J Geograph Syst (2000) 2:141–156

Journal of Geographical Systems


# A linear regression solution to the spatial autocorrelation problem

**Daniel A. Griffith**

Department of Geography and Interdisciplinary Statistics Program, Syracuse University, Syracuse, NY 13244-1020, USA (e-mail: griffith@maxwell.syr.edu) and ASA/USDA-NASS Fellow, National Agricultural Statistics Service, Fairfax, VA, USA

Received: 18 February 1999 / Accepted: 17 September 1999

**Abstract.** The Moran Coefficient spatial autocorrelation index can be decomposed into orthogonal map pattern components. This decomposition relates it directly to standard linear regression, in which corresponding eigenvectors can be used as predictors. This paper reports comparative results between these linear regressions and their auto-Gaussian counterparts for the following georeferenced data sets: Columbus (Ohio) crime, Ottawa-Hull median family income, Toronto population density, southwest Ohio unemployment, Syracuse pediatric lead poisoning, and Glasgow standard mortality rates, and a small remotely sensed image of the High Peak district. This methodology is extended to auto-logistic and auto-Poisson situations, with selected data analyses including percentage of urban population across Puerto Rico, and the frequency of SIDs cases across North Carolina. These data analytic results suggest that this approach to georeferenced data analysis offers considerable promise.

**Key words:** Eigenfunction, spatial autocorrelation, spatial autoregression, geographic weights matrix, georeferenced data

**JEL classification:** C49, C13, R15

## 1 Motivation

As with time series modelling, spatial statistical modelling involves nonlinear model specifications and estimation in order to account for auto- or self-correlation. One source of this nonlinearity is the complicated normalizing factor that appears in spatial statistical likelihood function estimation equations. This factor ensures that when moving from a spatially autocorrelated to a spatial unautocorrelated mathematical attribute space, probabilities sum/integrate to 1. The expression for this normalizing constant in the auto-Gaussian case – the prefix *auto-* signifying the acknowledgment of self-correlation – is cumbersome, relates directly to eigenfunctions (Ord 1975), and can be accurately approximated (Griffith and Sone 1995) or sometimes efficiently handled through the use of special matrix operations (Barry and Pace

1997). Unfortunately the same is not true for the auto-logistic or auto-Poisson specifications (Cressie 1991, pp. 424–440).

Normalizing factors for all spatial autoregressive models are problematic, in that they are complex and often have no closed form solutions. One solution to this problem has been to avoid computing an analytical probability density altogether; rather, a researcher could simply draw a very large random sample from a target distribution using Markov Chain Monte Carlo (MCMC) procedures (e.g., Gibbs sampling). If the sample is large enough, the distribution can be investigated empirically (e.g., the probability density could be approximated using kernel density estimators or histograms). Of note is that the normalizing constant problem has plagued Bayesian statistics until recently, with this MCMC approach supplying a means to circumvent it. LeSage (1996) outlines how this computer intensive approach can be implemented; Conlon and Waller (1999) outline implementation specifically for the spatial statistical conditional autoregressive (CAR) prior. The primary goal of this paper is to evaluate another possible solution to this problem by making selected empirical comparisons. More specifically, it seeks to circumvent computational difficulties associated with the complicating normalizing factors, reducing the numerical intensity of spatial statistical procedures and making results more directly comparable with those of more traditional statistical methods. Freeing scientists from having to contend with such numerical difficulties is a considerable advantage. The methodology presented and assessed here is especially of importance to geographers and regional scientists who deal with percentage, counts, and/or large georeferenced data sets.

## 2 Background

The most common interpretation of spatial autocorrelation is in terms of trends, gradients, or patterns across a map. Unlike conventional correlation coefficients, however, the Moran Coefficient (MC) is not restricted to the range $[-1, 1]$; rather, its range is dictated by what essentially are the extreme eigenvalues of the geographic connectivity matrix $\mathbf{C}$ (the matrix depicting the geometric arrangement of areal units). $\mathbf{C}$ is constructed from $n^2$ 0/1 binary values, where $c_{ij} = 1$ if row areal unit $i$ and column areal unit $j$ share a common boundary and $c_{ij} = 0$ otherwise, and consequently can go slightly beyond these two usual endpoint values (de Jong et al. 1984). In fact, the accompanying eigenvectors represent a kaleidoscope of orthogonal map patterns of possible spatial autocorrelation (Griffith 1996). Tiefelsdorf and Boots (1995, 1996) uncover part of this relationship between the eigenfunctions and MC values. It is this set of eigenvectors that offers a possible solution to the tractability of the normalizing constant problem, in a context similar to REML (restricted maximum likelihood) methodology. Additionally, a theorem by Rao (1967) confirms the equivalency between the OLS and spatial autoregressive solutions here, since the spatial autoregressive covariance matrix contained in the log-likelihood auto-Gaussian specification [see expression (3.1)] can be rewritten in the form he specifies, namely

$$\left[(\mathbf{I} - \mathbf{1}\mathbf{1}'/n) + \sum_{i=1}^{n} \mathbf{E}\rho^k \mathbf{\Lambda}^k \mathbf{E}'\right]\sigma^2, \tag{2.1}$$

where $\mathbf{I}$ is an $n$-by-$n$ identity matrix, $\mathbf{1}$ is an $n$-by-1 vector of ones, $\mathbf{E}$ is an $n$-by-$n$ matrix of orthogonal eigenvectors, $\rho$ is the spatial autocorrelation parameter, $\mathbf{\Lambda}$ is an $n$-by-$n$ diagonal matrix of eigenvalues (corresponding to the eigenvectors in matrix $\mathbf{E}$), $\mathbf{C} = \mathbf{E\Lambda E}'$, $'$ denotes the operation of matrix transpose, and $\sigma^2$ denotes the standard constant variance parameter.

## 3 Problem statement

The presence of non-zero spatial autocorrelation introduces a number of complications into the statistical analysis of georeferenced data. Foremost are the necessary modifications of probability density/mass functions when moving from a spatially autocorrelated to a spatially unautocorrelated mathematical space, one of which is the normalizing constant. Second, because model specifications are nonlinear in nature, nonlinear estimation techniques must be used. The third concerns attaining a better and deeper understanding of the spatial autocorrelation phenomenon. These problems give rise to the following research questions:

How can the normalizing constant complication be avoided?
How can spatial statistical models be equated with conventional statistical models?
What does the spatial autocorrelation term in a spatial statistical model mean?

The general problem addressed here seeks to supply at least some answers to each of these three questions.

### *3.1 The normalizing constant problem*

Most regression analyses that take into account latent spatial dependency are based upon the assumption of normally distributed error terms, and involve maximum likelihood estimation (MLE) for which the log-likelihood function for variable $Y$ is specified as

$$\text{constant} - \frac{n}{2} LN(\sigma^2) + \frac{1}{2} LN[\det(\mathbf{V})] - (\mathbf{Y} - \mathbf{X\beta})'\mathbf{V}(\mathbf{Y} - \mathbf{X\beta})/(2\sigma^2), \tag{3.1}$$

where $\mathbf{Y}$ is an n-by-1 vector of map values for variable $Y$, $\det(\mathbf{V})$ denotes the operation of matrix determinant being performed on matrix $\mathbf{V}$, $LN$ denotes the natural logarithm, $\mathbf{X\beta}$ denotes the standard nonconstant mean, $\boldsymbol{\beta}$ is a $(p+1)$-by-1 vector of regression coefficient parameters, $\mathbf{X}$ is an $n$-by-$(p+1)$ matrix of predictor variables, and matrix $\mathbf{V}$ is a function of the connectivity matrix $\mathbf{C}$ and $\rho$. Essentially expression (3.1) is a standard probability density function expression found in introductory multivariate textbooks. The normalizing constant (the Jacobian of the transformation from a spatially autocorrelated to a spatially unautocorrelated mathematical space, a concept discussed both in calculus and in introductory mathematical statistics textbooks), is $\frac{1}{2} LN[\det(\mathbf{V})]$ in this case. The term $\det(\mathbf{V})$ can be rewritten in terms of the eigenvalues of matrix $\mathbf{V}$, which for the popular spatial autoregressive models

translates this normalizing factor into

conditional autoregressive (CAR) model: $\frac{1}{2}\sum_{j=1}^{n} LN(1-\rho\lambda_j)$

simultaneous autoregressive (SAR)/autoregressive response (AR) model: $\sum_{j=1}^{n} LN(1-\rho\lambda_j)$

where the $\lambda_j$s are the $n$ eigenvalues of matrix $\mathbf{C}$ or its row-standardized version, matrix $\mathbf{W}$ $\left(\text{i.e., } w_{ij} = \frac{c_{ij}}{\sum_{j=1}^{n} c_{ij}}\right)$, depending upon which of these two geographic weights matrices is used in a spatial analysis. Of note is that the AR specification differs slightly from expression (3.1).

The auto-logistic specification (assuming pairwise-only dependence between sites), which can be extended to the auto-multinomial, for binary variable $Y$ is of the form

$$\Pr(\mathbf{Y}) = \frac{\text{EXP}(\mathbf{Y}'\mathbf{X}\boldsymbol{\beta} + \rho\mathbf{Y}'\mathbf{C}\mathbf{Y})}{\text{normalizing constant}}, \tag{3.2}$$

where $\Pr(\mathbf{Y})$ denotes the probability of vector $\mathbf{Y}$, EXP denotes the base-$e$ (natural) anti-logarithm, and the normalizing constant in the denominator is an infinite sum for an infinite region, one that has no closed form since it is a function of the unknown parameters. One cumbersome way to handle this term is through simulation work that allows the normalizing constant to be approximated (e.g., MCMC). Another way is by using the pseudolikelihood, which trades away statistical efficiency in exchange for a closed-form expression that avoids working with this unwieldy normalizing constant (Cressie 1991, p. 461). And, Heagerty and Lele (1998) make the very appealing suggestion of using a composite likelihood approach.

The auto-Poisson specification (assuming pairwise-only dependence between sites) results from specifying that all components of a variable have Poisson conditional probability mass functions, and for integer counts variable $\mathbf{Y}$ is of the form

$$\Pr(\mathbf{Y}) = \frac{\text{EXP}\left[\mathbf{Y}'\mathbf{X}\boldsymbol{\beta} - \sum_{i=1}^{n} LN(y_i!) + \rho\mathbf{Y}'\mathbf{C}\mathbf{Y}\right]}{\text{normalizing constant}}, \tag{3.3}$$

where $y_i$ denotes the $i$-th element of vector $\mathbf{Y}$, $y_i!$ denotes the number of permutations of size $y_i$, and the normalizing constant is a sequence of $n$ infinite sums, one that is dependent upon the parameters. Here the normalizing constant is intractable (Cressie 1991, p. 462). In addition, spatial dependence is restricted to being negative in nature, a result that is counter to the real world. Kaiser and Cressie (1997) propose a clever solution to this last problem, one that differs qualitatively from the solution suggested in this paper.

For these last two cases, Cressie emphasizes that the "real hurdle to obtaining the likelihood (from which inference on the parameters can proceed)

remains the normalizing constant" (1991, p. 440). In addition, both can be approximated with an auto-log-Gaussian specification.

## 4 A conceptual framework

Consider the constant mean specification of expression (3.1), for which $\mathbf{X\beta} = \mu\mathbf{1}$ and $\left(\mathbf{I} - \frac{\mathbf{11}'}{n}\right) = \mathbf{M}$. To begin, the Moran Coefficient may be rewritten, using both matrix and summation notation, as

$$\mathrm{MC} = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{n}{\mathbf{1}'\mathbf{C1}} \frac{\mathbf{Y}'\mathbf{MCMY}}{\mathbf{Y}'\mathbf{MY}}, \tag{4.1}$$

where the projection matrix $\mathbf{M}$, commonly appearing in statistical theory (Healy, 1986), centers the variable $\mathbf{Y}$ (i.e., subtracts the sample mean, $\bar{y}$, from each observed value, $y_i$). Of note is that this matrix is idempotent, meaning $\mathbf{M}^2 = \mathbf{M}$.

de Jong et al. (1984) establish the exact extremes of the MC in terms of part of the numerator of the right-hand fraction of expression (4.1), namely

$$\mathbf{MCM}. \tag{4.2}$$

These extremes approximately equal $\lambda_2$ and $\lambda_n$ of matrix $\mathbf{C}$. Tiefelsdorf and Boots (1995) extend de Jong et al.'s results, demonstrating that each of the $n$ eigenvalues of expression (4.2) is a MC value, once it is multiplied by the left-hand term of expression (4.1), namely $\frac{n}{\mathbf{1}'\mathbf{C1}}$.

In extending the findings of Tiefelsdorf and Boots (1995), and linking them to principal components analysis (PCA; Griffith 1984), the eigenvalues of expression (4.2) may be interpreted in the context of map pattern as follows:

The first eigenvector, $\mathbf{E}_1^*$, of expression (4.2) is the set of numerical values that has the largest MC achievable by any set of numerical values, for the given geographic connectivity matrix $\mathbf{C}$. The second eigenvector is the set of numerical values that has the largest achievable MC by any set of numerical values that is uncorrelated with $\mathbf{E}_1^*$. This sequential construction of eigenvectors continues through $\mathbf{E}_n^*$, which is the set of numerical values that has the largest negative MC achievable by any set of numerical values that is uncorrelated with the preceding $(n-1)$ eigenvectors.

Hence these $n$ eigenvectors describe the full range of all possible mutually orthogonal map patterns, and may be interpreted as synthetic map variables. Constructing a Moran scatterplot by plotting $\mathbf{CE}_j^*$ versus $\mathbf{E}_j^*$ always reveals a linear alignment of points, but with a slope that varies with $j$. The slope defined by $\mathbf{CE}_1^*$ versus $\mathbf{E}_1^*$ is the maximum positive one, and for $\mathbf{CE}_n^*$ versus $\mathbf{E}_n^*$ is the maximum negative one, with the $(n-2)$ intervening slopes sequentially rotating from this first to this nth case.

Paralleling PCA, the eigenvectors of expression (4.2) can be used to construct linear combinations with prespecified levels of spatial autocorrelation. In fact, suppose $\mathbf{b} = \mathbf{E}^*\mathbf{M}^2\mathbf{Y}$. Then expression (4.1) may be rewritten as

$$\frac{\mathbf{n}}{\mathbf{1}'\mathbf{C}\mathbf{1}}\frac{\mathbf{b}'\mathbf{\Lambda}\mathbf{b}}{\mathbf{b}'\mathbf{b}} = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}}\frac{\sum_{j=1}^{n} b_j^2\lambda_j}{\sum_{j=1}^{n} b_j^2}, \tag{4.3}$$

where the vector $\mathbf{b}$ actually is an estimate of a regression coefficient vector $\boldsymbol{\beta}$. In other words, judiciously selected weights used to construct a linear combination of eigenvectors allow a map pattern to be constructed having a prespecified MC value. Because these weights are squared, and any eigenvector is determined up to a multiplicative factor of $-1$, the signs of the coefficients are irrelevant. Rather, the resulting target MC value is a weighted average, and hence requires eigenvectors to be combined that have MC values both greater than and less than the target MC value.

Expression (4.3) alludes to spatial autocorrelation being directly linked to conventional OLS (ordinary least squares) regression in the auto-normal case. Consider the following pure SAR model:

$$\mathbf{MY} = \rho\mathbf{CMY} + \varepsilon, \tag{4.4}$$

where $\varepsilon$ is an $n$-by-1 vector of errors. This equation may be rewritten with a spectral decomposition as a spatially filtered OLS specification:

$$\begin{aligned}\mathbf{MY} &= \rho\mathbf{ME}\mathbf{\Lambda}\mathbf{E}'\mathbf{MY} + \varepsilon \\ &\approx \rho\mathbf{E}^*\mathbf{\Lambda}\mathbf{E}^{*\prime}\mathbf{Y} + \varepsilon = \mathbf{E}^*\boldsymbol{\beta} + \varepsilon.\end{aligned} \tag{4.5}$$

Of note is that $\mathbf{b} = \mathbf{E}^{*\prime}\mathbf{Y}$ is the standard OLS estimator when vector $\mathbf{Y}$ is regressed on matrix $\mathbf{E}^*$, as well as it appears in both the numerator and the denominator of the MC as written in expression (4.3). This solution is analogous to using PCA synthetic variables in regression in order to handle multicollinearity[1]. But here the eigenvectors of expression (4.2) are introduced in order to account for spatial autocorrelation. By accounting for spatial autocorrelation with judiciously selected eigenvectors (i.e., ones with significant regression coefficients, representing compelling degrees of spatial autocorrelation, and the spatial autocorrelation nature indicated by the MC), behavior of the error term will mimic that of independence. Of note is that matrix $\mathbf{ME}$ approximately equals $\mathbf{E}^*$, the eigenvectors of expression (4.2). Furthermore, this decomposition supplies a foundation for the derivation of expression (2.1).

[1] PCA orthogonalizes the covariance matrix for a set of $p$ variables, in $R$-mode, and uses eigenvectors of this covariance matrix to construct synthetic variates that are orthogonal and uncorrelated and can be used as predictors in a regression analysis. The solution presented here orthogonalizes the covariance matrix for a set of $n$ observations, similar to what is done by PCA in $Q$-mode, and uses the eigenvectors themselves as synthetic variates that are uncorrelated and can be used as predictors in a regression analysis.

## 5 Empirical demonstrations

Four data sets whose analyses are reported in the literature have been chosen for comparative purposes here. The first three provide a comparison over a wide range of sample sizes, while the fourth furnishes a comparison in the presence of negative spatial autocorrelation. Estimation results for these analyses are tabulated in Table 1. Two additional comparisons have been completed that are not summarized in detail here; one is for the Griffith and Amrhein Syracuse pediatric lead poisoning data (1997, p. 313), and the other is for the Haining Glasgow cancer data (1990, p. 366). These demonstrations emphasize advantages and disadvantages of using judiciously selected eigenvectors to filter spatial autocorrelation out of the error term posited for georeferenced data.

For the first demonstration consider Anselin's (1988, pp. 193, 197) proposal for an initial regression model that describes crime (the combined total of residential burglaries and vehicle thefts per 1,000 households) in Columbus, Ohio, as a function of income and housing value. These raw data conform reasonably well to a normal distribution [Shapiro-Wilk (S-W) = 0.96409, probability under the null hypothesis ($p$) > 0.10]. Variation across the four quadrants of the Columbus geographic landscape displays a tolerable degree of heterogeneity[2] (Bartlett = 4.651, $p > 0.10$). Marked positive spatial autocorrelation is displayed by the geographic distribution (MC = 0.52064, $z \approx 5.6$).

Because the detected latent positive spatial autocorrelation should be linked to a spatial process, such as crime displacement, and because variables more than likely are missing from the equation specification, spatial autocorrelation is viewed here as entering into the analysis through the error term; as such, an SAR model specification has been estimated. Data analysis results reported in Table 1 differ slightly from those reported by Anselin because of the different spatial autoregressive model specification employed here. Comparison of the autoregressive and filtering results reveals: (1) marked positive spatial autocorrelation present in the raw crime data persists in the OLS residuals, and essentially is completely accounted for by the spatial SAR model and eigenvector filtered OLS specifications; (2) the filtered OLS specification may do a modestly better job than the SAR one in accounting for spatial autocorrelation; and, (3) the non-normality and variance heterogeneity detected for the filtered OLS solution is attributable to the presence of an outlier (Neighborhood #4), which also is conspicuous in the SAR residual results but only potentially so in the traditional OLS results. Of note is that the three eigenvectors were selected from the set identified as having MC > 0.

For the second demonstration consider Griffith's (1992) analysis of the geographic distribution of 1986 median family income across the Ottawa-Hull metropolitan region as a function of population density of census tracts. These raw data fail to adequately conform to a normal distribution (S-W = 0.96254,

[2] Variance heterogeneity may be quantified with Bartlett's statistic, which assumes normality, as well as Levene's statistics, which assumes a continuous variable. While both statistics were computed as part of the analyses undertaken in this research, only Bartlett statistics are reported. For the case of perfectly homogeneous variance, Bartlett = 0; the parent populations have exactly the same variance.

**Table 1.** Comparative estimation results for selected georeferenced data sets having spatial autoregressive results reported in the literature

| | OLS | s.e. | S/AR(**W**) | s.e. | filtered OLS | s.e. |
|---|---|---|---|---|---|---|
| Anselin's Columbus crime data ($n = 49$) | | | | | | |
| $\hat{\rho}$ | × | × | 0.56163 | 0.20641 | × | × |
| $b_0$ | 68.61886 | 4.73547 | 59.89390 | 5.98608 | 59.43748 | 4.29781 |
| $b_{\text{income}}$ | −1.59730 | 0.33413 | −0.94132 | 0.36763 | −0.95530 | 0.28637 |
| $b_{\text{house value}}$ | −0.27393 | 0.10320 | −0.30226 | 0.09442 | −0.27516 | 0.08750 |
| $b_{\mathbf{E}_3^*}$ | × | × | × | × | −44.95254 | 9.75474 |
| $b_{\mathbf{E}_5^*}$ | × | × | × | × | −24.37356 | 9.35443 |
| $b_{\mathbf{E}_2^*}$ | × | × | × | × | −23.90644 | 9.95318 |
| (pseudo-)$R^2$ | 0.552 | | 0.658 | | 0.742 | |
| *residuals* | | | | | | |
| S-W | 0.98303 ($p > 0.10$) | | 0.96452 ($p > 0.10$) | | 0.95051 ($p < 0.10$) | |
| Bartlett | 6.154 ($p > 0.10$) | | 4.289 ($p > 0.10$) | | 8.911 ($p < 0.05$) | |
| MC | 0.24220 ($\lvert z\rvert \approx 2.6$) | | 0.03908 ($\lvert z\rvert \approx 0.4$) | | −0.02895 ($\lvert z\rvert \approx 0.3$) | |
| Griffith's Ottawa-Hull income data ($n = 192$) | | | | | | |
| $\hat{\rho}$ | × | × | 0.52168 | 0.12058 | × | × |
| $b_0$ | 43655 | 1018.85 | 42581 | 1641.19 | 41986 | 820.72 |
| $b_{\text{population density}}$ | −1.02916 | 0.26745 | −0.69221 | 0.31957 | −0.45751 | 0.22050 |
| $b_{v-\bar{v}}$ | −0.35936 | 0.07836 | −0.29755 | 0.12706 | −0.14358 | 0.06462 |
| $b_{I_{-\text{outlier(CT}\neq 50)}}$ | × | × | × | × | −20368 | 7352.01 |
| $b_{I_{+\text{outlier(CT}\neq 52)}}$ | × | × | × | × | −28345 | 7442.61 |
| $b_{\sum_{j=1}^{k} b_j \mathbf{E}_j^*}$ | × | × | × | × | 0.92056 | 0.08945 |
| (pseudo-)$R^2$ | 0.163 | | 0.335 | | 0.499 | |
| *residuals* | | | | | | |
| S-W | 0.98648 ($p > 0.10$) | | 0.98636 ($p > 0.10$) | | 0.98389 ($p > 0.10$) | |
| Bartlett | 5.555 ($p > 0.10$) | | 5.778 ($p > 0.10$) | | 5.437 ($p > 0.10$) | |
| MC | 0.24110 ($\lvert z\rvert \approx 5.4$) | | 0.00716 ($\lvert z\rvert \approx 0.2$) | | −0.00118 ($\lvert z\rvert \approx 0.0$) | |

$p < 0.01$). Variation across the four quadrants of the Ottawa-Hull geographic landscape displays a modest degree of heterogeneity (Bartlett = 1.703, $p > 0.10$). And, moderate positive spatial autocorrelation is displayed by this geographic distribution (MC = 0.35839, $z \approx 8.1$).

Because the detected latent positive spatial autocorrelation should be linked to a spatial process, such as similar socioeconomic groups clustering together, and because variables most certainly are missing from the equation specification, spatial autocorrelation is viewed here as entering into the analysis through the error term; as such, again an SAR model specification has been estimated. As the number of eigenvectors increases, an attempt to search through them for compelling relationships becomes more problematic. The exploratory strategy employed with this example is to select a threshold MC value below which all eigenvectors are weighted zero, and then construct a linear combination of the remaining ones to include as a synthetic variable in the regression equation; the threshold used for these Ottawa-Hull data is 0.5; $b_j = \mathbf{E}_j^{*T}\mathbf{Y} \approx 0$ will have the effect of essentially excluding a nonsignificant eigenvector from this linear combination. Accordingly, this synthetic variate has multiple degrees of freedom associated with it ($k = 28$ in this Ottawa-Hull case). Comparison of the autoregressive and filtering results reveals: (1) positive spatial autocorrelation present in the raw income data persists in the OLS

**Table 1.** (continued)

| | OLS | s.e. | S/AR(**W**) | s.e. | filtered OLS | s.e. |
|---|---|---|---|---|---|---|
| Griffith & Can's Toronto population density data ($n = 731$) | | | | | | |
| $\hat{\rho}$ | × | × | 0.60796 | 0.04850 | × | × |
| $b_0$ | 9.03133 | 0.02885 | 9.10907 | 0.06342 | 8.62482 | 0.02636 |
| $b_{\text{distance from CBD}}$ | −0.02850 | 0.00130 | −0.01076 | 0.00178 | −0.00550 | 0.00131 |
| $b_{I_{\text{excessdensity}}}$ | 2.07654 | 0.46657 | 1.98728 | 0.38943 | 1.91199 | 0.33930 |
| $b_{I_{\text{near-zero density}}}$ | −1.43890 | 0.26986 | −1.46195 | 0.22522 | −1.31746 | 0.19627 |
| $b_{\sum_{j=1}^{k} b_j \mathbf{E}_j^*}$ | × | × | × | × | 0.89494 | 0.03513 |
| (pseudo-)$R^2$ | 0.419 | | 0.600 | | 0.693 | |
| *residuals* | | | | | | |
| S-W | 0.98471 (p > 0.10) | | 0.98353 (p > 0.10) | | 0.98749 (p > 0.10) | |
| Bartlett | 17.628 (p < 0.01) | | 6.948 (p < 0.10) | | 10.173 (p < 0.05) | |
| MC | 0.40656 ($\lvert z\rvert \approx 7.6$) | | 0.04976 ($\lvert z\rvert \approx 0.9$) | | −0.01359 ($\lvert z\rvert \approx 0.2$) | |
| Anselin's southwest Ohio unemployment data ($n = 25$) | | | | | | |
| $\hat{\rho}$ | × | × | −0.62307 | 0.21957 | × | × |
| $b_0$ | 0.96650 | 0.02219 | 0.98796 | 0.01334 | 0.98856 | 0.01602 |
| $b_{\text{1983 unemployment}}$ | 0.96033 | 0.31775 | 1.01484 | 0.27580 | 0.62394 | 0.23207 |
| $b_{\text{1983 net migration}}$ | −1.02676 | 0.34417 | −0.88492 | 0.30476 | −0.72395 | 0.26389 |
| $b_{I_{\text{SMSAnear-zero density}}}$ | −0.02609 | 0.01406 | −0.01659 | 0.01294 | −0.01304 | 0.01050 |
| $b_{\mathbf{E}_{21}^*}$ | × | × | × | × | −0.06216 | 0.01557 |
| $b_{\mathbf{E}_{14}^*}$ | × | × | × | × | 0.03325 | 0.01491 |
| $b_{\mathbf{E}_7^*}$ | × | × | × | × | 0.03378 | 0.01516 |
| (pseudo-)$R^2$ | 0.486 | | 0.636 | | 0.789 | |
| *residuals* | | | | | | |
| S-W | 0.98071 ($p > 0.10$) | | 0.96826 ($p > 0.10$) | | 0.92218 ($p < 0.10$) | |
| Bartlett | 6.987 ($p < 0.10$) | | 2.628 ($p > 0.10$) | | 1.248 ($p > 0.10$) | |
| MC | −0.27733 ($\lvert z\rvert \approx 2.1$) | | −0.15546 ($\lvert z\rvert \approx 1.2$) | | −0.17362 ($\lvert z\rvert \approx 1.3$) | |

residuals, and essentially is well accounted for by the spatial SAR model and eigenvector filtered OLS specifications; (2) the filtered OLS specification may do a slightly better job accounting for the geographic variation in median family income; (3) coefficient and standard error estimates for population density are volatile across these implementations; and, (4) the filtered OLS solution identifies two prominent outliers, one being an extreme in the right-hand tail and the other in the left-hand tail of the frequency distribution. Of note is that the goodly number of degrees of freedom associated with the synthetic variate is consistent with the notion of spatial autocorrelation indexing redundant information in georeferenced data.

For the third demonstration consider Griffith and Can's (1996) formulation of a spatial autoregressive population density model for the 1986 Toronto-centered region in terms of a negative exponential function of distance separating census tracts (CTs) from the CBD (central business district). The raw data fail to adequately conform to a normal distribution (S-W = 0.78683, $p < 0.01$). These researchers have uncovered several outliers, one being an excessively high density, and several others being zero or near-zero densities. Results reported here differ slight from those of Griffith and Can because the constant 1604.6 was added to population density before a logarithmic transformation was performed on this variable, and because the CBD coordinate is

estimated from the data rather than being set in this exercise; the principle consequence is the removal of a weak east-west trend from the data.

Because the detected latent positive spatial autocorrelation should be linked to a spatial response mechanism, such that localities with similar population densities will tend to cluster together in urban space, spatial autocorrelation is viewed here as entering into the analysis as a direct response to density itself. Therefore, an AR model specification has been estimated. The exploratory strategy again is used here, employing a threshold MC value of 0.5, which in turn results in construction of the synthetic variate using $k = 118$ eigenvectors. Comparison of the autoregressive and filtering results reveals: (1) marked positive spatial autocorrelation present in the raw population density data persists in the OLS residuals, and essentially is adequately accounted for by the spatial AR model and eigenvector filtered OLS specifications; (2) just as Griffith and Can report, these data contain considerable spatial heterogeneity; and, (3) as Curry (1972) shows, estimation procedures can confuse spatial autocorrelation and distance decay effects.

For the fourth demonstration consider Anselin's (1988, 206–210) spatial econometrics analysis of a georeferenced socioeconomic attribute data set that relates to a spatial economics version of the classical economics Philips curve (i.e., a graph depicting the relationship between inflation and unemployment). His study area comprises the 25 counties of southwest Ohio. The dependent variable in his analysis is the change in wage rates for 1983. He reports results from fitting both a conventional OLS and a spatial AR model to this variable, using the inverse unemployment rate, a net migration rate, and a standard metropolitan statistical area (SMSA) indicator variable as predictors. This example is of particular interest because it involves modest negative spatial autocorrelation.

The filtered OLS analysis employed here restricts attention to eigenvectors depicting negative spatial autocorrelation. Comparison of the autoregressive and filtering results reveals: (1) negative spatial autocorrelation lurking in the raw county unemployment rates, which is largely accounted for by the spatial AR model and eigenvector filtered OLS specifications; and, (2) a reversal of the inference concerning the SMSA indicator variable based upon either the AR or the filtered OLS estimation results. Of note is that theoretically negative spatial autocorrelation is affiliated with a suppression of the number of degrees of freedom, whereas the filtered OLS results reported here suggest that degrees of freedom are lost in its presence.

Noteworthy results reported in Table 1 for comparative purposes concern the estimated regression coefficients and their corresponding standard errors. For Anselin's crime data there is little difference between the income and house value coefficient estimates across the three specifications, and little difference between the intercept estimates obtained for the SAR and spatial filtering specifications. The standard errors obtained with the spatial filtering specification change disproportionately to the change in $R^2$. For Griffith's income data there are substantial differences between the non-intercept regression coefficients and the standard errors across the three specifications. For Griffith and Can's population density data results for the outlier indicator variables appear to be the most stable across the three specifications. Finally, for Anselin's unemployment data the estimated regression coefficient for the unemployment rate differs most noticeably between the SAR and spatially filtered specifications. Volatilities displayed by these results more than likely are

largely attributable to spatial autocorrelation latent in the **X** matrix. Modifying the eigenvector spatial filtering approach to more closely parallel Getis's (1995) spatial filtering approach would acknowledge and clarify this situation.

## 6 Implications about the proposed specification for spatial auto-logistic and auto-poisson models

Especially the empirical demonstrations completed employing log-Gaussian analyses for percent of children having elevated blood lead levels, and for lung cancer rates, suggest that the eigenvector filtering methodology offers a convenient way to accommodate the presence of spatial autocorrelation in georeferenced data while retaining more conventional statistical model specifications. This implication is particularly encouraging for logistic and Poisson regression modeling, where the normalizing factors for the spatial auto- versions are analytically intractable, and because the spatial auto-Poisson specification cannot accommodate the almost always encountered case of positive spatial autocorrelation. Two additional empirical demonstrations are reviewed in this section, one for each of these two auto- situations. Their comparative estimation results are reported in Table 2.

The geographic distribution of the percentage of urban population by municipio across Puerto Rico in 1970 can be described with a logistic model specification (Griffith and Amrhein 1997, pp. 306–308). This geographic distribution should reflect the urban hierarchy on the island. Accordingly, the description devised here is based upon the five urban regions of Puerto Rico; moreover, it is an ANOVA. Based upon a similar argument to that presented for the preceding population density analysis, in this situation a spatial AR model has been estimated. Of note is that the log-Gaussian approximation involves the logarithmic transformation $LN\left(\frac{\%\text{ urban population} + 13.80998}{100 - \%\text{ urban population} + 13.80998}\right)$. Comparison of the autoregressive and filtering results reveals: (1) weak positive spatial autocorrelation present in the percentage urban population data ($MC = 0.16164$) is accounted for by the spatial AR model and eigenvector filtered OLS specifications as well as the filtered logistic regression specification; (2) as expected, the log-Gaussian approximation analyses are plagued by nonconstant variance; and, (3) because the logistic regression results are based upon actual population figures, conspicuous statistical inference differences appear between the log-Gaussian approximation and logistic regression estimations (i.e., the sample sizes used to compute the standard errors are tens of thousands rather than 72).

Repeating this logistic regression exercise with 1980 georeferenced data uncovers a third prominent eigenvector for filtering spatial autocorrelation effects. Because the eigenvectors do not change between 1970 and 1980, the statistical analysis links to a fixed effects model specification.

Meanwhile, Cressie (1991, pp. 386–389) reports georeferenced data for SIDs (sudden infant death syndrome) cases during 1974–1978 and 1979–1984 in North Carolina. Geographical clustering of SIDs cases already has been documented for this North Carolina data set. Cressie (1991, p. 429) comments that an auto-Poisson model would constitute a reasonable spatial autoregressive specification here and that a geographic trend in SIDs cases occurs across the set of 100 North Carolina counties. He also notes that a suitable

**Table 2.** Comparative estimation results for selected georeferenced data sets conforming to non-Gaussian distributions

| | log-Gaussian AR(**W**) | s.e. | log-Gaussian filtered OLS | s.e. | filtered logistic/ Poisson | s.e. |
|---|---|---|---|---|---|---|
| Griffith & Amrhein's Puerto Rico percent urban population data ($n = 73$) | | | | | | |
| $\hat{\rho}$ | 0.21545 | 0.23020 | × | × | × | × |
| $b_0$ | −0.54828 | 0.13032 | −0.55863 | 0.09482 | −0.3042 | 0.00158 |
| $b_{\text{San Juan-Caguas}}$ | 0.34719 | 0.19420 | 0.46193 | 0.17271 | 0.9445 | 0.00252 |
| $b_{\text{Arecibo-Caguas}}$ | −0.17356 | 0.23031 | −0.50777 | 0.22184 | −1.1052 | 0.00346 |
| $b_{\text{Mayaguez-Caguas}}$ | −0.19272 | 0.20197 | −0.21366 | 0.20159 | 0.2058 | 0.00350 |
| $b_{\text{Ponce-Caguas}}$ | 0.10840 | 0.20527 | 0.29648 | 0.20041 | 0.4381 | 0.00309 |
| $b_{\mathbf{E}_7^*}$ | × | × | −2.57779 | 0.88176 | 5.1415 | 0.01290 |
| $b_{\mathbf{E}_6^*}$ | × | × | 1.98345 | 0.86773 | −3.5605 | 0.01260 |
| (pseudo-)$R^2$ | 0.118 | | 0.251 | | 0.213 | |
| *residuals* | | | | | | |
| S-W | 0.96799 ($p > 0.10$) | | 0.97033 ($p > 0.10$) | | × | |
| Bartlett | 12.622 ($p < 0.05$) | | 8.759 ($p < 0.10$) | | 3.257 ($p > 0.10$) | |
| MC | 0.02287 ($\lvert z\rvert \approx 0.3$) | | −0.03776 ($\lvert z\rvert \approx 0.5$) | | 0.06530 ($\lvert z\rvert \approx 0.9$) | |
| Cressie's North Carolina SIDs data ($n = 100$) | | | | | | |
| $\hat{\rho}$ | 0.23625 | 0.18362 | × | × | × | × |
| $b_0$ | −6.24821 | 0.07560 | −6.26602 | 0.0497 | −6.31137 | 0.0460 |
| $b_{u-\bar{u}}$ | 0.00167 | 0.00064 | 0.00180 | 0.0004 | 0.00248 | 0.0005 |
| $b_{v-\bar{v}}$ | −0.00364 | 0.00186 | −0.00599 | 0.0017 | −0.00457 | 0.0017 |
| $b_{I_{\text{low value outlier}}}$ | −1.71068 | 0.57770 | −1.81613 | 0.5052 | −13.38106 | 346.0466 |
| $b_{I_{\text{high value outlier}}}$ | × | × | 1.42739 | 0.5043 | 1.51990 | 0.2803 |
| $b_{\mathbf{E}_4^*}$ | × | × | 1.45647 | 0.6323 | 1.28682 | 0.5797 |
| $b_{\mathbf{E}_{15}^*}$ | × | × | −1.06673 | 0.4946 | −0.86467 | 0.3708 |
| $b_{\mathbf{E}_{16}^*}$ | × | × | −2.20995 | 0.4943 | −2.43255 | 0.4652 |
| $b_{\mathbf{E}_{19}^*}$ | × | × | −1.14049 | 0.5060 | −1.12825 | 0.4585 |
| (pseudo-)$R^2$ | 0.265 | | 0.482 | | 0.884 | |
| *residuals* | | | | | | |
| S-W | 0.98539 ($p > 0.10$) | | 0.98313 ($p > 0.10$) | | × | |
| Bartlett | 7.813 ($p < 0.10$) | | 3.568 ($p > 0.10$) | | × | |
| MC | −0.01148 ($\lvert z\rvert \approx 0.2$) | | −0.07178 ($\lvert z\rvert \approx 1.1$) | | −0.01321 ($\lvert z\rvert \approx 0.2$) | |

power transformation – suggesting a preference for the logarithmic one – can be applied to the rates version of these data so that a Gaussian autoregressive specification can be estimated (p. 434). Cressie identifies an outlier in the 1974–1978 data set, and adds unity to each value to better computationally handle 0 values (p. 391). Results reported here differ slightly from those reported by Cressie because the optimal constant to add to the rates was found to be 0.34235, rather than 1.

An SAR spatial autoregressive specification has been estimated here because missing variables, such as ethnicity, local health care practices, and socioeconomic attributes of local areas, appear to account for geographic clustering of cases. Of note is that the statistical description furnished by a Poisson specification incorporating the OLS-identified prominent eigenvectors, to adjust for spatial autocorrelation, is considerably better than the description furnished by a log-Gaussian approximation. Comparison of the autoregressive and filtering results reveals: (1) modest positive spatial autocorrelation present in the raw SIDs data (MC = 0.24465) is well accounted

for by the spatial SAR model and eigenvector filtered OLS specification as well as the filtered Poisson regression specification; (2) as might be expected, the log-Gaussian approximation SAR analysis is characterized by nonconstant variance; and, (3) the low value outlier may well be an artifact of model misspecification.

Repeating this Poisson regression exercise with the 1979–1984 georeferenced data results in disappearance of two of the prominent eigenvectors for filtering spatial autocorrelation effects, as well as failure for extremely low values to emerge as outliers. Again because the eigenvectors do not change between the two time periods, the statistical analysis links to a fixed effects model specification.

## 7 Large georeferenced data sets

A conspicuous feature of each of the six preceding empirical analyses is their respective relatively small $n$ value. One of the sources of massively large georeferenced data sets is remote sensing. Bailey and Gatrell (1995, p. 254) analyze data from a one-square-kilometer, 30-by-30 pixels portion of the High Peak District in England; in practice a serious spatial scientist would rarely analyze a LANDSAT image this small, but this data set is invaluable for illustrative purposes. The focus of the analysis here is on the ratio of Band #4 to Band #3, bands that, respectively, represent the *near-infrared* and *red* wavelengths of the electromagnetic spectrum. This ratio provides a good picture of spatial variation in biomass, with healthy green vegetation reflecting strongly in Band #4, which measures the vigor of vegetation, whereas its energy absorption is strongly sensed in Band #3, which aids in the identification of plant species. The frequency distribution for this ratio deviates from a normal frequency distribution (S-W $= 0.94904$, $p < 0.01$), with marked variance fluctuation across the four quadrants of the region (Bartlett $=$ 91.561, $p < 0.01$). Such variance heterogeneity often typifies remotely sensed data. In addition, as is characteristic of most remotely sensed data, this ratio displays strong positive spatial autocorrelation (MC $= 0.88047$, $z \approx 36.7$).

An SAR spatial autoregressive specification has been estimated here because variables most certainly are missing for this simple analysis. The SAR and filtered OLS estimation results together with diagnostic statistics for this data set include

| | SAR (W) | s.e. | filtered OLS | s.e. |
|---|---|---|---|---|
| $\hat{\rho}$ | 0.98812 | 0.00886 | × | |
| $b_0$ | 2.10031 | 0.33445 | 2.10730 | 0.00300 |
| $b \sum_{j=1}^{k} b_j E_j$ | × | | 1.00000 | 0.00586 |
| (pseudo-) $R^2$ | 0.953 | | 0.970 | |
| *residuals* | | | | |
| W-S | 0.95641 | (p < 0.01) | 0.98300 | (p < 0.10) |
| Bartlett | 39.201 | (p < 0.01) | 22.670 | (p < 0.01) |
| MC | 0.16838 | | 0.14206. | |

The exploratory strategy was adopted here, and employs a threshold MC value of 0.25, which in turn results in construction of the synthetic variate using $k = 279$ eigenvectors. Comparison of these results reveals: (1) latent spatial autocorrelation is extremely strong, and contributes substantially to the variance heterogeneity across this geographic landscape; (2) pronounced positive spatial autocorrelation present in the raw spectral bands ratio data (MC = 0.38469) is not totally accounted for by either the spatial SAR model or the eigenvector filtered OLS specification; and, (3) the SAR and filtered OLS residuals are ill behaved. Of note is that the eigenvectors included here are the centered analytical ones for matrix **C**, which approximate those for expression (4.2); the first eigenvector has not been removed from this set. These eigenvectors are known analytically (Griffith 1999), which allows any size remotely sensed image to be analyzed using the filtered-OLS methodology.

## 8 Conclusions

The empirical demonstrations explored in this paper provide evidence that the eigenvectors of matrix expression (4.2) can be used with OLS regression to account for spatial autocorrelation during the statistical modelling of georeferenced data. This methodological finding is especially useful in the cases of auto-logistic and auto-Poisson analyses, for which the necessary autoregressive normalizing constant remains elusive. Future research should focus, at least in part, on deriving statistical distribution theory, more than likely within the context of fixed effects, for such model specifications, and to further refine the exploratory strategy repeatedly employed in this paper.

Stepwise regression analysis, which is illuminating here because the eigenvectors of expression (4.2) are uncorrelated, frequently indicates that many significant eigenvectors account for only a fraction of a percent of variance in a given dependent georeferenced variable. This is one reason why the exploratory composite index approach appears to be successful. Several recent scientific discoveries suggest that this may be the expected, rather than a concern, in attaining a better understanding of phenomena. For example, Hadi and Ling (1998) argue that PCA components associated with near-zero eigenvalues should not necessarily be routinely discarded. Recent experiments in physics reveal that neutrinos, which historically had been assigned a mass of zero, actually have a minute mass that, when accumulated across the trillions and trillions and trillions that exist, adds up to a substantial mass, explaining at least some of the missing mass puzzle that has bedeviled astrophysicists for decades. And, psychologists recently found that a battery of genes (there might be 50 or more) appears to influence intelligence, with the one discovered on the long arm of chromosome 6 of the human genome accounting for only about 2% of the variance in IQ.

Finally, the empirical demonstrations reviewed in this paper offer at least partial answers to the posed research problem questions; these may be summarized as follows:

How can the normalizing constant complication be avoided? Potentially by implementing the eigenvector filtering OLS methodology.
How can spatial statistical models be equated with conventional statistical models? Again, potentially by implementing the eigenvector filtering OLS methodology.

What does the spatial autocorrelation term in a spatial statistical model mean? It often means that very complex map pattern effects are complicating the statistical analysis of georeferenced data. Rarely do simple linear gradient trends fully account for these effects; rather, often some combination of numerous eigenvector map patterns is needed to account for latent spatial autocorrelation.

Therefore, the verdict is that the eigenvector filtering OLS methodology offers considerable promise for the proper statistical analysis in a regression context of spatial data.

## References


Anselin L (1988) *Spatial econometrics*. Kluwer, Dordrecht

Bailey T, Gatrell A (1995) *Interactive spatial data analysis*. Longman Scientific, Harlow

Barry R, Pace R (1997) Quick computation of spatial autoregressive estimators. *Geographical Analysis* 29:232–247

Conlon E, Waller L (1999) Flexible neighborhood structures in hierarchical models for disease mapping, submitted to *Biometrics*

Cressie N (1991) *Statistics for spatial data*. Wiley, New York

Curry L (1972) A spatial analysis of gravity flows. *Regional Studies* 6:131–147

de Jong P, Sprenger C, van Veen F (1984) On extreme values of Moran's I and Geary's *c*. *Geographical Analysis* 16:17–24

Getis A (1995) Spatial filtering in a regression framework. In: Anselin L, Florax R (eds) *New directions in spatial econometrics*, pp 172–185. Springer, Heidelberg Berlin New York

Griffith D (1984) Measuring the arrangement property of a system of areal units generated by partitioning a planar surface. In: Bahrenberg G, Fischer M, Nijkamp P (eds) *Recent developments in spatial analysis: Methodology, measurement, models*, pp 191–200. Gower, Aldershot, UK

Griffith D (1992) Estimating missing values in spatial urban census data. *Operational Geographer* 10(2):23–26

Griffith D (1996) Spatial autocorrelation and eigenfunctions of the geographic weights matrix accompanying geo-referenced data. *Canadian Geographer* 40:351–367

Griffith D (1999) Eigenfunction properties and approximations of selected incidence matrices employed in spatial analyses. Submitted to *Linear algebra and its applications* (revised version resubmitted)

Griffith D, Amrhein C (1997) *Multivariate statistical analysis for geographers*. Prentice-Hall, Englewood Cliffs, NJ

Griffith D, Can A (1996) Spatial statistical/econometric versions of simple urban population density models. In: Arlinghaus S (ed) *Practical handbook of spatial statistics*, pp 231–249. CRC Press, Boca Raton, FL

Griffith D, Sone A (1995) Trade-offs associated with normalizing constant computational simplifications for estimating spatial statistical models. *Journal of Statistical Computation and Simulation* 51:165–183

Hadi A, Ling R (1998) Some cautionary notes on the use of principal components regression. *The American Statistician* 52:15–19

Haining R (1990) *Spatial data analysis in the social and environmental sciences*. Cambridge University Press, Cambridge, UK

Heagerty P, Lele S (1998) A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association* 93:1099–1111

Healy M (1986) *Matrices for statistics*. Oxford University Press, New York

Kaiser M, Cressie N (1997) Modeling Poisson variables with positive spatial dependence. *Statistics and Probability Letters* 35:423–432

LeSage J (1997) Bayesian estimation of spatial autoregressive models. *International Regional Science Review* 20:113–129

Ord J (1975) Estimating methods for models of spatial interaction. *Journal of the American Statistical Association* 70:120–126

Rao C (1967) Least squares theory using an estimated dispersion matrix and its application to measurement of signals. In: LeCam L, Neyman J (eds) *Proceedings* of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1965/66, vol I, pp 355–372. University of California Press, Berkeley

Tiefelsdorf M, Boots B (1995) The exact distribution of Moran's I. *Environment and Planning A* 27:985–999

Tiefelsdorf M, Boots B (1996) Letters to the editor: The exact distribution of Moran's I. *Environment and Planning A* 28:1900