Spatial Autocorrelation and Association Measures

J. Negreiros
Universidade Lusófona, Portugal

M. Painho
Instituto Superior de Estatística e Gestão de Informação - Universidade Nova de Lisboa, Portugal

I. Lopes
Universidade Lusófona, Portugal

A. C. Costa
Instituto Superior de Estatística e Gestão de Informação - Universidade Nova de Lisboa, Portugal

GIS AND SPATIAL ANALYSIS: AN OVERVIEW

Several classical statements relating to the definition of GIS can be found in specialized literature such as the *GIS International Journal*, expressing the idea that spatial analysis can somehow be useful. GIS is successful not only because it integrates data, but it also enables us to share data in different departments or segments of our organizations. I like this notion of putting the world's pieces back together again (ArcNews, 2000). "GIS is simultaneously the telescope, the microscope, the computer and the Xerox machine of regional analysis and the synthesis of spatial data" (Abler, 1988). "GIS is a system of hardware, software and liveware implemented with the aim of storing, processing, visualizing and analyzing data of a spatial nature. Other definitions are also possible" (Painho, 1999). "GIS is a tool for revealing what is otherwise invisible in geographical information" (Longley, Goodchild, Maguire, & Rhind, 2001). Certainly, GIS is not a graphic database.

According to Unwin (1997) and Bailey (1994), "the spatial analysis concept has nothing to do with the general ability to describe spatial data, generally incorporated in all commercial GIS but it presents the challenge, given a spatial pattern, to explore it with an appropriate model and represent it with a graphical display." Certainly, to describe is not to explore. As Murteira (1993) confirms, "exploring is a detective work, a search for clues and evidence, while describing is a job of judgment, a job of analyzing and evaluating clues."

Rossiter (1999) includes spatial flow modeling and deterministic processes like groundwater movement and environmental quality management based on economic criteria such as land use and transportation. It seems that geographical analysis comprises GIS (an applied computer-science view), spatial statistics including uncertainty issues (spatial autocorrelation, spatial autoregression, Kriging, stochastic simulation, morphologic geostatistics, and space-time processes), classical spatial statistics, remote detection, and deterministic spatial analysis such as optimization routing, B-Splines, overlay, buffering and DEM (Digital Elevation Model) operations (cartographic modeling).

Recently, Wikle and Cressie (1999) reported new advances in spatiotemporal prediction for large datasets using Kalman filter modeling to reduce the time dimension in an empirical method of moments implementation, while Hopkins et al. (1999) mentioned Kriging use for (x,y,t) ozone concentration estimation with GMS®. It is therefore implicit that spatial analysis is a GIS component to support decision making for solving problems with a spatial component.

SPATIAL AUTOCORRELATION

The role of this spatial component holds two major implications for the way statistical analysis should be carried out. Location leads to spatial dependence (correlation or variation that each neighbor holds in relation to a particular point) and spatial heterogeneity (clustering, concentration, or proportion of neighborhood average in relation to a specific point) established by Tobler's First Law of Geography: "everything is related to everything but closer things tend to be more related". Since regional differentiation respects the intrinsic uniqueness of each location, spatial autocorrelation can be viewed, hence, as a map pattern descriptor.

Classical statistics offer a wide range of inferential methods based on restricted assumptions. The 95% confidence level intervals for the true unknown population mean and standard deviation are only valid if samples are independent and homocedastic with uncorrelated error. Another classical approach to test if there is an association between two variables where each variable consists of a number of classes is the chi-square (χ^2) contingency test, which should be used for sufficiently large independent random samples. Classical statistical tests for the difference between the means of two sub-populations are based on independent random samples and normal distribution assumptions. Furthermore, linear and multiple regression models, including the global F and individual t-Student tests, assume a linearity relationship between the random variables and the independency, homocedasticity, and normality of the errors. "If these assumptions are verified then the ordinary least squares estimator is the best unbiased estimator" (Griffiths, Hill, & Judge, 1993, p. 301).

Yet, spatial data do not follow statistician's rules, and spatial patterns must be taken into account. As Griffith and Layne (1999) point out, "spatial autocorrelation is quite often positive with a correlation range between 0.3 and 0.5." Spatial ANOVA of SpaceStat® regards spatial differentiation as a regression model whose spatial W matrix can be included, and explanatory variables are converted to categorical ones such as North/South subregions. "This assumption of indicator variables in the form of a dummy variable can measure the difference between the sub-region mean and the overall one" (Anselin, 1992). A positive and high t-Student indicates a strong discrepancy among subregions. Furthermore, "if spatial residual dependency exists then biased estimates can be assumed and a spatial lag autocorrelation regression can be applied" (Anselin, 1992).

"Autocorrelation damages the ability to perform standard statistical hypothesis tests because the confidence interval estimated by the classical Pearson product moment is narrow inducing, thus, biased conclusions" (Legendre, 1994; Nass & Garfinkle, 1992). "The estimator standard errors are not minimized and regression coefficients from least squares are unbiased while their variances are underestimated" (Clark & Hosking, 1986; Ebdon, 1998). This occurs because new observations under the lack of independence do not each bring one full degree of freedom since the observer holds some prior knowledge at new locations reflecting information loss. As spatial autocorrelation approaches 1, the effective degrees of freedom approach 0. As spatial autocorrelation approaches 0, the effective degrees of freedom tend to the total sampling number. As spatial autocorrelation approaches -1, the effective degrees of freedom increase beyond the number of observations.

Fewer samples are therefore needed as positive spatial autocorrelation increases, a possible reduction solution for the huge spatial data computation issue. Under this viewpoint, there is a better than equal chance of predicting neighboring values if this information is already available. As spatial autocorrelation also decreases, error prediction increases. Spatial autocorrelation is the spine of spatial interpolation. If someone records downtown air pollution in Madrid, Spain, monitoring devices will report the same levels adding almost no new information to the previous samples. Information redundancy is the logical consequence with autocorrelated georeferenced data because samples collected for two juxtaposed points in space tend to represent essentially the same information. Thus, spatial autocorrelation may be defined as a measure of the true but masked information content in spatial data.

"If spatial autocorrelation exists, the standard error of the linear estimators will not be minimized, the dependent variable estimation will be inefficient, the confidence intervals will be larger, R^2, t and F tests will no longer be valid and the residuals may be highly correlated" (Dutilleul, 1993). The independent zero spatial autocorrelation assumption conflicts with Tobler's Law: if strong residual autocorrelation exists, then ordinary least squares (OLS) inference breaks down because near things hold the same weight as distant ones.

With independent data, the two-sided 95% confidence interval for the mean μ is [$\bar{x}$ $-1.96\sigma/\sqrt{n}$, $\bar{x}$ $+1.96\sigma/\sqrt{n}$], where $\bar{x}$ is the sample mean and σ the population standard deviation, but with positive *n* correlated samples, Cressie (1993) proved that the previous interval equals [$\bar{x}$ $-2.485\sigma/\sqrt{n}$, $\bar{x}$ $+2.485\sigma/\sqrt{n}$] which leads to a wider confidence interval. Moreover, the estimated variance of the mean with a population of *n* samples is not σ^2/n but σ^2/n' where $n'=n/[1+2\{\rho/(1-\rho)\}\{1-(1/n)\}-2\{\rho/(1-\rho)\}^2(1-\rho^{n-1})/n]$ and ρ is the Pearson's correlation

Related Content