Computational Simplifications Needed for Efficient Implementation of Spatial Statistical Techniques in a GIS

Daniel A. Griffith* and Zhiqiang Zhang†

*Department of Geography, Syracuse University
Syracuse, NY 13244-1020
†Phone.com, 800 Chesapeake Drive
Redwood City, CA 94063

Abstract

This paper contributes to the ongoing debate about which spatial analysis functions should be coupled with a GIS by identifying research problems that need to be solved before a richer toolbox of spatial statistical techniques can be implemented in a GIS. Three general problem areas are addressed. The first replaces a sequential ordinary least squares linear regression implementation with a single regression analysis. The second establishes the effective sample size for a single variable in a georeferenced data set, a result useful when calculating confidence intervals for means. The third establishes the effective sample size for pairs of variables in a georeferenced data set, a result useful when calculating the significance of correlation coefficients. These three general problems allow four more specific research problems to be identified that are in need of definitive solutions before a richer toolbox of spatial statistical techniques can be relatively easily implemented in a GIS. Their complete solutions will involve both empirical assessments and simulation experiments. These four problems are represented by four principal equations posited in this paper, equations that offer considerable computational simplification for the implementation of spatial statistical techniques within a GIS. Sufficient evidence in support of them is presented here to allow their implementation at this time on an experimental basis. These equations remove the need for eigenfunction and nonlinear optimization routines, and maintain the standard linear regression technique as the workhorse of a GIS statistical analysis. They also strengthen the inferential basis for a spatial scientist.

I. INTRODUCTION

Zhang and Griffith (1997, 2000) contribute to the ongoing debate about which spatial analysis functions should be coupled with a GIS. These researchers outline how user-friendly spatial statistical analysis modules can be developed, illustrating such developments with ArcView scripts. In tandem they demonstrate how to implement a spatial statistical/GIS module in Access with component software technology. These two articles reflect upon the issue of which spatial statistical functions serve well as a bridge between GISs and classical statistical analysis systems, especially spatial autocorrelation testing and spatial autoregression modeling. This paper extends their discussion by identifying research problems that need to be solved before a richer toolbox of spatial statistical techniques can be implemented in a GIS. The main focus is on algorithms and methods that can simplify the spatial statistical computing in a GIS environment. This is critical for turning a basic desktop GIS into an efficient spatial analytical system, and is becoming increasingly important as spatial statisticians realize that simplifications seem to be the only choice when the spatial dataset being analyzed is massive in size.

Three general problem areas are addressed in this paper. The first replaces the sequential ordinary least squares linear regression implementation promoted by Zhang and Griffith (2000) with a single regression analysis. The second establishes the effective sample size—the equivalent sample size for independent observations—for a single variable in a georeferenced data set. This result is useful when calculating confidence intervals for means. The third establishes the effective sample size for pairs of variables in a georeferenced data set. This result is useful when calculating the significance of correlation coefficients. Of note is that the geographic connectivity matrix is assumed, as it can be extracted from spatial topology data contained in a GIS using either boundary files (for surface partitionings), Thiessen polygon arcs (for point data), or inter-point distances.

II. ESTIMATING THE AUTOREGRESSIVE PARAMETER $\hat{\rho}$ FROM A MORAN COEFFICIENT

A simple and precise way to circumvent the numerical intensity and computer memory requirements associated with implementing a simultaneous autoregressive model (SAR; see Griffith, 1988) involves exploiting spatial dependency latent in a Moran Co-

1082-4006/99/0502-97$5.00

efficient (MC). The MC is relatively simple to calculate, even for massively large georeferenced data sets and can be calculated using a linear regression algorithm (Griffith and Amrhein, 1997, pp. 44-45). The SAR spatial autocorrelation parameter, ρ, requires calculating the eigenvalues of a geographic contiguity matrix as well as nonlinear regression involving iterative nonlinear optimization. But the covariation between MC and $\hat{\rho}$ for the SAR model displays a strong and reasonably precise linear relationship for normally distributed variables, which is portrayed in Figure 1; the observed relationship is denoted by o's (o), whereas the predicted values are denoted by pluses (+). This graph was constructed from empirical data results associated with acceptable Shapiro-Wilk (S-W) statistics (i.e., values close to 1), which indicate conformity with a normal frequency distribution, based upon a number of empirical data sets (see Griffith and Layne, 1999).

The graph appearing in Figure 1 reveals a rather smooth curve depicting the relationship between MC and $\hat{\rho}$. Fitting an equation to this curve yields a useful approximation for the SAR autocorrelation parameter, namely

$$\hat{\hat{\rho}} = \frac{2}{1+e^{-4\frac{MC}{MC_{max}}}} - \frac{2}{1+e^{\frac{4}{(n-1)MC_{max}}}} , \qquad (1)$$

where MC_{max}denotes the maximum positive MC value that can be calculated with a given surface partitioning, and e is the base of the system of natural logarithms ($e \approx 2.71828$). This extreme MC value roughly equals $\frac{n}{\mathbf{1}^T\mathbf{C1}}\lambda_2$, where λ_2 is the second eigenvalue of the geographic weights matrix **C** used to calculate MC, and n denotes the sample size. Here **1** is an n-by-1 vector of ones, and T denotes the matrix transpose operation. This equation was specified in such a way that $\hat{\hat{\rho}} \equiv 0$ when MC = $-\frac{1}{n-1}$. For the empirical cases used to construct Figure 1, pseudo-R^2 = 0.878. The predicted values align reasonably well with their empirical counterparts. A 15-by-15 lattice simulation experiment—for which MC_{max} = 1.05084—further confirmed this equational form, yielding pseudo-R^2 = 0.999. The graph for this result appears in Figure 2; the simulated relationship is denoted by o's (o), whereas the predicted values are denoted by pluses (+).

Two assumptions underlie equation (1). The first is that the frequency distribution for the georeferenced data being analyzed conforms to a bell-shaped curve (i.e., a normal frequency distribution). The second is that λ_2 is easy to calculate.

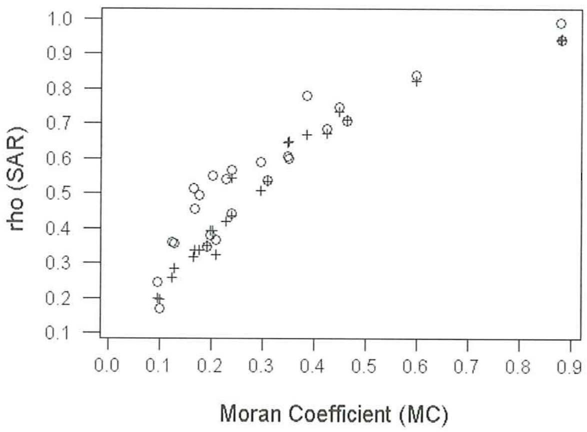

Figure 1. Predicting rho from a MC

Box-Cox Transformations

Griffith et al. (1998) discuss a quantile approach to selecting a power transformation to convert a non-normal variable to one that more closely mimics normally distributed data. Often georeferenced environmental data reflect a log-normal rather than a normal distribution. In this case the transformation of interest for some georeferenced variable Y is LN(Y + δ), where LN denotes the natural logarithm and is equivalent to a power transformation whose exponent is zero. A quick way to estimate δ is given by the following procedure:

> Select a systematic sample of size m, say m = 50, from across the data range, [$-y_{min}$, y_{ma}]. Let this set of values be the candidate set for parameter δ. Calculate the S-W statistic for each of the m transformations, LN(Y + $\hat{\delta}_j$), j = 1, 2, ..., m. Equate $\hat{\delta}$ to that δ_j having the largest S-W statistic.

This procedure tends to render a good estimate for $\hat{\delta}$, but not necessarily the optimal estimate since it fails to include the Jacobian term included in a Box-Cox specification. Fortunately, this deviation seems to disappear as n increases. An investigation of Haining's (1990, 1991) georeferenced Glasgow epidemiological data yields, for three variables having significant S-W statistics, those illustrative tabulations appearing in Table 1.

Clearly this selection method works well for log-normal transformations, and requires that only the traditional S-W statistic be added to a GIS toolbox. The

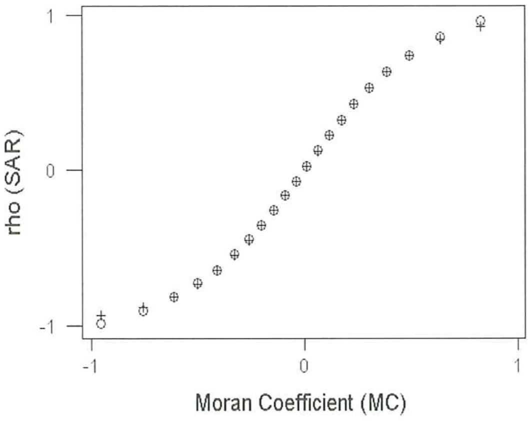

Figure 2. Predicting rho from a MC: 15-by-15 lattice simulation results

plot of S-W versus $\hat{\delta}$ portrays a hyperbola curve, which could be used for local interpolation in a way that parallels the local quadratic interpolation procedure outlined by Zhang and Griffith (2000).

The remaining problem pertains to the case of a non-zero power transformation. A procedure could be devised that systematically evaluates an exponent, say γ, across the interval [-2, 2] as well as δ across the data range [$-y_{min}$, y_{max}]. But future research is needed to discover a more efficient approach.

Estimating MC_{max} for a Geographic Connectivity Matrix

In general MC_{max} approximately equals 1. But in many cases this quantity exceeds 1, whereas in some cases it is less than 1. Precisely, MC_{max} is determined by the maximum eigenvalue of matrix

$$(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^T}{n})\mathbf{C}(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^T}{n})$$, where **I** is the identity matrix.

This matrix appears in the numerator of the MC formula. This eigenvalue is approximately and asymptotically equivalent to the second largest eigenvalue of matrix **C**.

The principal eigenvalue of binary matrix **C** can be quickly calculated, even for very large n, using one of the oldest and the well-known method of

$$\operatorname*{limit}_{k\to\infty} \frac{\mathbf{1}^T\mathbf{C}^{k+1}\mathbf{1}}{\mathbf{1}^T\mathbf{C}^k\mathbf{1}} = \lambda_1$$, where k is a positive integer,

for matrix **C** (Chatelin, 1993, p. 213). Of note is that this ratio is very easy to program, especially since matrix **C** contains mostly zeroes. In fact, the computation of this ratio can be greatly simplified in a GIS environment by using a polygon neighbors list file (Zhang and Griffith, 1997), with each row of this file recording the identification numbers of all the spatial neighbors of a specific polygon. An n-by-1 vector **A**, when right-multiplied by matrix **C**, results in a new n-by-1 vector whose i-th element is simply the sum of the elements of vector **A** that are spatially connected to polygon i. Since a polygon neighbors file can be easily generated from the topologic information stored in a GIS data set, the principal eigenvalue calculation algorithm illustrates just another advantage of combining GIS and spatial statistical computing. This advantage becomes even more pronounced in cases when the numerical intensity involved in solving eigensystems of a connectivity matrix based upon a large sample size is so massive that even powerful supercomputers cannot supply sufficient computing resources. The following is an Avenue script that illustrates how to implement this algorithm in ArcView.

```
'* Script: EigenValue.Principal
'* Description: Calculate the principal eigenvalue of
           a given binary matrix C (expressed as a neigh
           bor list)
'* Parameters: input — C (the neighbor list)

C = self.get(0)

'verify the inputs
if (C.Is(List).Not) then
  return nil
end

'generate an n-by-1 vector of ones
temp1 = list.make
n = C.count
for each i in 1..n
  temp1.add(1)
end
'find the principal eigenvalue using the Chatelin formula
```

Table 1. Sample Box-Cox transformation results

Standardized mortality rates	original	quantile selected $\hat{\delta}$	S-W$_{max}$ selected $\hat{\delta}$
Accidents	0.93773	$\hat{\delta}$ = 89; S-W = 0.9862	$\hat{\delta}$ = 73; S-W = 0.9864
Respiratory	0.95244	$\hat{\delta}$ = 32; S-W = 0.9740	$\hat{\delta}$ = 26; S-W = 0.9740
Cancer	0.96229	$\hat{\delta}$ = -23; S-W = 0.9787	$\hat{\delta}$ = -31; S-W = 0.9790

```
precision = 10 ^ (-5)
diff = n
count = 1
lamlag = -1

'denominator in the first loop (total number of connections)
denom = 0
for each i in 1..n
  denom = C.get(i-1).count + denom
end

temp = list.make
denom = 0
for each i in 1..n
  denom = C.get(i-1).count + denom
  temp.add(C.get(i-1).count)
end

while(true)
  for each i in 1..n
    sum = 0
    for each j in 1..(C.get(i-1).count)
      index = C.get(i-1).get(j-1)
      sum = temp.get(index) + sum
    end
    temp1.set( (i-1), sum )
  end

  numer = 0
  for each m in 1..n
    numer = temp1.get(m-1) + numer
  end

  lamda = numer / denom
  diff = lamda - lamlag
  if (diff > precision) then
    lamlag = lamda
    denom = numer
    temp = temp1.DeepClone
    continue
  else
    break
  end

end

return lamda
```

Expansion of the matrix expression $(\mathbf{I}-\frac{\mathbf{11}^T}{n})\mathbf{C}(\mathbf{I}-\frac{\mathbf{11}^T}{n})$ suggests that the principal eigenvalue of matrix **C** is being reduced by functions of the maximum row sum for matrix **C** and the total sum of the entries of matrix **C**, namely $\mathbf{1}^T\mathbf{C1}$. An empirical analysis of twelve judiciously selected surface partitionings, whose n values range from 75 to 513, yields

$$\hat{\lambda}_2 = \lambda_1 - \frac{3\times c_{max}}{2n} - \frac{7\times\mathbf{1}^T\mathbf{C1}}{3n^2}, \qquad (2)$$

where c_{max} denotes the maximum row sum. Although this result is based upon a rather small sample size, it is bolstered by its accompanying conceptual expectation as well as a very favorable performance in a simulation experiment involving 10,000 randomly selected regular square tessellations of pixels forming rectangular regions ranging in size from 10-by-10 to 10,000-by-10,000. The accompanying R^2 value for equation (2) is 0.978; the accompanying residuals appear to conform to a bell-shaped curve, but display considerable heteroscedasticity. Again, future research is needed in order to confirm, and possibly refine, this formula.

III. THE EFFECTIVE SAMPLE SIZE FOR A SINGLE GEOREFERENCED VARIABLE

Much of conventional statistical theory was developed using the independent and identically distributed (*iid*) assumption. This theory is replete with formulae including sample size (n) and degrees of freedom (df) terms. Accounting for the redundant information contained in georeferenced data can involve calculating the equivalent sample size and/or degrees of freedom for an *iid* data set. These equivalencies may be defined as effective sample size (n*) and effective degrees of freedom(df*); respectively they equal n and df when spatial autocorrelation is zero, and n* equals 1 when spatial autocorrelation is 1.

One simple statistical test commonly employed to zoom in on an average, namely the t-test performed on single sample means, can be modified to properly account for the presence of spatial autocorrelation in georeferenced data. Complications emerging in this statistical problem that lie dormant with *iid* data include: variance inflation/deflation factors (VIF/VDF), effective sample size differing from n, and autocorrelation estimation (denoted by $\hat{\rho}$). Of note is that the notion of a VIF frequently is encountered when studying linear regression.

The most popular specification of the SAR model is based upon a row-standardized geographic weights matrix. This specification casts each locationally tagged attributed value, y_i, as a function of the weighted average of its nearby, surrounding attribute values, namely $\rho_Y \sum_{j=1}^{n} w_{ij} y_j$ $(0 \le w_{ij} \le 1; \sum_{j=1}^{n} w_{ij} = 1)$. This specification also conceptualizes spatial autocorrelation as being contained in the regression model error term. And, theoretically this specification relates to the Bessell function semivariogram model in geostatistics (Griffith and Layne, 1999, Ch. 3). The equational form of the SAR model is, using matrix notation,

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{ß} + (\mathbf{I} - \rho_Y \mathbf{W})^{-1}\boldsymbol{\varepsilon}\ ,$$

where **Y** is an n-by-1 vector for the dependent variable (i.e., the regressand), **Z** is an n-by-(p+1) matrix of predictor variables (i.e., the regressors), **ß** is a (p+1)-by-1 vector of regression parameters, **ε** is an n-by-1 vector of *iid* N(0, σ_ε^2) error terms, **W** is an n-by-n row-standardized geographic weights matrix—often converted from a binary, 0-1 contiguity matrix **C** based upon "rook's" adjacencies (drawing an analogy with chess moves)—and ρ_Y is the spatial autoregressive parameter for variable Y. In general, though, let the spatial covariance matrix be denoted by $\mathbf{V}^{-1}\sigma^2$, which for this SAR model yields $\mathbf{V}^{-1} = (\mathbf{I} - \rho_Y\mathbf{W})^T(\mathbf{I} - \rho_Y\mathbf{W})$.

Now, consider the sampling distribution of the mean, $\bar{y}$. For *iid* data, the variance of this sampling distribution is given by $\frac{\sigma^2}{n}$. In contrast, for georeferenced data this variance is given by $\frac{\mathbf{1}^T\mathbf{V}^{-1}\mathbf{1}}{n^2}\sigma^2$. Of note is that for *iid* data, $\mathbf{V}^{-1} = \mathbf{I}$ and this latter expression reduces to $\frac{\sigma^2}{n}$, whereas if spatial autocorrelation equals 1, conceptually speaking this latter expression reduces to $\frac{\sigma^2}{1}$. Meanwhile, positive spatial autocorrelation causes an inflation of the variance. The accompanying variance inflation factor (VIF) is given by $\frac{TR(\mathbf{V}^{-1})}{n}$, where *TR* denotes the matrix trace operator. Combining this result with the conventional adjustment for variance, and taking into account the estimation of ρ_Y, renders $\dfrac{TR(\hat{\mathbf{V}}^{-1}) - \dfrac{\mathbf{1}^T\hat{\mathbf{V}}^{-1}\mathbf{1}}{n}}{n-2}$ as the VIF estimator.

Determining the effective sample size for this univariate case requires the following algebraic manipulations:

$$\frac{\sigma_\varepsilon^2}{\dfrac{n^2}{\mathbf{1}^T\mathbf{V}^{-1}\mathbf{1}}} = \frac{TR(\mathbf{V}^{-1})\sigma_\varepsilon^2}{TR(\mathbf{V}^{-1})\dfrac{n^2}{\mathbf{1}^T\mathbf{V}^{-1}\mathbf{1}}} = \frac{\dfrac{TR(\mathbf{V}^{-1})}{n}\sigma_\varepsilon^2}{n\dfrac{TR(\mathbf{V}^{-1})}{\mathbf{1}^T\mathbf{V}^{-1}\mathbf{1}}}\ .$$

Hence the effective sample size is given by n* = $n\frac{TR(\mathbf{V}^{-1})}{\mathbf{1}^T\mathbf{V}^{-1}\mathbf{1}}$. Of note is that if spatial autocorrelation is zero, this expression reduces to n; if spatial autocorrelation is 1, conceptually this expression reduces to 1.

Consequently, when computing the t-statistic for the mean of a georeferenced variable, where the spatial autoregressive parameter ρ_Y has been estimated, the effective degrees of freedom used should be df* = $n\frac{TR(\mathbf{V}^{-1})}{\mathbf{1}^T\mathbf{V}^{-1}\mathbf{1}}$ - 2, where 1 df is lost for estimation of the sample standard deviation, s, and 1 df is lost for estimation of the autoregressive parameter, $\hat{\rho}_Y$.

Calculating the effective sample size for a univariate georeferenced data analysis involves matrix inversion. But the relationship between this quantity and the SAR spatial autocorrelation parameter exhibits a very strong and very precise nonlinear trend, which is portrayed in Figure 3. This graph was constructed from both empirical data results and simulation experiment results; in Figure 3 the empirical relationship is denoted by o's (o), the simulation relationship is denoted by dots (.), and the predicted values are denoted by pluses (+). The simulation results follow a smooth curve, whereas the empirical data results slightly scatter about this smooth curve.

Figure 3 implies the following equation:

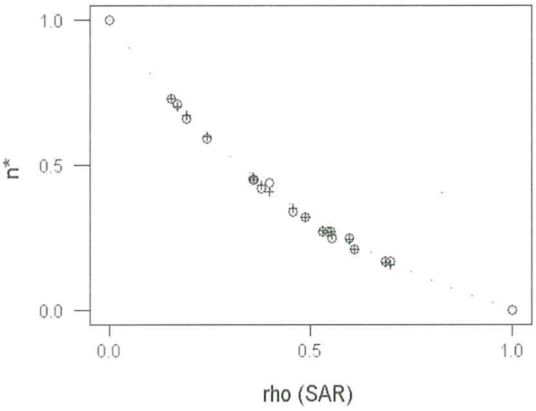

Figure 3. Predicting n* from n and rho

$$\hat{n}* = n \times [1 - 2.67978 \frac{n-1}{n} (1 - e^{-0.74555\hat{\rho}(1-0.37352\hat{\rho})})] . \quad (3)$$

For the empirical cases used to construct Figure 3, the pseudo-$R^2 = 0.998$. The specification of this equation has been formulated in order to insure that when $\hat{\rho} = 0$ then $\hat{n}* = n$, and when $\hat{\rho} = 1$ then $\hat{n}* = 1$.

An Example: Haining's Georeferenced Glasgow Data

Standard mortality rates for both "all deaths" and "cerebro-vascular disease" appear to conform to a normal frequency distribution in the population; their respective S-W statistics are 0.97771 (prob = 0.46) and 0.97296 (prob = 0.26). Three of the remaining standardized rates require a power transformation:

LN(accidents + 89): S-W = 0.98615 (prob = 0.85)
LN(respiratory + 32): S-W = 0.97398 (prob = 0.30)
LN(cancer – 23): S-W = 0.97869 (prob = 0.50)

Meanwhile, the standardized mortality rate for "ischaemic heart disease" has one extremely high and one extremely low rate, with these two outliers causing deviation from normality. Once these two anomalies are accounted for, the frequency distribution conforms acceptably well to a bell-shaped curve [i.e., S-W = 0.98178 (prob = 0.66)].

To simplify comparisons here, all variables have been converted to z-scores, resulting in $s_Y^2 = 1$. For Glasgow, MC_{max}= 1.07829; from equation (2), $\hat{MC}_{max} = 1.08247$. Univariate analysis results are summarized in Table 2.

Multiplying an error variance, s_e^2, by its corresponding theoretical VIF results in a variance value for the associated georeferenced variable of approximately 1 (e.g., 0.59889× 1.65020 = 0.98829). Clearly, using the values for $\hat{\hat{\rho}}$ rendered by equation (1), which almost equal their $\hat{\rho}$ counterparts, is far better than using ρ= 0. Similarly, using the values for $\hat{n}*$ rendered by equation (3), which also deviate little from their n* counterparts, is far better than using n = 87.

IV. THE EFFECTIVE SAMPLE SIZE FOR A PAIR OF GEOREFERENCED VARIABLES

An even more popular statistical test is the t-test performed on bivariate correlation coefficients. Consider the sampling distribution of the bivariate correlation coefficient, r. For *iid* sample data, the variance of this sampling distribution is given by $\frac{1}{n-2}$. In contrast, for georeferenced data this variance may be approximated with a simulation experiment.

As with the univariate case, the covariation for two georeferenced variables also has a VIF associated with it. In its simplest form, this VIF is given by $\frac{TR[(\mathbf{V}_X^{-1/2})^T(\mathbf{V}_Y^{-1/2})]}{n}$; in its sample statistics form, this VIF is given by $\frac{TR[(\hat{\mathbf{V}}_X^{-1/2})^T\hat{\mathbf{V}}_Y^{-1/2}]}{n-2} - \frac{\mathbf{1}^T(\hat{\mathbf{V}}_X^{-1/2})^T(\hat{\mathbf{V}}_Y^{-1/2})\mathbf{1}}{n(n-2)}$, for which both the two means and the two spatial autoregressive parameters have been estimated. But when this VIF is divided by the VIFs for the variances of X and Y, it becomes a variance deflation faction (VDF). In its simplest form, for example, it becomes $\frac{TR[(\mathbf{V}_X^{-1/2})^T\mathbf{V}_Y^{-1/2}]}{\sqrt{TR(\mathbf{V}_X^{-1})\times TR(\mathbf{V}_Y^{-1})}}$. In other words, the covariation inflation for two georeferenced variables tends to be more than compensated for by the product of the individual VIFs, yielding a VDF, which is sensible since a correlation coefficient cannot exceed 1. In practice this VDF tends to be close to 1.

As often is the case, the major impact of spatial

Table 2. Univariate results for Haining's Glasgow data

variable	$\hat{\rho}$	MC	$\hat{\hat{\rho}}$	VIF	s_e^2	n*	$\hat{n}*$
All deaths	0.70108	0.42670	0.68079	1.65020	0.59889	12.9	13.2
Accidents	0.51561	0.29308	0.51727	1.24176	0.79733	25.0	25.5
Respiratory	0.54626	0.30537	0.53427	1.28423	0.76515	22.7	23.2
Cancer	0.63716	0.39776	0.64935	1.45876	0.67531	16.6	17.0
Heart disease	0.50405	0.23218	0.42742	1.22733	0.81576	25.9	26.4
Cerebro-vascular	0.35573	0.17466	0.33465	1.09805	0.92260	39.1	39.7
Social class	0.70581	0.46609	0.72013	1.66780	0.56334	12.6	13.0

autocorrelation with regard to a correlation coefficient is on the variance of its sampling distribution. Suppose the variance of the sampling distribution of r is given by σ_r^2. Then given that $\sigma_r^2 = \frac{1}{n-2}$, the effective sample size, say n^*_{XY}, may be defined as $n^*_{XY} = \frac{1}{\sigma_r^2} + 2 = \sigma_r^{-2} + 2$ This effective sample size can be approximated by simulating the sampling distribution, such that $\hat{n}^*_{XY} = \hat{\sigma}_r^{-2} + 2$. The steps of the necessary simulation experiment used to calculate $\hat{\sigma}_r^2$ may be summarized as follows:

> Begin an experiment by generating two n-by-1 vectors of *iid* N(0,1) pseudo-random variables, one for variable X and one for variable Y; respectively denote these pseudo-random variables by, say, ε_X and ε_Y. Next, embed spatial autocorrelation into both X and Y as follows:
>
> $\mathbf{X} = (\mathbf{I} - \rho_X\mathbf{W})^{-1}\varepsilon_X$ and $\mathbf{Y} = (\mathbf{I} - \rho_Y\mathbf{W})^{-1}\varepsilon_Y$,
>
> where ρ_X and ρ_Y respectively denote prespecified levels of spatial autocorrelation (ranging from 0 to 1 for positive spatial autocorrelation). Third, calculate the correlation coefficient for X and Y. Fourth, replicate this procedure several thousand times, keeping ρ_X and ρ_Y the same across replications, and repeating these sets of replications for a systematic sample of ρ_j values from $(\rho_{min}, 1)$. Fifth, calculate the variance of r, say s_r^2. Finally, calculate the approximate effective sample size as $n^*_{XY} = \frac{1}{s_r^2} + 2 = s_r^{-2} + 2$.

The resulting equation should reduce to $\frac{1}{n-2}$ when $\rho_X = \rho_Y = 0$, and should increase from this value as both ρ_X and ρ_Y increase.

Calculating the effective sample size for a bivariate georeferenced data analysis involves matrix inversion coupled with extensive simulation work. The relationship between this measure and each of the SAR spatial autocorrelation parameters exhibits a very strong and very precise nonlinear trend, which is portrayed in Figure 4. This graph was constructed from a series of simulation experiments employing the geographic weights matrix for the Glasgow geographic landscape; the simulated relationship with regard to ρ_X is denoted by o's (o), whereas the simulated relationship with regard to $\rho_Y$is denoted by pluses (+). The simulation results follow the same scatterplot for variables X and Y, which is sensible. The horizontal alignment of points in the upper right-hand quadrant of the graph are for cases where only one of the autocorrelation parameters equals 0; the cluster of points on the right-hand part of the graph are for cases where $\rho_X \neq \rho_Y$. Of note is that the range of negative spatial autocorrelation is governed by the smallest eigenvalue of the geographic weights matrix (i.e., $\frac{1}{-0.63602} = -1.57228$ for the Glasgow surface partitioning).

Figure 4 implies the following equation:

$$\hat{n}^*_{XY} = n \times \frac{1 - \dfrac{\dfrac{\rho_X\rho_Y}{|\lambda_{extreme}|\times|\lambda_{extreme}|}}{\dfrac{n}{n-3.5}}}{1 + \dfrac{\dfrac{\rho_X\rho_Y}{|\lambda_{extreme}|\times|\lambda_{extreme}|}}{6}} - 2$$

$$= 1 + (n-3) \times \frac{1 - \dfrac{\rho_X\rho_Y}{|\lambda_{extreme}|\times|\lambda_{extreme}|}}{1 + \dfrac{\rho_X\rho_Y}{6\times|\lambda_{extreme}|\times|\lambda_{extreme}|}}, \qquad (4)$$

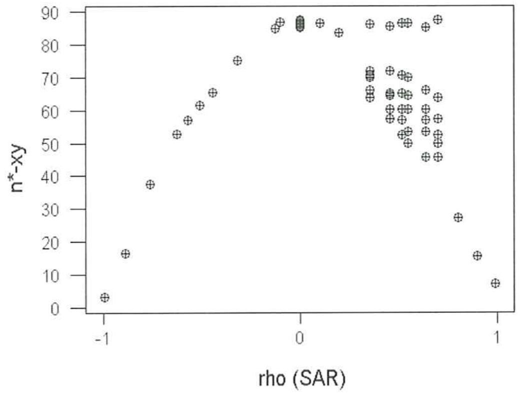

Figure 4. Predicting n*-xy from rho-x and rho-y

where the extreme eigenvalues are λ_{max} for positive spatial autocorrelation and λ_{min} for negative spatial autocorrelation, and are extracted from the row-standardized geographic weights matrix **W**. Because matrix **W** is row-standardized, λ_{max} is theoretically known to always equal 1; hence for the very common case of both X and Y containing positive spatial autocorrelation, the equation reduces to $\hat{n}^*_{XY} = 1 + (\text{n-3}) \times \frac{1-\rho_X\rho_Y}{1+\frac{\rho_X\rho_Y}{6}}$. This equation is similar to that known for time series (see Haining, 1990, p. 314), and that reported by Richardson and Hémon (1982). Two noteworthy differences are the appearance of λ_{min} in the case of negative spatial autocorrelation, and the appearance of 6 in the denominator. This value of 6 may be specific to the Glasgow medical districts surface partitioning; further research is need to confirm or revise this finding. The simulation experiments used to construct Figure 4 also were used to evaluate the equation for $\hat{n}^*_{XY}$; the accompanying scatterplot appears in Figure 5, for which pseudo-R^2 = 0.996.

An Example: Haining's Data Revisited

Again using the Glasgow geographic landscape, consider the "cancer mortality rates" variable together with the "percentage of households living in poor accommodation lacking basic amenities" (a "social class" surrogate). Let this first variable be Y and this second variable be X. This first variable can be transformed in order to obtain an acceptable S-W statistic using *LN*(Y – 23). Unfortunately, a power transformation does not exist for variable X that yields an acceptable S-W statistic. For purposes of analysis, the logarithmic transformation *LN*(X + 0.01) has been employed here.

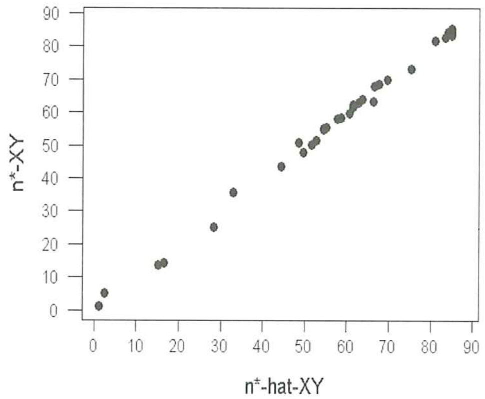

Figure 5. Relationship between n*-XY and n*-hat-XY

First, the correlation coefficient between the two selected transformed georeferenced variables is 0. 206, which is slightly less than the r = 0.245 value reported by Haining for Y with *LN*(X + 1). The VDF here is 0.99722, which essentially is 1. A simulation experiment for this case—for which ρ_X = 0.70581 andρ_Y= 0.63716, and involving 5,000 replications—yields$\hat{\sigma}_r$ = 0.15391 for the autocorrelated data and 0.10814 for the unautocorrelated data. This second value compares very favorably with its theoretical counterpart of 0.10847. Hence, $\hat{n}^*_{XY} = \frac{1}{0.15391^2} + 2 = 44.2$.

Now the corresponding significance tests here may be written as follows, with regard to spatial autocorrelation acknowledgement:

ignoring: $\frac{0.206-0}{0.10847} = 1.89914 < t_{85,0.975} = 1.9883$,

accounting for: $\frac{0.206-0}{0.15391} = 1.33844 << t_{42.2,0.975} = 2.0178$.

Therefore, taking spatial autocorrelation into account strongly implies that there is no relationship between Glasgow "cancer mortality rates" and "percentage of households living in poor accommodation lacking basic amenities" *in the population.*

V. IMPLICATIONS FOR FUTURE RESEARCH

Therefore, four research problems are identified in this paper that need to be definitively solved before a richer toolbox of spatial statistical techniques can be relatively easily implemented in a GIS. The complete solutions will involve both empirical assessments and simulation experiments. These four problems are represented by the four principal equations posited in this paper.

Equation (1) describes the relationship between the SAR spatial autocorrelation parameter computed using a row-standardized binary geographic weights matrix, and the MC for a given georeferenced data set that is normally distributed. But does this relationship persist when the geographic weights matrix is based upon inter-point distance, or some other measure of nearness? Does this relationship persist when the underlying frequency distribution is non-normal (e.g., uniform, exponential, or sinusoidal)? Does this relationship persist if the autoregressive response (AR) model, a popular specification in spatial econometrics, is employed? And, how does this relationship change when the conditional autoregressive

(CAR) model, a popular specification in image analysis, is employed?

Equation (2) allows the maximum MC value to be predicted. It needs to be more thoroughly evaluated with regard to a much larger set of empirical surface partitionings, including those associated with hexagonal tessellations. It also needs to be more thoroughly evaluated through a wider range of simulation experiments. Evidence reported here suggests equation (2) holds considerable promise for implementation purposes.

Equation (3) describes the effective sample size for a single georeferenced variable. Considerable evidence has been accumulated in support of its specification. Additional assessment should be in terms of a wider range of simulation experiments, including ones that involve hexagonal tessellations. It also needs to be related to equation (4).

Equation (4) describes the effective sample size for a pair of georeferenced variables. This equation also needs to be subjected to a more thorough assessment, primarily using simulation experiments. Is the factor of 6 appearing in it universal? Or does this value somehow relate to the minimum eigenvalue of the associated matrix **C**, or to the number of areal units involved? In addition, this result needs to be assessed within the context of differing natures of spatial autocorrelation (i.e., $\rho_X > 0$ and $\rho_Y < 0$).

In conclusion, equations (1)-(4) offer considerable computational simplification for the implementation of spatial statistical techniques within a GIS. In fact, sufficient evidence in support of them is presented in this paper to allow their implementation at this time on an experimental basis. These equations can be easily implemented in a desktop GIS system like ArcView, as is illustrated in this paper. The same implementation also can be easily replicated in other GIS systems, such as Arc/Info using AML or VBA (which requires Arc/Info 8), or even in a desktop DBMS, such as Access (which requires interoperability with GIS components). They remove the need for eigenfunction and nonlinear optimization routines, and maintain the standard linear regression technique as the workhorse of a GIS statistical analysis. Finally, they strengthen the inferential basis for a spatial scientist. These equations clearly are worthy of the subsequent attention needed to confirm their respective utilities.

REFERENCES

[1] Chatelin, F., 1993, *Eigenvalues of Matrices* (translated by W. Ledermann), NY: Wiley.

[2] Griffith, D., 1988, *Advanced Spatial Statistics*, Dordrecht: Martinus Nijhoff.

[3] Griffith, D., and C. Amrhein, 1997, *Multivariate Statistical Analysis for Geographers*, Englewood Cliffs, NJ: Prentice Hall.

[4] Griffith, D., and L. Layne, 1999, *A Casebook for Spatial Statistical Data Analysis*, NY: Oxford University Press.

[5] Griffith, D., J. Paelinck, and R. van Gastle, 1998, The Box-Cox transformation: computational and interpretation features of the parameters, in D. Griffith and C. Amrhein, *Advances in Spatial Modelling and Methodology: Essays in Honor of Jean Paelinck*, Dordrecht: Kluwer, pp. 45-56.

[6] Haining, R., 1990, *Spatial Data Analysis is the Social and Environmental Sciences*, NY: Cambridge University Press.

[7] Haining, R., 1991, Bivariate correlation with spatial data, *Geographical Analysis*, 23: 210-227.

[8] Richardson, S., and D. Hémon, 1982, On the variance of the sample correlation between two independent lattice processes, *J. of Applied Probability*, 18: 943-948.

[9] Zhang, Z. and D. Griffith, 1997, Developing user-friendly spatial statistical analysis modules for GIS: an example using ArcView, *Computers, Environment and Urban Systems*, 21: 5-29.

[10] Zhang, Z., and D. Griffith, 2000, Integrating GIS components and spatial statistical analysis in DBMS, *International J. of Geographical Information Science*, forthcoming.