Griffith, Daniel A.

Article

Selected Challenges from Spatial Statistics for Spatial Econometricians

Comparative Economic Research. Central and Eastern Europe

Provided in Cooperation with:
Institute of Economics, University of Łódź

Suggested Citation: Griffith, Daniel A. (2012) : Selected Challenges from Spatial Statistics for Spatial Econometricians, Comparative Economic Research. Central and Eastern Europe, ISSN 2082-6737, Łódź University Press, Łódź, Vol. 15, Iss. 4, pp. 71-85, https://doi.org/10.2478/v10103-012-0027-5

This Version is available at:
https://hdl.handle.net/10419/259119

10.2478/v10103-012-0027-5

DANIEL A. GRIFFITH*

Selected Challenges from Spatial Statistics for Spatial Econometricians

Abstract

Griffith and Paelinck (2011) present selected non-standard spatial statistics and spatial econometrics topics that address issues associated with spatial econometric methodology. This paper addresses the following challenges posed by spatial autocorrelation alluded to and/or derived from the spatial statistics topics of this book: the Gaussian random variable Jacobian term for massive datasets; topological features of georeferenced data; eigenvector spatial filtering-based georeferenced data generating mechanisms; and, interpreting random effects.

1. Introduction

Geography experienced a quantitative revolution in the 1950s and 1960s (Curry 1967). Work generated by this movement initially analyzed distances from locations of privilege as well as attribute variables whose observations were distinguished merely by a locational index. Especially statistical decisions spawned by these analyses proved to display far more variability than indicated by classical statistical distribution theory; this increased variability is attributable to positive spatial autocorrelation (SA) latent in almost all georeferenced data. Addressing these inadequacies, Cliff and Ord[1] (1969) and Besag (1974), among

* Ph.D., University of Texas at Dallas

[1] *Geographical Analysis* celebrated the major contributions to science of this cluster of research with a special issue in 2009.

others, commenced a formal development of autoregression-based spatial statistics that popularized model specifications accounting for latent SA. This line of work soon eclipsed the point pattern analysis work that, until then, typified much of quantitative spatial analyses. Meanwhile, parallel spatial econometric developments flourished after the introduction of Paelinck and Klaassen's (1979) seminal book, followed by Anselin's (1988) classic book. Paelinck (2012) addresses this historical trajectory.

The purpose of this paper is to highlight selected challenges posed by SA alluded to and/or derived from the spatial statistics literature and contextualized in Griffith and Paelinck (2011). One challenge arises from the increasing size of georeferenced datasets, some of which are massive today. Calculating maximum likelihood estimates (MLEs) requires computing the determinant of an n-by-n spatial covariance matrix—the Jacobian of a transformation in calculus terms—which becomes excessively numerically intensive or even infeasible for massive georeferenced datasets. This paper outlines an alternative MLE solution to nonlinear regression, which is new, couched in the existing spatial statistics literature about approximating the Jacobian term. A second challenge stems from topological considerations accompanying georeferenced datasets. This paper focuses on a mistake appearing in the earlier literature, and describes a modified version of the well-known matrix powering algorithm that successfully computes the principal eigenfunction for a periodic matrix. A third challenge concerns georeferenced data generating mechanisms involving eigenvector spatial filtering, and further develops contributions in Griffith (2011a,b). A fourth challenge furnishes additional insight into the meaning of spatially structured random effects. Successful engagement of these challenges poses a potential to improve both spatial statistical and spatial econometric work.

2. The spatial statistical Jacobian term for Gaussian model specifications

In part because normal curve theory was the best developed probability model-based analysis of the time, most early spatial statistics assumed a bell-shaped curve. Gaussian spatial autoregressive model specifications to describe n georeferenced sample values include a Jacobian term, which is: (1) the determinant of an n-by-n matrix; (2) the normalizing constant ensuring that the probability density function integrates to 1; and, (3) a function of the SA parameter(s). Computational difficulties introduced into calculating MLEs of model parameters by the logarithm of this determinant has generated a body of literature addressing its simplification and approximation (Ord, 1975; Griffith,

1992, 2004a; Barry and Pace, 1999; Smirnov and Anselin, 2001, 2009; Pace and LeSage, 2004; Zhang and Leithead, 2007; Walde et al., 2008).

The likelihood function is equivalent to a multivariate normal probability density function with a sample of size 1 and n variables:

$$L = (2\pi)^{-n/2} \mid \mathbf{V} \mid^{1/2} (\sigma^2)^{-n/2} e^{-(\mathbf{Y}-\mu\mathbf{1})^T \mathbf{V} (\mathbf{Y}-\mu\mathbf{1})/(2\sigma^2)} \quad (1)$$

where $\mathbf{V}^{-1}\sigma^2$ is the SA variance-covariance matrix that is a function of the spatial autoregressive parameter ρ in a single-parameter model specification, Y is a normally distributed random variable, **Y** is an n-by-1 vector of random variable values, **1** is an n-by-1 vector of ones, T denotes the matrix transpose operation, and μ and σ^2 respectively are the constant mean and the variance of Y. When $\rho = 0$, $\mathbf{V} = \mathbf{I}$, the n-by-n identity matrix.

But all of the more recent literature overlooks the useful simplicity of the approximation developed by Griffith (1992, 2004a), with a special case for regular square tessellations (Griffith, 2004). The appeal of this latter approximation is that it can be employed efficiently and effectively with a dataset whose size is in the millions or billions—a massive dataset. For a symmetric distribution of eigenvalues, such as that for a regular square tessellation, the Jacobian approximation given by Griffith (2004) reduces to

$$-\sum_{i=1}^{n} LN(1-\rho\lambda_j)/n \approx 2\omega LN(\delta) - \omega LN(\delta+\rho) - \omega LN(\delta-\rho) \quad (2)$$

where λ_j is the j^{th} eigenvalue of matrix **V**, and ω and δ are coefficients to be calibrated. When $\rho = 0$, both $2\omega LN(\delta) - \omega LN(\delta+\rho) - \omega LN(\delta-\rho) = 0$ and $-\sum_{i=1}^{n} LN(1-\rho\lambda_j) = 0$. Griffith (2004b) shows that the Jacobian term associated with a regular square tessellation forming a complete rectangular region also can be approximated by

$$-\sum_{i=1}^{n} LN(1-\rho\lambda_j)/n \approx LN(1+q_2\rho^2+q_4\rho^4+q_{20}\rho^{20}) \quad (3)$$

where q_2, q_4, and q_{20} are coefficients to be calibrated. When $\rho = 0$, $LN(1+q_2\rho^2+q_4\rho^4+q_{20}\rho^{20}) = 0$.

MLEs for the three parameters of equation (1) are

$$\hat{\mu} = \mathbf{1}^T\mathbf{V}\mathbf{Y}/(\mathbf{1}^T\mathbf{V}\mathbf{1}), \tag{4}$$

$$\hat{\sigma}^2 = (\mathbf{Y} - \mu\mathbf{1})^T\mathbf{V}(\mathbf{Y} - \mu\mathbf{1})/n, \tag{5}$$

and for a spatial simultaneous autoregressive (SAR; the spatial error model in the spatial econometrics literature) model specification, for which $\mathbf{V} = (\mathbf{I} - \rho\,\mathbf{W})^T(\mathbf{I} - \rho\,\mathbf{W})$, where $\mathbf{W}$ is an n-by-n geographic weights matrix, ρ may be calculated by solving the differential equation

$$\partial LN(L)/\partial\rho = \left[\sum_{j=1}^{n}(-\lambda_j)/(1-\rho\,\lambda_j)\right]\Big/ n + (\mathbf{Y} - \hat{\mu}\mathbf{1})^T(\mathbf{W}^T + \mathbf{W} - \rho 2\mathbf{W}^T\mathbf{W})(\mathbf{Y} - \hat{\mu}\mathbf{1})/(2n\,\hat{\sigma}^2) = 0 \tag{6}$$

where λ_j are the n eigenvalues of matrix $\mathbf{W}$.

Equations (2) and (3) respectively result in $\left[\sum_{j=1}^{n}(-\lambda_j)/(1-\rho\,\lambda_j)\right]\Big/ n \approx$

$-2\,\alpha\,\rho/(\delta^2 - \rho^2)$, and

$-(2q_2\rho + 4q_4\rho^3 + 20q_{20}\rho^{19})/(1 + q_2\rho^2 + q_4\rho^4 + q_{20}\rho^{20})$.

These two substitutions dramatically simplify equation (6).

For a regular square tessellation forming a complete rectangular region (i.e., a remotely sensed image whose data may be important for an environmental economics analysis), with P > 3, Q > 3, and PQ ≤ 5,625, numerical experiments yield the following large sample results:

$\omega \approx 0.16361 - 0.00457(1/P + 1/Q) - 0.47594/(PQ)$

$\delta \approx 1.17583 - 0.33691[1/(P+1) + 1/(Q+1)] - 1.08316/[(P+1)(Q+1)]$, and

$q_2 \approx 0.11735 + 0.10091(1/P^{5/4} + 1/Q^{5/4}) + 0.42844/(PQ)$

$q_4 \approx 0.07421 + 0.05730(1/P^{2/3} + 1/Q^{2/3}) - 0.66001/(PQ)$

$q_{20} \approx 0.05221 + 0.52467(1/P^{7/4} + 1/Q^{7/4}) + 2.48015/(PQ)$.

The computation of Table 1 results utilized these numerical generalizations for a massive 3,000-by-5,000 pixels georeferenced dataset

collected for the Florida Everglades[2]. Relatively few computational resources are needed to analyze n = 15,000,000 observations in this case.

A challenge for spatial econometricians suggested by this spatial statistical work is the generalization of coefficients for equation (2) for massive georeferenced datasets based upon the type of irregular surface partitioning that characterize administrative units. The popular autoregressive response (AR; the spatial error model in the spatial econometrics literature) model specification has the following MLEs:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}(\mathbf{I} - \rho\,\mathbf{W})\mathbf{Y} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} - \rho\,(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{Y}$$

$$\hat{\sigma}^2 = [(\mathbf{I} - \rho\,\mathbf{W})\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}]^{\mathrm{T}}[(\mathbf{I} - \rho\,\mathbf{W})\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}]/\mathrm{n}\,, \text{ and}$$

$$\left[\sum_{\mathrm{j}=1}^{\mathrm{n}} (-\lambda_{\mathrm{j}})/(1 - \rho\,\lambda_{\mathrm{j}})\right] \Big/ \mathrm{n} +$$

$$[\mathbf{Y}^{\mathrm{T}}(\mathbf{W}^{\mathrm{T}} + \mathbf{W} - \rho 2\mathbf{W}^{\mathrm{T}}\mathbf{W})\mathbf{Y} - 2\mathbf{Y}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}})/(2\mathrm{n}\,\hat{\sigma}^2) = 0\,.$$

Because $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ needs to be inverted only once, this model specification involves relatively little increase in computational intensity vis-à-vis the constant mean case. Consequently, timing results appearing in Table 1 remain informative for the nonconstant mean case.

3. The topology of georeferenced data

SA and autoregression analyses frequently articulate the topological structure of georeferenced data with a simple binary 0-1 n-by-n geographic weights matrix **C** based upon connectivity/contiguity. The row and column labels of matrix **C** are the ordered locations in a geographic landscape, with this ordering being the same for both the rows and the columns for the sake of convenience. The common definitions of contiguity for surface partitioning are based upon analogies with chess moves: the rook when non-zero length common boundaries, and the queen when both zero (i.e., points) and non-zero length common boundaries, determine contiguity. If a row and a column location are contiguous, then the corresponding matrix cell is coded 1; otherwise, it is coded

[2] A January 1, 2002, 28.5-meter resolution LANDSAT 7 Enhanced Thematic Mapper Plus (ETM+) image.

0. Consequently, matrix **C** is sparse and symmetric. Often the preceding matrix **W** is a row-standardized version of this matrix **C**.

Table 1. Spatial autocorrelation parameter estimation

band	estimated eigenvalues				coefficient equations			
	equation (2)		equation (3)		equation (2)		equation (3)	
	$\hat{\rho}$	CPU time[1]	$\hat{\rho}$	CPU time[1]	$\hat{\rho}$	CPU time[1]	$\hat{\rho}$	CPU time[1]
1	0.9248	0.03	0.9236	0.04	0.9249	0.03	0.9231	0.01
2	0.9072	0.01	0.9097	0.03	0.9073	0.01	0.9094	0.04
3	0.8452	0.03	0.8621	0.04	0.8454	0.03	0.8639	0.03
4	0.4060	0.03	0.3990	0.03	0.4067	0.03	0.3991	0.01
5	0.4773	0.01	0.4672	0.03	0.4776	0.03	0.4683	0.03
7	0.6299	0.01	0.6255	0.03	0.6302	0.03	0.6288	0.03

Notes: [1] measured in seconds.

Source: own calculations.

Consider the case where matrix **C** is irreducible (i.e., it cannot be permuted into disjoint block diagonal submatrices). Powers of matrix **C** can be interpreted as follows (Maćkiewicz and Ratajczak 1996):

> Matrix $\mathbf{C}^k$ yields a count in cell c_{ij} that indicates the number of ways of moving from row location i to column location j crossing exactly k boundaries.

This combinatorial interpretation motivated the use of the row sums of a power sum of matrix **C**, standardized by the largest element in the resulting n-by-1 vector, say $\mathbf{E}_A$, as an index of topological accessibility for a network or surface partitioning. Relatively large values denote locations that are better connected (directly and indirectly) and more centrally located (i.e., more accessible) within the topology represented by the graph associated with matrix **C**, whereas relatively small values denote topologically peripheral locations within the graph (frequently those positioned on the boarder of the associated geographic landscape). The diameter of the graph counterpart of matrix **C** (i.e., the maximum number of links to be crossed when moving from any of the n nodes—areal units in the case of spatial analysis—to reached any of the other nodes) is a common stopping exponent for the power sum. Paths between nodes become redundant beyond the diameter, but do account for some detail in terms of topological structure. All entries in the summation matrix for this exponent have non-zero entries.

Vector $\mathbf{E}_A$ relates to the principal eigenvector of matrix **C**, say $\mathbf{E}_1$, and tends to converge upon it; Maćkiewicz and Ratajczak (1996, p. 78) argue that

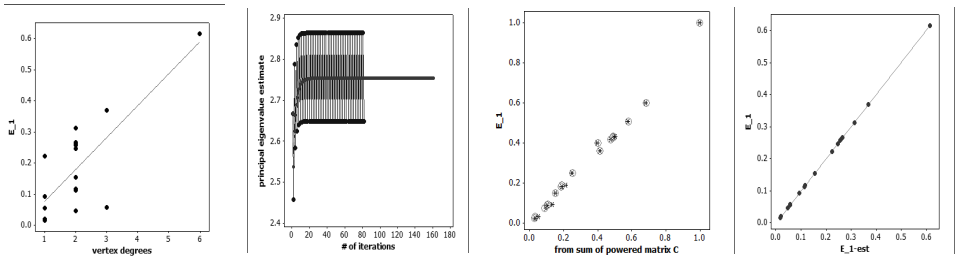
computing vector $\mathbf{E}_1$ is the proper way to define topological accessibility. Mass (1985) criticizes some of the earlier discussion by geographers concerning this accessibility index, challenging conjectures about the relationship between the principal eigenfunction and the row/column sums of matrix **C** (Figure 1a). Cvetković and Rowlinson (1990) echo Maas' discussion, but without contributing to resolving the controversy. Mass employs the geographic connectivity matrix for the 1929 Uganda road network reported by Gould (1967, p. 67). But this graph is periodic (the well-know matrix powering algorithm oscillates between 2.64892 and 2.86332; Figure 1b), and hence Maas reports incorrect eigenvalues, having obtained the solution for the lower bound in the oscillation (he reports only the first 10 of the 18 eigenvalues):

Table 1a. Maas' and actual eigenvalues

eigenvalue	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}
from Maas	2.652	1.828	1.618	1.447	0.9170	0.7035	0.6180	0.4419	0	0
actual	2.754	1.839	1.639	1.414	1.0000	0.8718	0.6787	0.3525	0	0

Source: Mass C. (1985).

Figure 1. Accessibility index scatterplots



Notes: Left (a): the principal eigenvector $\mathbf{E}_1$ versus the number of neighbors. Left middle (b): trajectories of the old and new algorithm estimates. Right middle (c): the principal eigenvector $\mathbf{E}_1$ versus the sum of the powers of matrix **C** through its diameter (open circle) and through 200 (*). Right (d): the principal eigenvector $\mathbf{E}_1$ versus its estimate produced by the new algorithm.

Source:own calculations.

The problematic periodicity of the 1929 Uganda road network graph can be resolved by modifying the well-know matrix powering algorithm to estimate the first eigenfunction so that it includes an iteration lag:

$$\lim_{k\to\infty}[\mathbf{1}^T(\mathbf{C}^k+\mathbf{C}^{k+1})\mathbf{1}/\mathbf{1}^T(\mathbf{C}^{k-1}+\mathbf{C}^k)\mathbf{1}]\to\lambda_1 .$$

This modification produces the proper convergences of the principal eigenvalue (Figure 1b) and its corresponding eigenvector (Figure 1d).

A challenge for spatial econometricians here is to determine the value of eigenvector $\mathbf{E}_1$ for empirical analyses. Is $\mathbf{E}_1$ a useful spatial analysis covariate? Is the correspondence between the row/column sums of matrix **C** sufficiently close to $\mathbf{E}_1$ that their vector is a useful spatial analysis covariate?

4. Georeferenced data generating mechanisms

Griffith and Paelinck (2011) present salient features of georeferenced data, including how variance inflation occurs through, and how correlation coefficients are impacted by, SA. They incorporate eigenvector spatial filters (ESFs) into georeferenced data generating mechanisms (i.e., a selected probability model that includes both random and SA components that combine together to yield individual observations) in some of their demonstrations.

Eigenvector spatial filtering methodology employs the eigenvectors extracted from a modified version of the geographic connectivity matrix **C**, namely $(\mathbf{I}-\mathbf{11}^{\mathrm{T}}/\mathrm{n})\,\mathbf{C}\,(\mathbf{I}-\mathbf{11}^{\mathrm{T}}/\mathrm{n}) = \mathbf{MCM}$, where **M** is the standard projection matrix commonly encountered in multivariate statistics, and ${}_{\mathbf{C}}\mathbf{E}_j$ is its j^{th} eigenvector. This matrix expression comes from the numerator of the Moran Coefficient (MC), whose matrix version for response vector **Y** adjusted only for its constant mean is given by

$$\mathrm{MC} = [\mathrm{n}/(\mathbf{1}^{\mathrm{T}}\mathbf{C1})][\mathbf{Y}^{\mathrm{T}}\mathbf{MCMY}/(\mathbf{Y}^{\mathrm{T}}\mathbf{MY})] .$$

Substituting the eigenvectors into this expression results in a Rayleigh quotient, with vector ${}_{\mathbf{C}}\mathbf{E}_1$ maximizing the expression. Accordingly, these n eigenvectors can be interpreted as follows:

> the first eigenvector, say ${}_{\mathbf{C}}\mathbf{E}_1$, is the set of real numbers that has the largest MC achievable by any set for the geographic arrangement defined by the spatial connectivity matrix **C**; the second eigenvector is the set of real numbers that has the largest achievable MC by any set that is orthogonal and uncorrelated with ${}_{\mathbf{C}}\mathbf{E}_1$; and so on through ${}_{\mathbf{C}}\mathbf{E}_n$, the set of real numbers that has the largest negative MC achievable by any set that is orthogonal and uncorrelated with the preceding (n – 1) eigenvectors.

As such, these eigenvectors furnish n distinct map pattern descriptions of latent SA in geographically distributed variables because they are mutually orthogonal and uncorrelated. An ESF is constructed from some linear combination of a subset of these eigenvectors, and serves as a spatial proxy

variable capturing SA effects in a model specification. This control variable embeds stochastic spatial dependencies among location-indexed observations into the parameters of a probability density/mass function.

All but one of the matrix **MCM** normalized eigenvectors have means of 0 and variances of 1/n. In other words, all of their 1st and 2nd moments match. Consider the following linear combination of these vectors:

$$W\sum_{j=1}^{K}(2q-1)\,a_j\mathbf{E}_j/\sqrt{\sum_{j=1}^{K}a_j^2}=W\mathbf{E}_K\mathbf{B}, \qquad (7)$$

where scaling coefficient W is some positive real number, binary 0-1 variable q is a Bernoulli RV, and a_j is a positive coefficient for eigenvector j. The term **B** describes the nature, whereas the scalar W describes the degree (i.e., the relative amount of variance accounted for), of SA. The mean of linear combination (7) is 0, whereas its variance is W^2/n, and the term (2q – 1) makes no difference because each eigenvector $\mathbf{E}_j$ is unique to a multiplicative factor of –1.

The central limit theorem governs expression (7), which implies that the ESFs described by it are approximately normally distributed.

Lemma 1: K << n eigenvectors of matrix **MCM** are independent and are not necessarily identically distributed RVs. As K goes to infinity, expression (7) converges on a normal distribution.

PF: Because all of the K means are 0, and

$\lim_{K\to\infty} a_j^2/\left(n\sum_{j=1}^{K}a_j^2\right)\to 0$ (Bentkus et al., 1996; Chaidee and Tuntapthai, 2009).

Simulation experiments furnish evidence corroborating this lemma (Table 2); it can be tested for in practice with a normal quantile plot. Therefore, an individual observation of a georeferenced Gaussian random variable may be written as

$$Y_i = Y_i^* + W\mathbf{e}_{iK}\mathbf{B}, \qquad (8)$$

where the n Y_i^* are iid $N(\mu, \sigma^2)$, and the $W\mathbf{e}_{iK}\mathbf{B}$ is an ESF.

Equation (8) comprises two normal components, one of which is equivalent to a spatially structured random mean response (i.e., $W\mathbf{e}_{iK}\mathbf{B}$ creates random deviations about μ), and furnishes the data generating mechanism in terms of the following parametric mixture distribution:

$$\left.\begin{array}{l} Y \mid SA \sim N(\mu, \sigma^2) \\ SA \sim N(0, W^2/n) \end{array}\right\} \Rightarrow Y \sim N(\mu, \sigma^2 + W^2/n).$$

In other words, the distribution at each location i is conditional on $W\mathbf{e}_{iK}\mathbf{B}$, and SA inflates variance while not affecting an average mean response across a map.

A Manly transformation[3] modifies the geographic distribution of population density across Poland (Figures 2a and 3a), based upon communes, so that it approximates a bell-shaped curve (Figure 2b). The ESF (a linear combination of 22 of 591 candidate vectors) for this geographic distribution closely conforms to be bell-shaped curve (Figures 2c and 3b), and accounts for roughly 27% of the geographic variation in the transformed population density. The random variable Y* approximates a normal distribution, deviating from a bell-shaped curve with one heavy tail (Figures 2d and 3c). The estimated mixture distribution is

$$\left.\begin{array}{l} Y \mid SA \sim N(0.93211, 0.03911^2) \\ SA \sim N(0, 0.02417^2) \end{array}\right\} \Rightarrow Y \sim N(0.93211, 0.04598^2).$$

The variance inflation factor is 1.38192; SA introduces an additional nearly 40% geographic variability into the transformed population density.

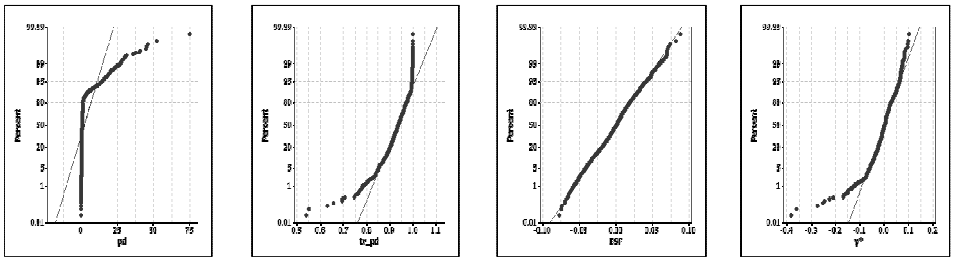
Table 2. Comparisons of eigenvectors and of ESFs: Poland surface partitionings

moments	communes	counties	viovodeships
n	2,468	369	16
mean	0	0	0
standard deviation	0.02013	0.05206	0.25000
skewness: mean	0.05806	0.09079	0.20614
standard deviation	0.20666	0.14942	0.21966
excess kurtosis: mean	0.63105	0.17750	-1.14832
standard deviation	1.07388	0.42391	0.33501
# eigenvectors with MC > 0.25	591	85	4
eigenvectors: % with Pr(normality) < 0.01	0	18.82	0
simulated ESFs: % with Pr(normality) < 0.01	0	16.62	0.62

Source: own calculations.

[3] The Manly transformation completes the family of Box-Cox power transformations. Here the empirically calibrated transformation is $\exp[-0.3319/(\text{population/area})^{0.7800}]$.

Figure 2. Normal quantile plots



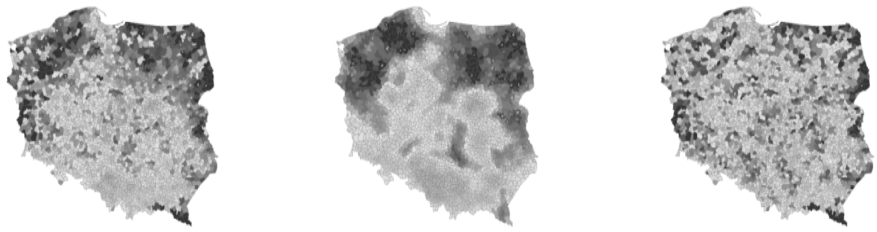
Notes: Left (a): population/area Middle left (b): transformed population density. Middle right (c): ESF. Right (d): Y*.

Source: own calculations.

One advantage of the ESF specification is that it very accurately captures spatial structure as reflected in map pattern (compare Figures 3a and 3b). Thus, ESF supports simulation experiments for which Y* can be determined with a pseudo-random number generator, and then added to the ESF, resulting in the spatial structure being held constant across simulation replications—the parametric mixture distribution specifies the geographic distribution of Y as being conditional on this map pattern.

One challenge for spatial econometricians suggested by this spatial statistical conceptualization is the establishment of ESF properties vis-à-vis spatial autoregression model specifications. Another is to fully develop the mathematical statistical theory associated with ESFs.

Figure 3. The geographic distribution of population density



Notes: scale from light gray to black is proportional to density. Left (a): population/area. Middle (b). ESF. Right (c): Y*.

Source: own calculations.

5. Explicating spatially structured random effects

Random effects model specifications address samples for which observations are selected in a highly structured rather than random way. Frequently random effects can be estimated in one of two ways: employing repeated measures in a frequentist context, or employing priors in a Bayesian context. This first conceptualization directly links random effects to means of time series for individual locations. Because the random effects term is a constant through time, its spatial structure can be captured by an ESF. In turn, this ESF can be estimated with only one slice of time (i.e., a map for a specified point in time), as in the preceding section, revealing that an ESF model specification is able to uncover at least part of a random effects term without repeated measures. Priors in a Bayesian analysis also allow the estimation of a random effects term with only one slice of time.

An average exists for each time series in a space-time dataset. This average ignores both spatial and serial correlation in the space-time series. A random effects model essentially works with these averages, adjusting them in accordance with the correlational structure latent in their parent space-time series, as well as their simultaneous estimation. The random effects model specification achieves this by fitting a distribution with a few parameters (e.g., a mean and a variance for a bell-shaped curve), rather than n individual means (fixed effects) for the n locations. Consequently, a relationship exists between the time series means and the random effects. This random effects specification relates to a fixed effects specification that includes n indicator variables, each for a separate district specific local intercept (one local intercept is arbitrarily set to 0 to eliminate perfect multicollinearity with the global mean).

A challenge for spatial econometricians suggested by this spatial statistical analysis concerns a need to better understand the number of degrees of freedom associated with a random effects term. Another challenge is to better understand random effects terms in the presence of covariates. Given that estimation of the spatially structured part of a random effects term is possible with a single map, a third challenge is to investigate whether or not estimation of the spatially unstructured component of a random effects term can be simplified.

6. Implications and conclusions

Spatial statistics and spatial econometrics are kindred spirits in terms of empirical analysis methodologies. Problems and challenges faced by one of

these fields reveals parallel problems and challenges already faced, or to be faced, by the other field. Griffith and Paelinck (2011) present a number of contemporary non-standard sources and treatments of such problems and challenges. This paper builds on their work, extending and identifying other problems and challenges. No doubt the future will produce new problems and challenges, too. Paelinck (2012) points out that spatial econometrics seeks to obtain a better understanding of the workings of spatial economies. Similarly, spatial statistics seeks to obtain a better understanding of the workings of geographic landscapes, some of which constitute space economies. This paper crystallizes the following challenges for spatial econometricians suggested by contemporary spatial statistical work: (1) formulating efficient and effective spatial autoregressive implementations for massive georeferenced datasets; (2) determining the utility of the principal eigenvector of a geographic weights matrix for empirical analyses; (3) casting georeferenced data generating mechanisms in terms of parametric mixture models involving ESFs; and, (4) improving our understanding of spatially structured and unstructured random effects terms that may appear in spatial statistical/econometric model specifications. New insights about these issues offer the potential to improve both spatial statistical and spatial econometric work.

Recerence

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

Bentkus V., Bloznelis M., Götze F. (1996), *A Berry-Esséen bound for Student's statistic in the non-i.i.d. case*, 'J. of Theoretical Probability', Springer, New York, 9

Besag J. (1974), *Spatial interaction and the statistical analysis of lattice systems*, 'J. of the Royal Statistical Society B', Wiley, New York, 36

Chaidee N., Tuntapthai M. (2009), *Berry-Esséen bounds for random sums of non-i.i.d. random variables*, 'International Mathematical Forum', m-Hikari, Ruse, 4

Cliff\ A., Ord J. (1969), *The Problem of Spatial Autocorrelation*, [in:] A. Scott (ed.) *London Papers in Regional Science*, Pion, London

Curry L. (1967), *Quantitative geography, 1967*, 'The Canadian Geographer', Wiley, New York, 11

Cvetković D., Rowlinson P. (1990), *The largest eigenvalue of a graph: a survey*, 'Linear and Multilinear Algebra', Taylor & Francis, Abingdon, 28

Gould P. (1967), *On the geographical interpretation of eigenvalues*, 'Transactions, Institute of British Geographers', Wiley, New York, 42

Griffith D. (1992), *Simplifying the normalizing factor in spatial autoregressions for irregular lattices*, 'Papers in Regional Science', Wiley, New York, 71

Griffith D. (2004a), *Extreme eigenfunctions of adjacency matrices for planar graphs employed in spatial analyses*, 'Linear Algebra & Its Applications', Elsevier, Amsterdam, 388

Griffith G. (2004b), *Faster maximum likelihood estimation of very large spatial autoregressive models: an extension of the Smirnov-Anselin result*, 'J. of Statistical Computation and Simulation', Taylor & Francis, Abingdon, 74

Griffith D. (2011a), *Positive spatial autocorrelation, mixture distributions, and geospatial data histograms*, [in:] Y. Leung, B. Lees, C. Chen, C. Zhou, and D. Guo (eds.), 'Proceedings 2011 IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM 2011)' , IEEE, Beijing

Griffith D. (2011b), *Positive spatial autocorrelation impacts on attribute variable frequency distributions*, 'Chilean J. of Statistics', Sociedad Chilena de Estadística, Valparaiso, 2 (2)

Griffith D., Paelinck J. (2011), *Non-standard Spatial Statistics and Spatial Econometrics*, Springer-Verlag, Berlin

Maćkiewic, A., Ratajczak W. (1996), *Towards a new definition of topological accessibility*, 'Transportation Research B', Elsevier, Amsterdam, 30

Mass C. (1985), *Computing and interpreting the adjacency spectrum of traffic networks*, 'Journal of Computational and Applied Mathematics', Elsevier, Amsterdam, 12&13

Ord J. (1975), *Estimation methods for models of spatial interactions*, 'Journal of the American Statistical Association', Taylor & Francis, Abingdon, 70

Pace R., LeSage J. (2004), *Chebyshev approximation of log-determinants of spatial weight matrices*, 'Computational Statistics and Data Analysis', Elsevier, Amsterdam, 45

Paelinck J. (2012), *Some challenges for spatial econometricians*, paper presented at the 2nd International Scientific Conference about Spatial Econometrics and Regional Economic Analys is, University of Lodz, Poland

Paelinck J., and Klaassen L. (1979), *Spatial Econometrics*, Saxon House, Farnborough

Smirnov O., Anselin L. (2001), *Fast maximum likelihood estimation of very large spatial autoregressive models: a characteristic polynomial approach*, 'Computational Statistics and Data Analysis', Elsevier, Amsterdam,, 35

Smirnov O., Anselin L. (2009), *An O(N) parallel method of computing the log-Jacobian of the variable transformation for models with spatial interaction on a lattice*, 'Computational Statistics and Data Analysis', Elsevier, Amsterdam, 53

Walde J., Larch M., Tappeiner G. (2008), *Performance contest between MLE and GMM for huge spatial autoregressive models*, 'J. of Statistical Computation and Simulation', Taylor & Francis, Abingdon, 78

Zhang Y., Leithead W. (2007), *Approximate implementation of the logarithm of the matrix determinant in Gaussian process regression*, 'J. of Statistical Computation and Simulation', Taylor & Francis, Abingdon, 77

Streszczenie

WYBRANE WYZWANIA STATYSTYKI PRZESTRZENNEJ DLA EKONOMETRYKÓW PRZESTRZENNYCH

Artykuł prezentuje wybrane, niestandardowe statystyki przestrzenne oraz zagadnienia ekonometrii przestrzennej. Rozważania teoretyczne koncentrują się na wyzwaniach wynikających z autokorelacji przestrzennej, nawiązując do pojęć Gaussowskiej zmiennej losowej, topologicznych cech danych georeferencyjnych, wektorów własnych, filtrów przestrzennych, georeferencyjnych mechanizmów generowania danych oraz interpretacji efektów losowych.