,\ldots,x_p) = \beta_0+x_1\beta_1+\ldots+x_p\beta_p+\rho w[y-(\beta_0+x_1\beta_1+\ldots x_p\beta)]+\varepsilon, \qquad (6)$$

which can be surmised as

$$\eta = E(y\backslash x_1,\ldots,x_p) = \Sigma x\beta+\rho w(y-\Sigma x\beta)+\varepsilon\ , \qquad (7)$$

and simplified in matrix notation as $\mathbf{Y = X\boldsymbol{\beta} +\boldsymbol{\rho} W(Y-X\boldsymbol{\beta})+\boldsymbol{\varepsilon}}$. The error terms assumed normally distributed with zero mean and variance-covariance matrix Σ_σ, i.e. $\boldsymbol{\varepsilon} \sim \mathbf{N}\ (\mathbf{0},\boldsymbol{\Sigma}_\sigma)$ expressed in terms of the spatial connectivity/structure of the data. Thus, $\boldsymbol{\Sigma}_\sigma= \boldsymbol{\sigma}^2\,(\mathbf{I}-\boldsymbol{\rho}\mathbf{W})$, where $W = w_{ij}$ is a spatial weight matrix that describes the spatial connectivity/dependency between the locations *i* and *j*. several specifications of elements in w_{ij} may be constructed including:

$$\mathbf{W} = w_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ share common boundary, otherwise } 0 \\ d^{-1} & \text{where } d \text{ is the distance between } i \text{ and } j \\ 1 & \text{if distance between } i \text{ and } j \text{ is} < \text{some threshold, otherwise } 0 \\ 1 & \text{for } k \text{ nearest neighbors, otherwise } 0 \end{cases}$$

CAR models restrict the spatial weight matrix to be symmetrical and therefore not suitable for modeling directional processes. Also, the *k* nearest neighbor connectivity option for w_{ij} generates as asymmetric neighborhood structure and therefore not suitable for CAR models.

SAR model on the other hand can be specified under three different variants. As spatial lagged model, as spatial error model oras spatial lagged mixed-model. Unlike CAR models, the neighborhood connectivity matrix **W** in the SAR model need not be symmetrical.

For a spatial lagged model, spatial autocorrelation is included as an additional predictor in the form of spatially lagged dependent variable. Thus $\mathbf{Y = \boldsymbol{\rho} Y^{*}+X\boldsymbol{\beta}+\boldsymbol{\varepsilon}}$, where the lagged dependent variable is $\mathbf{Y^{*} = WY}$, which finally yields

$$\mathbf{Y =(1- \boldsymbol{\rho} Y)^{-1}X\boldsymbol{\beta}+(1- \boldsymbol{\rho} Y)^{-1}\ \boldsymbol{\varepsilon}}. \qquad (8)$$

Where it is believed that the autoregressive process occur only in the error terms rather than either the in the response or in the predictor, the OLS model $\mathbf{Y = X\boldsymbol{\beta}+\boldsymbol{\varepsilon}}$ is complemented by a spatially lagged error term of the form $\boldsymbol{\varepsilon}=\boldsymbol{\lambda}\mathbf{W}\boldsymbol{\varepsilon}+\boldsymbol{\acute{\varepsilon}}$. This yield

$$\mathbf{Y = \boldsymbol{\lambda} W\boldsymbol{\varepsilon}+X\boldsymbol{\beta}+\boldsymbol{\acute{\varepsilon}}}, \qquad (9)$$

Where $\boldsymbol{\acute{\varepsilon}} \sim \mathbf{N}\ (\mathbf{0},\boldsymbol{\sigma}\mathbf{I}_\mathbf{n})\ \boldsymbol{\lambda}$ and is the lagged-error variable.

Where it is believed that spatial autocorrelation affects both the response and predictor variables, then another term $\mathbf{WX}\boldsymbol{\gamma}$ which expresses a lagged-decency of the predictor variables is added to the model. This results in a spatial lagged mixed model of the form:

$$Y = \lambda W\varepsilon+X\beta+WX\gamma+\varepsilon, \qquad (10)$$

Where γ expresses the regression coefficient of the lagged-response variable.

Haining [74] expresses the facts that ever SAR model is also a CAR model with $\mathbf{K = S+S^T\text{-}S^TS}$, where **K** is the $\boldsymbol{\rho}\mathbf{W}$ of the CAR model and **S** is the $\boldsymbol{\rho}\mathbf{W}$ of the SAR model.

Numerous software packages have been developed for implementing spatial regression models. Typical amongst them is the free software GeoDa [75] which easily fits both spatial lag and error models. The sped package in R software has vigorous functions for fitting spatial regression models. The comprehensive econometric toolbox developed by LeSage and Pace [76] in MATLAB has numerous functions for fitting spatial regression models.

Generalized structured additive regression

Generalized Additive Models (GAM) also provides a powerful class of models for modeling nonlinear effects of continuous covariates in regression models with non-Gaussian responses. Modeling the nonlinear effects of continuous covariates may be based on smoothing splines [77], local polynomials [78], regression splines with adaptive knot selection [79-81] and P-splines [82,83].

Fahrmeir et al. [84], Brezger [85] and Kneib [86] present a detailed description of Bayesian P-Splines and mixed model based inference in generalized Structured Additive Regression (STAR) based on Bayesian P-Splines. Generalized STAR models are extensions of GAM models which allow one to incorporate small area spatial effects, nonlinear effects of risk factors, and the usual linear or fixed effects in a joint model. Typically, a generalized STAR model is parameterized as:

$$\eta = f_1(x_{i1})+\cdots+f_p(x_{ip})+ f_{spat}(s_i)+ u'_i\gamma\ , \qquad (11)$$

Where $f_1,\ldots,f_p$ are nonlinear functions of the covariates $x_1,\ldots,x_p$. In such models, covariates of the parametric or fixed effects are subsumed in the term $u'_i\gamma$, where γ is an estimate of the fixed effect covariate u_i. The linear combination $u'_i\gamma$ corresponds to the usual parametric part of the predictor. The function f_{spat} (S_i) accounts for spatial effects of the data.

Bayesian Estimation

STAR models are highly parameterized; therefore, inference is based on a fully Bayesian estimation of the posterior distribution of the model parameters rather than maximum likelihood estimation methods. Since the posterior is analytically intractable, the parameter estimates are generated by drawing random samples from the posterior via MCMC simulation techniques.

Bayesian estimation and inference in statistical modeling provides a number of advantages over the classical approaches. This includes a more natural interpretation of parameter intervals, and the ease with which the true parameter density may be obtained. Bayesian approach has recently been given intense focus due to the widespread adoption of Markov Chain Monte Carlo (MCMC) methods. In the past, Bayesian estimation and inference was often daunting due to the requirement of numerical integration. The MCMC estimation method decomposes complicated estimation problems into simpler problems that rely on conditional distributions for each parameter in the model [87]. In classical approaches such as maximum likelihood estimation, inference is based on the likelihood of the data alone. In Bayesian approach, the likelihood of the observed data y given a d dimensional parameter set $\theta = (_1,\ldots, \theta_d)$, denoted as $p(y/\theta)$, is used to modify the prior beliefs $p(\theta)$ with the updated knowledge summarized in a posterior density $p(y/\theta)$. Applying Bayes theorem, $p(\theta/y)= p(y/\theta)$ $p(\theta)$ $p(y)$ is found, where the marginal likelihood $p(y)$ is obtained by integrating the likelihood over the prior densities, i.e. $p(y)=\int p(y/\theta)$ $p(\theta)d()$. Since $p(y)$ can be regarded as a normalizing constant, the posterior density can be simplified as $p(\theta/y)\alpha$ $p(y/\theta)$ $p(\theta)$.

Priors for unknown functions and fixed effects

The unknown functions $f_1,\ldots,f_p$, f_{spat} (S) and the fixed effects γ are considered as random variables and must be supplemented by appropriate prior assumptions. In the absence of any prior knowledge, diffuse prior $p(\gamma)$ $\alpha const$ (may be assigned for the fixed effects. Alternatively, a weak informative multivariate Gaussian distribution may be assigned. For modeling the unknown functions $f_1,\ldots,f_p$, there exists a variety of different approaches. Polynomials of degree l are often not flexible enough for small l, yet estimates become more flexible but also rather unstable for large l, especially at the boundaries [85]. Eilers and Marx [82] suggest specific forms of polynomial regression splines which are parameterized in terms of B-spline basis functions together with a penalization of adjacent parameters, also known as P-splines. For instance, following Eilers and Marx [82], $f(x)$ can be approximated by a polynomial spline of degree l with equally spaced knots $x_j^{\min} = \zeta_{j,0} < \zeta_{j,1} < \cdots < \zeta_{j,s-1} < \zeta_{j,s} = x_j^{\max}$ within the domain of x_j. The assumption that $f(x)$ can be approximated by a polynomial spline leads to a representation in terms of a linear combination of $d=s+l$ basic functions B_m, i.e. $f_j(x_j) = \sum_{m=1}^{d} \xi_{j,m} B_m(x_j)$. Thus, the estimation of $f(x)$ is reduced to the estimation of the vector of unknown regression coefficients $\xi = (\xi_1,\ldots,\xi_m)'$ from the data. Detailed description of Bayesian P-Splines in STAR models can be found in Brezger [85].

Priors for spatial effects

The spatial effect is commonly introduced in a hierarchical fashion via prior distributions of location-specific random effects. Unlike the SAR, CAR, or SMA models, spatial dependencies are estimated for each spatial unit. A major significance of STAR modeling approach is that the spatial effect can be split into spatially structured (correlated) and a spatially unstructured (uncorrelated) effects. Thus, $f_{spat}(s) = f_{str}(s)+f_{unstr}(s)$ where the function $f_{str}(s)$ accounts for spatially correlated effects of the data, whereas the function $f_{unstr}(s)$ accounts for unobserved heterogeneity, occurring locally or at a large scale. The most common prior for modeling the structured spatial effects $f_{str}(s)$ is the Markov random field prior pioneered by Besag [88,89]:

$$p\left(f_{str}(s) \middle| f_{str}(s'), s' \neq s, \tau_{str}^2 \right) \sim N\left(\frac{1}{N_s} \sum_{s' \sim s} f_{str}(s'), \frac{\tau_{str}^2}{N_s} \right). \qquad (12)$$

Here $s\,\varepsilon\,\{1,\ldots,S\}$ represents the locations of connected geographical regions, N_s is the number of geographical neighbors and $s' \sim s$ denotes that geographical locations s' and s are neighbors. The uncorrelated $f_{unstr}(s)$ part may be estimated based on location-specific Gaussian random effects $p(f_{unstr}(s)\backslash\tau^2_{unstr})\sim N(0,\tau^2_{unstr})$. In a fully Bayesian estimation, hyper-priors for the variance parametersτ^2_j, $j=1,\ldots,p$, τ^2_{str} and τ^2_{unstr} are also considered as unknown; therefore, appropriate hyper-parameters have to be assigned. Commonly, highly dispersed, but proper, inverse Gamma distributions $p(\tau^2_j)\sim IG(a_j,b_j)$ with known hyper-parameters a_j and b_j with density function $p(\tau^2_j)$ α $(\tau^2_j)^{-aj-1}exp(-bj/\tau^2_j)$ are assigned in the second stage of the hierarchy.

Different forms of STAR models may be structured for both cross-sectional and longitudinal data. Well known models that can be structured include GAM, Generalized Additive Mixed Models (GAMM), spatial regression models, generalized geoadditive mixed models (GGAMM), dynamic models, varying coefficient models, and geographically weighted regression [90] may be useful within a unifying framework. Detailed description of these models and their applications can be found in Fahrmeir and Lang [91,92], Lang and Brezger [93], Brezger and Lang [94], Eilers and Marx [82], Marx and Eilers [83], Wahba [95], and Hastie and Tibshirani [77].

Much literature has been developed around methodological issues relating to the Bayesian approach [96-103]. Bayesian approaches to GAM are currently either based on regression splines with adaptive knot selection [104-110], or on smoothness priors [77,91,92]. The development and implementation of Markov Chain Monte Carlo (MCMC) methods in software such as WinBUGS [111] and BayesX [112] have made Bayesian estimation approaches simpler.

Conclusion

Space has become, and would continue to be, an essential dimension in epidemiology. This is mainly due to the availability and quality of geographically referenced health data. Thus the relevance of space is unlimited, both in theory and in practice. However, statistical methods for spatial epidemiologic data are limited in the mainstream statistics literature. Many texts in the field of spatial statistics and related fields address the significance of space and theoretical approaches in diverse forms. This manuscript has discussed a wide range of statistical methods useful in spatial epidemiology; focusing on those relevant under two main themes; cluster analysis and ecological analysis. The availability of open source software packages designed to facilitate such methods and techniques has been key and resourceful in their implementation. However, their implantations must be guided by good practice theories within the epidemiologic principles. With these, spatial epidemiologic studies will continue to play a critical key role in the understanding of disease epidemiology, especially the complex relationship between population, health and environment.

References

1. Snow J. On the Mode of Communication of Cholera, 2^{nd} edn. London: John Churchill. 1855.
2. Kulldorff M. A spatial scan statistic. Commun Stat. Theory & Methods. 1997; 269: 1481-1496.
3. Rosenberg MS, Sokal RR, Oden NL, DiGiovanni D. Spatial autocorrelation of cancer in Western Europe. Eur J Epidemiol. 1999; 15: 15-22.
4. Ali M, Emch M, Yunus M, Sack RB. Are the environmental niches of Vibrio cholerae O139 different from those of Vibrio cholerae O1 El Tor? Int J Infect Dis. 2001; 5: 214-219.
5. Ali M, Emch M, Donnay JP, Yunus M, Sack RB. Identifying environmental risk factors for endemic cholera: a raster GIS approach. Health Place. 2002; 8: 201-210.
6. Ali M, Emch M, Donnay JP, Yunus M, Sack RB. The spatial epidemiology of cholera in an endemic area of Bangladesh. Soc Sci Med. 2002; 55: 1015-1024.
7. Tobler W. “A computer movie simulating urban growth in the Detroit region”. Eco Geogr. 1970; 46: 234-240.
8. Besag J, Newell J. The detection of clusters in rare disease. JR. Stat Soc A. 1991; 154: 143-155.
9. Kulldorff M, Nagarwalla N. Spatial disease clusters: detection and inference. Stat Med. 1995; 14: 799-810.
10. Kulldorff M, Feuer EJ, Miller BA, Freedman LS. Breast cancer clusters in the northeast United States: a geographic analysis. Am J Epidemiol. 1997; 146: 161-170.
11. Kulldorff M. Software for the spatial space time statistics, SaTScan v6.0.Information Management Service Inc. 2005.
12. Lawson AB, Denison DGT. Spatial cluster modeling.London: Chapman & Hall/CRC. 2002.
13. Tango T. Statistical Methods for Disease Clustering. New York: Springer. 2010.
14. Alt KW, Vach W. The reconstruction of “genetic kinship” in prehistoric burial complexes: Problems and statistics. Bock HH, Ihm P, editors. In: Classification, Data Analysis, and Knowledge Organization: Models and Methods with Applications. Berlin: Springer. 1991; 299-310.
15. Cuzick JC, Edwards R. Spatial clustering for inhomogeneous populations. J. R. Stat. Soc. B.1990; 52:73-104.
16. Diggle PJ, Chetwynd AG. Second-order analysis of spatial clustering for inhomogeneous populations. Biometrics. 1991; 47: 1155-1163.
17. Grimson RC, Wang KC, Johnson PWC. Searching for hierarchical clusters of disease: spatial patterns of sudden infant death syndrome. Soc Sci Med D. 1981; 15: 287-293.
18. MORAN PA. Notes on continuous stochastic phenomena. Biometrika. 1950; 37: 17-23.
19. Tango T. A class of tests for detecting ‘general’ and ‘focused’ clustering of rare diseases. Stat Med. 1995; 14: 2323-2334.
20. Tango T. Comparison of general tests for disease clustering. Lawson AB, Biggeri A, Bohning, Lesaffre E, Viel J-F, Bertollini R, editors. In: Disease Mapping and Risk Assessment for Public Health. Chichester: John Wiley & Sons. 1999; 111-117.
21. Tango T. A test for spatial disease clustering adjusted for multiple testing. Stat Med. 2000; 19: 191-204.
22. Walter SD. The analysis of regional patterns in health data. I. Distributional considerations. Am J Epidemiol. 1992; 136: 730-741.
23. Walter SD. The analysis of regional patterns in health data: II. The power to detect environmental effects. Am J Epidemiol. 1992; 136: 742-759.
24. Walter SD. Assessing spatial patterns in disease rates. Stat Med. 1993; 12: 1885-1894.
25. Whittemore AS, Friend N, Brown BW, Holly EA. A test to detect clusters of disease. Biometrika. 1987; 74: 631-635.
26. Cliff AC, Ord J. Spatial processes: models and applications. London: Pion Limited.1980.
27. Boots BN, Getis A. Point pattern analysis. CA Newbury Park: Sage Publications. 1998.
28. Glavanakov S, White DJ, Caraco T, Lapenis A, Robinson GR, Szymanski BK, Maniatty WA. Lyme disease in New York State: spatial pattern at a regional scale. Am J Trop Med Hyg. 2001; 65: 538-545.
29. Perez AM, Ward MP, Torres P, Ritacco V. Use of spatial statistics and monitoring data to identify clustering of bovine tuberculosis in Argentina. Prev Vet Med. 2002; 56: 63-74.
30. Bellec S, Hémon D, Rudant J, Goubin A, Clavel J. Spatial and space-time clustering of childhood acute leukaemia in France from 1990 to 2000: a nationwide study. Br J Cancer. 2006; 94: 763-770.
31. Nødtvedt A, Guitian J, Egenvall A, Emanuelson U, Pfeiffer DU. The spatial distribution of atopic dermatitis cases in a population of insured Swedish dogs. Prev Vet Med. 2007; 78: 210-222.
32. Kitron U, Kazmierczak JJ. Spatial analysis of the distribution of Lyme disease in Wisconsin. Am J Epidemiol. 1997; 145: 558-566.
33. Odoi A, Martin SW, Michel P, Holt J, Middleton D, Wilson J. Geographical and temporal distribution of human giardiasis in Ontario, Canada. Int J Health Geogr. 2003; 2: 5.
34. Borroto RJ, Martinez-Piedra R. Geographical patterns of cholera in Mexico, 1991-1996. Int J Epidemiol. 2000; 29: 764-772.
35. Isaaks EH, Srivastava RM. An Introduction to Applied Geostatistics. New York: Oxford University Press.1989.
36. Liebhold AM, Rossi RE, Kemp WP. Geostatistics and geographic information systems in applied insect ecology. Annu. Rev. Entomol. 1993; 38: 303-327.
37. Rossi RE, Mulla DJ, Journel AG, Franz EH. Geostatistical tools for modeling and interpreting ecological spatial dependence. Ecol. Monogr. 1992; 62: 277-314.
38. Weisz R, Fleischer S, Smilowitz Z. Site-specific integrated pest management for high value crops: sample units for map generation using the Colorado potato beetle (Coleoptera: Chrysomelidae) as a model system. J Econ Entomol. 1995; 88: 1069-1080.
39. Upton GJG, B Fingleton. Spatial Data Analysis by Example. Volume 1: Point Pattern and Quantitative Data. Chichester: John Wiley & Sons. 1985.
40. Anselin L. Local Indicators of Spatial Association: LISA. Geogr.Anal. 1995; 27: 93-115.
41. Getis A, Ord JK. The analysis of spatial association by use of distance statistics. Geogrr. Anal. 1992; 24: 189-206.
42. Ripley B. The second-order analysis of stationary point processes. J. Appl. Prob. 1976; 13: 255-266.
43. Ripley B. Spatial Statistics. Chichester: John Wiley & Sons. 1981.
44. Ripley B. Statistical Inference for Spatial Processes. Cambridge: Cambridge University Press. 1988.
45. Turnbull BW, Iwano EJ, Burnett WS, Howe HL, Clark LC. Monitoring for clusters of disease: application to leukemia incidence in upstate New York. Am J Epidemiol. 1990; 132: S136-143.
46. Knox EG. The detection of space-time interaction. J. R. Stat. Soc. C. 1964; 13: 25-20.
47. Knox EG. Detection of clusters. Elliott P, editor. In: Methodologies of Enquiry into Disease Clustering. Small Area Health Statistics Unit, London. 1989; 17-22.

48. Mantel N. The detection of disease clustering and a generalized regression approach. Cancer Res. 1967; 27: 209-220.

49. Ederer F, Myers MH, Mantel N. A statistical problem in space and time: Do leukaemia cases come in clusters? Biometrics.1964; 20: 626-638.

50. Barton DE, David FN, Merrington M. A criterion for testing contagion in time and space. Ann. Hum. Genet. 1965; 29: 97-103.

51. Diggle PJ, Chetwynd AG, Häggkvist R, Morris SE. Second-order analysis of space-time clustering. Stat Methods Med Res. 1995; 4: 124-136.

52. Kulldorff M. SaTScan users guide for version 6.0. 2006.

53. Hjalmars U, Kulldorff M, Gustafsson G, Nagarwalla N. Childhood leukaemia in Sweden: using GIS and a spatial scan statistic for cluster detection. Stat Med. 1996; 15: 707-715.

54. Michelozzi P, Capon A, Kirchmayer U, Forastiere F, Biggeri A, Barca A, Perucci CA. Adult and childhood leukemia near a high-power radio station in Rome, Italy. Am J Epidemiol. 2002; 155: 1096-1103.

55. Viel JF, Arveux P, Baverel J, Cahn JY. Soft-tissue sarcoma and non-Hodgkin's lymphoma clusters around a municipal solid waste incinerator with high dioxin emission levels. Am J Epidemiol. 2000; 152: 13-19.

56. Sheehan TJ, DeChello LM. A space-time analysis of the proportion of late stage breast cancer in Massachusetts, 1988 to 1997. Int J Health Geogr. 2005; 4: 15.

57. Odoi A, Martin SW, Michel P, Middleton D, Holt J, Wilson J. Investigation of clusters of giardiasis using GIS and a spatial scan statistic. Int J Health Geogr. 2004; 3: 11.

58. Tiwari N, Adhikari CM, Tewari A, Kandpal V. Investigation of geo-spatial hotspots for the occurrence of tuberculosis in Almora district, India, using GIS and spatial scan statistic. Int J Health Geogr. 2006; 5: 33.

59. Green C, Hoppa RD, Young TK, Blanchard JF. Geographic analysis of diabetes prevalence in an urban area. Soc Sci Med. 2003; 57: 551-560.

60. Cousens S, Smith PG, Ward H, Everington D, Knight RS, Zeidler M, Stewart G. Geographical distribution of variant Creutzfeldt-Jakob disease in Great Britain, 1994-2000. Lancet. 2001; 357: 1002-1007.

61. Chaput EK, Meek JI, Heimer R. Spatial analysis of human granulocytic ehrlichiosis near Lyme, Connecticut. Emerg Infect Dis. 2002; 8: 943-948.

62. Sabel CE, Boyle PJ, Löytönen M, Gatrell AC, Jokelainen M, Flowerdew R, Maasilta P. Spatial clustering of amyotrophic lateral sclerosis in Finland at place of birth and place of death. Am J Epidemiol. 2003; 157: 898-905.

63. Takahashi K, Yokoyama T, Tango T. FleXScan. Software for the flexible spatial scan statistic. Japan: National Institute of Public Health. 2004.

64. Kim AY, Wakefield J. R Data and Methods for Spatial Epidemiology: The SpatialEpi Package. 2010.

65. Lawson AB, Biggeri A, Bohning, Lesaffre E, Viel J-F, Bertollini R. Introduction to spatial models in ecological analysis Disease. Lawson AB, Biggeri A, Bohning, Lesaffre E, Viel J-F, Bertollini R, editors. In: Disease Mapping and Risk Assessment for Public Health. Chichester: John Wiley & Sons. 1999; 181-191.

66. Lawson AB. Statistical Methods in Spatial Epidemiology. Chichester: John Wiley & Sons. 2001.

67. Elliott P, Wakefield JC, Best NG, Briggs DJ. Spatial Epidemiology: Methods and Applications. Elliott P, Wakefield J, Best NG, Briggs DJ, editors. In: Spatial Epidemiology: Methods and Applications. Oxford: Oxford University Press. 2000; 1-29.

68. Nelder JA, Wedderburn RWM. Generalized linear models. J. R. Stat. Soc. A. 1972; 135: 370-384.

69. Mc Cullagh P, Nelder JA. Generalized Linear Models. London: Chapman and Hall. 1989.

70. Fahrmeir L, Tutz G. Multivariate Statistical Modeling Based on Generalized Linear Models. New York: Springer. 2001.

71. Mc Culloch CE, Searle SR. Generalized, Linear, and Mixed Models. New York: John Wiley & Sons. 2001.

72. Draper NR, Smith H. Applied Regression Analysis. 3rdedn.New York. John Wiley & Sons. 1998.

73. Cressie N. Statistics for Spatial Data.New York: John Wiley & Sons. 1993.

74. Haining RP. Spatial Data Analysis in the Social and Environmental Sciences. Cambridge: Cambridge University Press. 1990.

75. Anselin L, Ibnu S, Youngihn K. GeoDa: An Introduction to Spatial Data Analysis. Geogr. Analy. 2006; 38: 5-22.

76. LeSage J, Pace RK. Introduction to Spatial Econometrics. London: CRC Press/Taylor & Francis Group. 1999.

77. Hastie T, Tibshirani R. Generalized Additive Models. London: Chapman and Hall. 1990.

78. Fan J, Gijbels I. Local Polynomial Modeling and Its Applications. London: Chapman and Hall. 1996.

79. Friedman JH, Silverman BL. Flexible Parsimonious Smoothing and Additive Modeling (with discussion). Technometrics. 1989; 31: 3-39.

80. Friedman JH. Multivariate adaptive regression splines (with discussion). Ann. Stat. 1991; 19:1-141.

81. Stone CJ, Hansen MH, Kooperberg C, Truong YK. Polynomial splines and their tensor products in extended linear modeling (with discussion). Ann. Stat. 1997; 25: 1371-1470.

82. Eilers PHC, Marx BD. Flexible smoothing with B-splines and penalties. Stat. Sci. 1996; 11: 89-121.

83. Marx B, Eilers PHC. Direct generalized additive modeling with penalized likelihood. Comput.Stat. Data Anal. 1998; 28: 193-209.

84. Fahrmeir L, Kneib T, Lang S. Penalized structured additive regression for space-time data: a Bayesian perspective. Stat. Sin. 2004; 14: 731-761.

85. Brezger A. Bayesian P-Splines in Structured Additive Regression Models [Dissertation] Universität München. 2004.

86. Kneib T. Mixed model based inference in structured additive regression [dissertation]. Universität München. 2005.

87. Gelfand AE, Smith AFM. Sampling-based approaches to calculating marginal densities. J. Am. Stat. Ass. 1990; 85: 398-409.

88. Besag J. Spatial interaction and the statistical analysis of lattice systems (with discussion). J. R. Stat. Soc. B. 1974; 36: 192-225.

89. Besag J. Statistical analysis of non-lattice data. J. R. Stat. Soc. D. 1975; 24: 179-195.

90. Fotheringham AS, Brunsdon C, Charlton ME. Geographically weighted regression: The analysis of spatially varying relationships. Chichester: John Wiley & Sons. 2002.

91. Fahrmeir L, Lang S. Bayesian semiparametric regression analysis of multicategorical time-space data. Ann. Inst. Stat. Math. 2001; 53: 11-30.

92. Fahrmeir L, Lang S. Bayesian inference for generalized additive mixed models based on Markov random field priors. J. R. Stat. Soc. C. 2001; 50: 201-220.

93. Lang S, Brezger A. Bayesian P-splines. J. Comput. Graph. Stat. 2004; 13: 183-212

94. Brezger A, Lang S: Generalized additive regression based on Bayesian P-splines [SFB 386 Discussion paper 321] Department of Statistics, Universitat Munchen. 2003

95. Wahba G. Improper Prior, Spline smoothing and the problem of guarding against model errors in regression. J. R. Stat. Soc. B. 1978; 44: 364-372.

96. Manton KG, Woodbury MA, Stallard E. A variance components approach to categorical data models with heterogeneous cell populations: analysis of spatial gradients in lung cancer mortality rates in North Carolina counties. Biometrics. 1981; 37: 259-269.

97. Tsutakawa RK. Mixed model for analyzing geographic variability in mortality rates. J Am Stat Assoc. 1988; 83: 37-42.

98. Besag J, York Y, Mollie A. Bayesian image-restoration, with two applications in spatial statistics. Ann. Inst. Stat. Math. 1991; 43: 1-20.

99. Clayton D, Kaldor J. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics. 1987; 43: 671-681.

100. Clayton D, Bernardinelli L. Bayesian methods for mapping disease risk. Elliott P, Cuzick J, English D, Stern R, editors. In: Geographical and Environmental Epidemiology: Methods for Small-Area Studies. Oxford: Oxford University Press. 1992; 205-220.

101. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit (with discussion). J. R. Stat. Soc. B. 2002; 64: 583-640.

102. Lawson AB, Browne, WJ, Vidal Rodeiro CL. Disease Mapping with WinBUGS and MLwiN. Chichester: Wiley and Sons. 2003.

103. Lawson AB. Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology. Boca Raton: Chapman and Hall/CRC. 2008.

104. Smith M, Kohn R. Nonparametric regression using Bayesian variable selection. J. Econom. 1996; 75: 317-343.

105. Smith M, Kohn R. A Bayesian Approach to Nonparametric Bivariate Regression. J. Am. Stat. Ass. 1997; 92: 1522-1535.

106. Denison DGT, Mallick BK, Smith AFM. Automatic Bayesian curve fitting. J. R. Stat. Soc. B. 1998; 60: 333-350.

107. Biller C. Adaptive Bayesian regression splines in semiparametric generalized linear models. J. Comput. Graph. Stat. 2000; 9: 122-140.

108. Biller C, Fahrmeir L. Bayesian varying-coefficient models using adaptive regression splines. Stat. Model. 2001; 1:195-211.

109. DiMatteo I, Genovese CR, Kass RE. Bayesian curve-fitting with free-knot splines. Biometrika. 2001; 88: 1055-1071.

110. Hansen MH, Kooperberg C. Spline adaptation in extended linear models (with discussion and rejoinder by the authors). Stat. Sci. 2002; 17: 2-51.

111. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. Stats.Compu. 2000; 10: 325-337.

112. Belitz C, Brezger A, Kneib, T, Lang S, Umlauf N: BayesX - Software for Bayesian inference in structured additive regression models. Version 0.9.

Austin Biom and Biostat - Volume 1 Issue 2 - 2014
Submit your Manuscript | www.austinpublishinggroup.com

Citation: Osei FB. Current Statistical Methods for Spatial Epidemiology: A Review. Austin Biom and Biostat. 2014;1(2): 7.