Autocorrelation, Autoregression and Kriging: Appraisal for The Spatial Interpolation Issue

Negreiros, João
Universidade Nova de Lisboa, ISEGI
Campus de Campolide
1070-312 Lisboa, Portugal
E-mail: c8057@isegi.unl.pt

Costa, Ana Cristina
Universidade Nova de Lisboa, ISEGI
Campus de Campolide
1070-312 Lisboa, Portugal
E-mail: ccosta@isegi.unl.pt

Painho, Marco
Universidade Nova de Lisboa, ISEGI
Campus de Campolide
1070-312 Lisboa, Portugal
E-mail: painho@isegi.unl.pt

Santos, José
Universidade Nova de Lisboa, ISEGI
Campus de Campolide
1070-312 Lisboa, Portugal
E-mail: jsantos@isegi.unl.pt

1. Introduction

According to Rogerson and Fotheringham (1994), there is an increased demand for systems that do more than display and organize data. The set of potential applications for spatial analysis is enormous, for example, accident patterns, victim profiles within a residential population and spread rates for pollution levels. Techniques are needed to let spatial data speak for themselves (Griffith and Layne, 1999). Thus, spatial statistics must hold a specific spatial framework to apply quantitative and statistical methods for a better understanding of spatial relationships.

If the finding of spatial structures is fundamental then spatial autocorrelation, spatial interpolation and spatial autoregressive models are the three major spatial methods that fulfill this constraint. Although these fields have been developed autonomously, this short-paper tries to review a particular pattern that underlies these three concepts: the spatial interpolation subject. In short, this involves the estimation of missing geo-referenced data when undertaking spatial autoregression and Kriging. However, spatial autocorrelation can also be exploited for spatial prediction purposes such as patterns and outliers. Spatial autocorrelation measures can also indicate if there is an equally likely chance of predicting neighboring values while the degree of redundancy indicates how much information is free to vary. As expected, certain topics emerge from these concepts such as the dependence of the Ordinary Kriging (OK) variance solely on geometry, the weight matrix codification of autoregressive approaches, the need of heavy spatial analysis computation, the non-standardization possibility of observation

distribution or the impact of sampling approaches. However, these last questions will not be covered here.

2. Kriging, Autocorrelation and Autoregression

The word Kriging is synonymous with the optimal prediction of unknown values from observed data at known locations (Auñón and Gómez-Hernández, 2000). After the variogram has been defined, the algebraic relationship between values at individual distances including clustering and direction is used to estimate Kriging weights. Mostly, four factors are taken into account in assigning weights: A) Closeness to the location being estimated; B) Redundancy between data values; C) Anisotropic continuity; D) Magnitude of continuity. Kriging is BLUP (best linear unbiased predictor) whether or not data are normally distributed. It is linear since estimations are weighted linear combinations of the available data. It is unbiased because the error mean is zero (no over- or under-estimates). It is best since its goal is to minimize error variance. Conceptually, the variogram curve has minimal error variance at any known data location (it is zero but if and only if there is no measurement error) and maximum value at some specified range away from that point. As expected, the predictable maximum error Kriging deviation occurs in areas where there are no points.

Spatial autocorrelation can be interpreted as a descriptive index, measuring aspects of the way things are distributed in space (clustering) but, at the same time, it can be seen as a casual process measuring the degree of influence exerted by something over its neighbors (correlation) Goodchild (1986). This happens because the role of location 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. Since regional differentiation respects the intrinsic uniqueness of each location, spatial autocorrelation can be viewed, hence, as a map pattern descriptor.

Autoregressive models (AR, CAR and SAR) expand the standard linear model regression with an additional term that accounts for patterns that are not predicted by local variables but are instead related to the residual of the neighbour locations. In the AR model, for instance, the autoregressive term is independent of the environmental predictors and is appropriately applied when correlations result from endogenous population processes that are unrelated to environmental conditions. One live example of this approach is given by Lee (2005) concerning the analysis of a watershed (the water quality effects of point and non-point pollution sources) in the Neuse River Basin, NC. It is based on the hypothesis that the spatially weighted sum of water quality in neighboring stations affects the water quality of each monitoring station (indirect effect), as do the standard explanatory variables of pollution sources (direct effects). The reaction function is in the form of a spatial AR model: $y = \rho Wy + X1\beta1 + X2\beta2 + \varepsilon$, where y represents the water quality indicator vector, W is a weight matrix, X1 is the matrix of pollution sources, X2 is the matrix of land cover and stream characteristics, ρ is a spatial autocorrelation parameter, and ε is a random error vector. The model specifies that the spatially weighted sum of neighbor water qualities affects the nutrient level of each downstream monitoring unit, as do the general covariates of pollution sources and heterogeneous characteristics in each geographical unit. It eventually states that the realization of y at hydrologic unit i is a function of its realization of X at i, plus realization of y at hydrologic unit j, plus an error. According to Lee (2005), this AR model type is useful for explaining the

phenomena in a watershed because it allows observing the distinct spatial externality of spillover effects.

3. Discussion

The threat of spatial autocorrelation and statistical heterogeneity hold the ability to compare two regions and to characterize texture differences. Quite often, distance location pairs are less similar (competitive spatial processes) than closer ones (cooperative spatial processes). As a consequence, indices of spatial autocorrelation calculated globally and locally are valuable for descriptive purposes because they provide a measure of how similar objects are to their spatial neighbors. This spatial dependence impact is also crucial for Kriging.

However, it is the notion that variation is spatially autocorrelated in some predictable fashion that creates a fundamental conflict between the goals of classical statistics and its assumptions because classical estimators tend to be inconsistent as the degree of spatial dependence increases (either positive or negative). This may come as a surprise to ecologists and geographers who have been trained in the belief that Mother Nature follows the assumptions of classical statistics (Levine, 1997).

The spatial autoregression plus spatial autocorrelation provides another linkage with Kriging: the spatial interpolation subject. This means that the estimation of absent spatial data can be undertake by spatial autoregressive methods, Kriging and autocorrelation measures. For instance, a high degree of spatial autocorrelation suggests an equally likely chance of predicting neighboring values. On the contrary, a low value reveals a low level of spatial data redundancy (Griffith and Layne, 1999). Thus, several relationships among these three concepts emerge:

- Spatial autocorrelation is the required condition of Kriging and autoregression models. In fact, spatial autocorrelation is a physical reality and is a necessity for interpolating nearby values.
- Spatial autocorrelation seeks spatial identification report while spatial regression and Kriging seeks spatial prediction.
- The variance-covariance matrix is included within spatial regression and Kriging (Griffith and Layne, 1999).
- Once a variogram is computed, Kriging can be used for spatial interpolation.
- With spatial regression models, spatial interpolation can be regarded as an interactive re-estimation solution fashioned with updated variable imputations based on R^2 and IC decision parameters in Maximum Likelihood, Ordinary Least Squares and bootstrap procedures (Griffith and Layne, 1999). According to these authors, incorporating spatial autocorrelation results in a 5-10% improvement in the statistical description for a given georeferenced dataset.
- Kriging is primarily applied with continuous regions while spatial autoregression involves aggregations of phenomena into discrete regions such as ward units (Griffith and Layne, 1999).
- Spatial autoregressive methods assume that spatial interpolation follow an underlying trend plus random residuals. However, Kriging can presume two views: A) If Universal Kriging is assumed then spatial interpolation is interrelated with a background trend; B) If Ordinary Kriging is chosen then spatial interpolation is consistent with the samples global average plus random residuals.
- Since Kriging is an exact spatial interpolator (if no samples measurement error is applied), autoregressive residuals can be used for a reasonable reality approximation test.

The chance to verify existing patterns among residuals provides a key information source on possible assumption violations, variable transformations, outliers, trends surface and inappropriate formulations for raw data (Clark and Hosking, 1986).

The computational burden underling a number of those approaches is a challenging problem, particularly with very large spatial-temporal datasets. For example, spatial interpolation simulations in earth sciences typically require huge models with millions of cells. One way to overcome this problem is the brute-force of parallel computers, as recently proposed by Vargas *et al.* (2006).

REFERENCES

Auñón, J. and Gómez-Hernández, J.J. (2000), Dual Kriging with Local Neighborhoods: Application to the Representation of Surfaces. *Mathematical Geology*, 32(1), 69-85.

Clark, W., Hosking, P. (1986), *Statistical Methods for Geographers*. New York: John Wiley & Sons, 528 pp.

Goodchild, M. (1986), *Spatial Autocorrelation*. CATMOG 47, GeoBooks, Norwich.

Griffith, D.A. and Layne, L.J. (1999), *A Casebook for Spatial Statistical Data Analysis: A Compilation of Analyses of Different Thematic Data Sets*. Oxford University Press, 506 pp.

Lee, J.-H. (2005), *Spatial Econometric Analysis of a Watershed Utilizing Geographic Information Systems: Water Quality Effects of Point and Non-Point Pollution Sources in the Neuse River Basin, NC.* PhD Dissertation, NCSU - North Carolina State University.

Levine, N. (1997), Spatial statistics and GIS; software tools to quantify spatial patterns. *Journal of the American Planning Association*, 62(3), 381-392.

Rogerson, P.A. and Fotheringham, A.S. (1994), "GIS and Spatial Analysis: Introduction and Overview". In: Fotheringham, S. and Rogerson, P. (eds.), *Spatial analysis and GIS*. Bristol:Taylor & Francis Inc., 1-10.

Vargas, H. S., Caetano, H. and Mata-Lima, H. (2006). “A new Parallelization approach for Sequential Simulation – Environmental Application”. Paper presented at *geoENV 2006 – 6th International Conference on Geostatistics for Environmental Applications*, Rhodes, Greece.

ABSTRACT

Conventional statistical models must be reformulated to properly account for spatial correlation and spatial heterogeneity within georeferenced data. For instance, if the autoregressive residuals reveal a medium-strong spatial autocorrelation then any missing variable within the initial regression model can be significant. In emphasizing classical statistical assumptions about spatial data, four major problems emerge. I) *Non-Gaussian distribution of spatial variables (trend removal and Box-Cox power transformations are the corrective actions before OK, SK and UK can be applied).* II) *Correlated random error with non-zero mean and non-Normal distribution (a low probability of the Kiefer-Salmon asymptotic regression indicates a rejection of the null hypothesis of Gaussian error).* III) *Non-homoscedasticity of error distribution (it is easier to handle nonstationary time series with changing means than nonstationary time series with changing variances).* IV) *Non-multicolinearity leads to a larger estimated variance of the regression coefficients (the combination of a high R^2 with a very low t statistic is often a good indicator that something is wrong in terms of colinearity.*

It is also important to stress the heavy computation issue concerning these three approaches. One way to overcome the mathematical design that underlies them is the brute-force of parallel computers, particularly with very large spatial-temporal datasets.