Chilean Journal of Statistics
Vol. 9, No. 1, April 2018, 3-17

Invited Paper

Impacts of negative spatial autocorrelation on frequency distributions

Yongwan Chun*, Daniel A. Griffith

School of Economic, Political and Policy Sciences, University of Texas at Dallas,
800 W. Campbell Road Richardson, Texas 75080, USA

(Received: September 4, 2017 · Accepted in final form: January 17, 2018)

Abstract

The literature contains treatments of various aspects of spatial autocorrelation. With its pervasiveness in empirical datasets, this literature focuses much more on positive spatial autocorrelation. In contrast, treatments of negative spatial autocorrelation scarcely appear in the literature. The purpose of this paper is to summarize impacts of negative spatial autocorrelation on data frequency distributions. Selection of this data visualization tool for study is because much empirical statistical research employs it in an initial data analysis step, and because findings here extend those reported in Griffith (2011) pertaining to positive spatial autocorrelation. This paper examines the first four moments of spatially autocorrelated random variables studied with simulation experiments. These simulations utilized a novel eigenvector spatial filtering based approach to generate spatially autocorrelation random variables.

Keywords: Histogram · Binomial random variable · Negative spatial autocorrelation · Normal random variable · Poisson random variable · Spatial autocorrelation.

1. Introduction

Spatial autocorrelation (SA)-the tendency for nearby values on a map to covary-has been a topic of interest in the geospatial sciences for a number of decades. Its treatment has followed the usual evolutionary trajectory of measurement [e.g., the Moran Coefficient (MC) and the Geary Ratio (GR)], then hypothesis testing (e.g., see Cliff and Ord, 1973), and finally modelling (e.g., spatial autoregression, Moran eigenvector spatial filtering, and geostatistical semivariogram functions). Positive SA (PSA), or the tendency for similar values to cluster on a map, has received most of the attention, largely because spatial researchers have found relatively few conspicuous empirical examples of negative SA (NSA), or the tendency for dissimilar values to cluster on a map. Spatial analysts tend to believe that NSA is a rare event; it is one of the most neglected topics in spatial statistics.

NSA refers to a geographic distribution of values, or a map pattern, in which, relatively speaking, the neighbors of locations with large values tend to have small values, the neighbors of locations with intermediate values tend to have intermediate values, and the neighbors of locations with small values tend to have large values. A Moran scatterplot

*Corresponding author. Email: ywchun@utdallas.edu (Y. Chun), dagriffith@utdallas.edu (D.A. Griffith)

ISSN: 0718-7912 (print)/ISSN: 0718-7920 (online)

http://www.soche.cl/chjs

for this situation, constructed by graphing the pairs of z-scores (z_i, sum of surrounding z_is), portrays a scatter of points that aligns along a trend line sloping from the upper left-hand to the lower right-hand quadrants of a graph. NSA naturally materializes with competitive locational processes, negative spatial externalities, the construction of spatial correlograms, the spectrum (i.e., eigenvalues) of a spatial weights matrix (SWM), the calculation of linear regression residuals, and the computation of LISA (local indicator of spatial association; see Anselin, 1995) statistics.

However, Spielman (2012) argues that NSA materializes in, for example, new forms of urban development, and that when NSA does occur, it often is of particular substantive interest. This situation is typical of time series analyses, too, in which most serial correlation empirically is found to be positive. Temporal autocorrelation exceptions include agricultural production (e.g., the cob-web model), credit spread return series, and high frequency financial data. Long-standing NSA exceptions include the regional Phillips curve reported by Anselin (1988), and relatively few intra-county population densities reported by Griffith et al. (2003). This situation may well be exemplified by the Wu et al. (1998) study of 361 agricultural plant breeding field trials, in which only eight trials (i.e., 2.2%) displayed NSA (the average value of the eight SA parameters is -0.28, for an NSA range of $[-1, 0]$).

The purpose of this paper is to summarize an investigation of the impact of NSA on data frequency distributions, extending Griffith's (2011) findings about how PSA can distort a frequency distribution for georeferenced data. This paper presents simulation experimental results based upon generated spatially autocorrelated random numbers, focusing on three popular distributions that are widely utilized in georeferenced data analyses (the normal, binomial, and Poisson). It examines the impacts of NSA with a focus on the first four moments (i.e., mean, variance, skewness, and kurtosis). This paper focuses on areal data (e.g., census data), which are very common for empirical datasets in the social sciences, including geography, regional science, economics, and demography.

2. Background

NSA is uncovered empirically by Saavedra (2000) as well as Boarnet and Glazer (2002) in both welfare and federal grants competition among local governments, by Montgomery and Chazdon (2001) in lowland Costa Rican second growth forest competition for light, by Irwin and Geoghegan (2001) in the Patuxent watershed land parcel development, by Stirböck (2002) in an index of investment specialization across Europe, by Garrett and Marsh (2002) in cross-border lottery shopping, and by Conley et al. (2003) in the spatial distribution of productivity in Malaysia. Furthermore, Gray and Shadbegian (2007) detect weak NSA in 102 industrial plant emissions of sulfur dioxide and nitrogen oxides across the medium-size United States (US) cities of St. Louis, Cincinnati, and Charlotte, whereas NSA is observed by Garretsen and Peeters (2009) in investment across OECD countries, by Basdas (2009) in the Turkish manufacturing industry, by both Filiztekin (2009) and Pavlyuk (2011) in regional employment, and by Elhorst and Zigova (2014) in research activity competition among a set of Economics Departments. Although this enumeration of occurrences appears to be sizeable, a comprehensive listing of PSA examples would eclipse it.

Nevertheless, the literature about NSA is still relatively scant.

3. Methodology

This section presents methodologies employed in the simulation experiments. First, it presents an overview of Moran eigenvector spatial filtering (MESF), and a discussion of an adjusted Moran Coefficient (MC) to expand the narrow range for NSA compared to PSA. Then, a MESF based approach is presented for generating SA embedded random numbers. Griffith (2017) comments that this method provides a single formula for drawing spatially autocorrelated random values from popular statistical distributions. In contrast, some popular parametric models do not support all aspects of SA. Especially, the popular auto-Poisson model can accommodate only negative spatial autocorrelation (Besag, 1974).

3.1 An overview of Moran eigenvector spatial filtering

MESF is a relatively novel methodology to handle SA contained in data (Griffith, 2003). MESF uses a set of synthetic proxy variables, which are extracted as eigenvectors from a n-by-n SWM, say $\boldsymbol{C}$, that ties n geographic objects together in space (indicating which are pairwise directly correlated), and then adds these vectors as control variables to a regression model specification. These control variables identify and isolate stochastic spatial dependencies among a set of georeferenced observations, filtering these dependencies out of a models residuals and adding them to the models mean response, thus allowing regression model building to proceed with observations that mimic being independent.

The MC SA index furnishes the basis for MESF; the GR or one of the other SA indices also could furnish this basis. The MC index can be written in matrix form as follows:

$$\frac{n}{\mathbf{1}^\top \boldsymbol{C}\mathbf{1}} \frac{\boldsymbol{Y}^\top(\boldsymbol{I}-\mathbf{1}\mathbf{1}^\top/n)\boldsymbol{C}(\boldsymbol{I}-\mathbf{1}\mathbf{1}^\top/n)\boldsymbol{Y}}{\boldsymbol{Y}^\top(\boldsymbol{I}-\mathbf{1}\mathbf{1}^\top/n)\boldsymbol{Y}},$$

where $\boldsymbol{Y}$ is a georeferenced variable, $\boldsymbol{I}$ is an n-by-n identity matrix, $\mathbf{1}$ is an n-by-1 vector of ones, n is the number of areal units, superscript $\top$ is the matrix transpose operator, and $\boldsymbol{C}$ is the binary 0-1 SWM. The eigenfunctions (i.e., the paired eigenvalues and eigenvectors) of interest are extracted from the modified SWM

$$(\boldsymbol{I}-\mathbf{1}\mathbf{1}^\top/n)\boldsymbol{C}(\boldsymbol{I}-\mathbf{1}\mathbf{1}^\top/n), \tag{1}$$

which appears in the numerator of this expression. When multiplied by $n/\mathbf{1}^\top\boldsymbol{C}\mathbf{1}$, an eigenvalue of this matrix is converted to the MC measuring the SA in its associated eigenvector (Tiefelsdorf and Boots, 1995; Griffith, 1996). The sign of an eigenvalue indicates the nature of SA represented by its corresponding eigenvector, whereas its magnitude indicates the degree of SA.

Extracting the eigenfunctions from SWMs constitutes a spectral decomposition of these matrices. The extracted eigenvectors with associated eigenvalues relatively far from zero, and hence representing other than negligible SA, may be viewed as portraying global, regional, or local components of SA because of the particular map patterns they take on when visualized. In other words, SA manifests itself in terms of similar (PSA) or dissimilar (NSA) values of variable Y clustering on a map.

MESF involves only the relevant eigenvectors when analyzing georeferenced data. A linear combination of these relevant eigenvectors is a constructed eigenvector spatial filter (ESF). Thus, the set of n eigenvectors needs to be reduced to this much smaller subset. The first screening is to set aside those eigenvectors portraying negligible degrees of SA; those set aside have eigenvalue absolute values less than some threshold value. If only PSA is of interest, then those eigenvectors with eigenvalues greater than this aforementioned

threshold constitute a candidate set; if only NSA is of interest, then those eigenvectors with eigenvalues less than the negative of this aforementioned threshold constitute a candidate set. A PSA-NSA mixture would include both of these preceding subsets. Chun et al. (2016) further discuss how a candidate set can be constructed. Once a candidate set is determined, then the selection of eigenvectors is with a stepwise regression procedure. This stepwise selection is legitimate because the n eigenvectors are mutually orthogonal and uncorrelated by construction. The resulting selected eigenvectors together with their regression coefficients allow the construction of an ESF.

3.2 The adjusted Moran Coefficient: initial developments

The MC NSA range roughly is between $(-0.5, -c)$ and $(-1.1, -c)$, where $-c = -1/(n-1)$ for a univariate case; most surface partitionings are closer to -0.5 than to -1.1. The adjusted MC, MC_{adj}, may be calculated as follows:

$$\text{MC}_{\text{adj}} = 2\Big(\frac{\text{MC} - \text{MC}_{\text{min}}}{\text{MC}_{\text{max}} - \text{MC}_{\text{min}}}\Big)^{\gamma} - 1,$$
$$\gamma = \frac{-\ln(2)}{\ln(-1/(n-1) - \text{MC}_{\text{min}}) - \ln(\text{MC}_{\text{max}} - \text{MC}_{\text{min}})}, \tag{2}$$

which has a range of $[-1, 1]$, with 0 [rather than $-1/(n-1)$] denoting no SA. The extreme MC values, MC_{max} and MC_{min}, are calculated with the extreme eigenvalues of matrix expression (1) (Jong et al., 1984). The exponent γ is calculated, not estimated, because all quantities in its definition are known. This exponent centers MC_{adj} at 0, in most cases stretching its NSA greatest lower bound to -1, and shrinking its PSA least upper bound to 1. The asymptotic standard error, σ_{MC}, is $\sqrt{2/\mathbf{1}^{\top}\boldsymbol{C}\mathbf{1}}$ (see Griffith, 2010). The corresponding adjusted derivation based upon mathematical statistical theory yields $\sigma_{\text{MC}_{\text{adj}}} = \big(2\sqrt{2/\mathbf{1}^{\top}\boldsymbol{C}\mathbf{1}}\big)/(\text{MC}_{\text{max}} - \text{MC}_{\text{min}})^{\gamma}$. Figure 1 portrays sampling distribution results based on a simulation experiment with 10,000 replications using a 70-by-68 $(= 4,760)$ regular square tessellation surface partitioning as well as the 2010 Dallas metroplex census tract (2,760 areal unit polygons) surface partitioning.

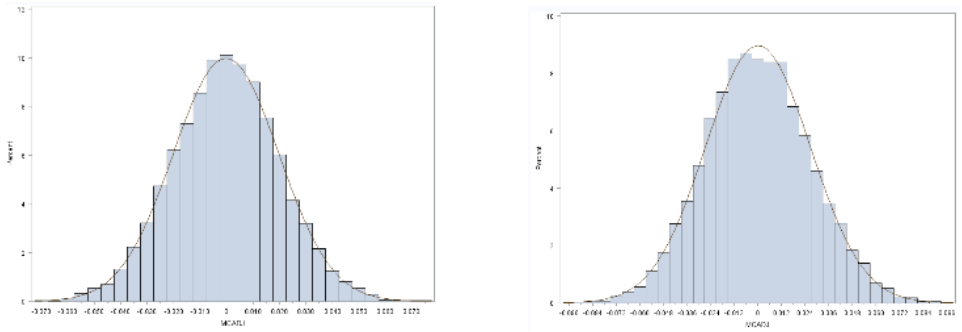
The square tessellation results, whose original MC range needs little adjustment, render $\gamma \approx 1$; the irregular surface partitioning for Dallas renders $\gamma \approx 0.7$. Both experiments render a simulated mean and variance that are almost identical to their theoretical counterparts (slight deviations are due to sampling error), and a normally distributed sampling distribution.

3.3 Eigenvector spatial filtering based random number generation

Because of many intractabilities associated with spatial statistics, spatial scientists use simulation experiments to establish many of its theoretical and conceptual properties and generalizations. Simulation experiments utilizing the auto-normal model employ the following relatively simple simultaneous autoregressive (SAR) mechanism to embed SA into univariate georeferenced data:

$$\boldsymbol{Y} = \mu\mathbf{1} + (\boldsymbol{I} - \rho\boldsymbol{W})^{-1}\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{I}), \tag{3}$$

where σ^2 denotes the landscape-wide constant variance of ϵ, which are independent and identically distributed (iid), and the SA parameter $\rho < 0$ denotes NSA. $\boldsymbol{W}$ is a row standardized SWM; its diagonal elements are zero. $(\boldsymbol{I} - \rho\boldsymbol{W})$ is positive definite so that the parameter space of ρ is $(1/\lambda_n, 1/\lambda_1)$, where λ_1 and λ_n are the largest and smallest



γ	$\text{MC}_{\max}$	$\text{MC}_{\min}$
1.0029876	1.0196336	-1.025495

Variable	Mean	Std. Dev.
MC	-0.000974264	0.0204344
MC_{adj}	-0.000153346	0.0200018

γ	$\text{MC}_{\max}$	$\text{MC}_{\min}$
0.7076671	-0.715145	1.1842969

Variable	Mean	Std. Dev.
MC	-0.0016676	0.0269278
MC_{adj}	0.000077291	0.0267067

Figure 1. Simulation experiment results. Left (a): for a 70-by-68 regular square tessellation. Right (b): for the 2010 Dallas metroplex census tract surface partitioning

eigenvalues of matrix $\boldsymbol{W}$. One advantage of this approach is that ρ in simulated values can be validated easily with popular estimation methods such as maximum likelihood, regardless of whether the SA is PSA or NSA. One weakness of this simulation approach is that it employs a prespecified nature and degree of SA, but without controlling for its associated map pattern (Griffith, 2017).

In contrast, MESF circumvents this pair of weaknesses. The following equation defines the mean response:

$$\boldsymbol{Y} = \alpha\mathbf{1} + \boldsymbol{E}_k\boldsymbol{\beta}_{E_k} + \boldsymbol{\epsilon}, \tag{4}$$

where $\boldsymbol{E}_k\boldsymbol{\beta}_{E_k}$ denotes ESF and $\boldsymbol{\epsilon}$ denotes a vector of non-spatial random errors. For generalized linear models involving binomial and Poisson regression, equation (4) can be expressed as

$$\boldsymbol{\mu} = g^{-1}(\alpha\mathbf{1} + \boldsymbol{E}_k\boldsymbol{\beta}_{E_k}), \tag{5}$$

where $g(\cdot)$ is a link function, with common links functions being the natural logarithm for Poisson models, and the logit for binomial models.

ESF based spatial random number generation can begin with construction of an ESF, $\boldsymbol{E}_k\boldsymbol{\beta}_{E_k}$, from a set of SAR-generated random numbers defined by equation (3), following Griffith (2011). That is, MESF is conducted with SAR random values in which SA is accounted for with $\boldsymbol{E}_k\boldsymbol{\beta}_{E_k}$. Furthermore, similar to the MC_{adj} in equation (2), the ESF_{adj} can be constructed to adjust for the unequal range of NSA vis-à-vis PSA. Then, the non-spatial random errors, $\boldsymbol{\epsilon}$, can be added for a normal distribution, $\mathcal{N}(\mathbf{0}, \sigma_\epsilon^2\boldsymbol{I})$, where σ_ϵ^2 is the standard deviation of $\widehat{\boldsymbol{\epsilon}}$. Meanwhile, Poisson random values are drawn with a vector $\boldsymbol{\mu}$, with each μ_i calculated as follows:

$$\mu_i = \frac{\exp[\ln(\alpha)]\exp[c_1\,\text{ESF}_{\text{adj},i}]}{\sum_i \exp[c_1\,\text{ESF}_{\text{adj},i}]/n}. \tag{6}$$

The SA level for a set of generated random values can be controlled with the weight term for the SA, c_1. Binomial random values can be drawn with a probability, p_i, and the

number of trials, n_{tr}. The probabilty term contains the SA; this quantity can be calculated as follows:

$$p_i = \frac{1}{1 + \exp(c_0 + c_1 \, \mathrm{ESF}_{\mathrm{adj},i})}. \tag{7}$$

The term c_0 is set to 0 for a specific scenario with an overall $p = 0.5$, and can be set to another value to obtain a different overall p level.

For all three of these random variables (RVs; i.e., normal, Poisson, binomial), the advantage of designing a simulation experiment with these equations is that the underlying map pattern is preserved, and, hence, the variance is preserved. For example, a spatial autoregressive simulation for normal RVs using the equation (3) SAR specification yields a $\widehat{\rho}$ equivalent to the input ρ, a constant average map, and a corresponding non-constant variance map (tending to decrease from the middle of this map to its periphery), whereas using the equation (5) MESF specification yields an average map that is the ESF, and a corresponding constant variance map.

4. Results

Simulation experiments, whose results are summarized in this paper, have been conducted with a 70-by-70 square tessellation. A spatial weights matrix is specified with the rook type contiguity. A total of 12 different levels of SA are employed for SAR random number generation: 0.1, 0.3, 0.5, 0.7, 0.9, and 0.95 for PSA, and $-0.1, -0.3, -0.5, -0.7, -0.9$, and -0.95 for NSA. Furthermore, 10 distinct map patterns for each of these PSA and NSA degrees of SA were generated; Appendix A presents one of these sets for each of the extreme cases of SA. Next, 10,000 random sets of numbers were generated for each of these distinct map patterns.

Figure 2 illustrates a simulated outcome for the three distributions with $\rho = -0.95$ for NSA (Figures 2a, 2c, and 2e), and $\rho = 0.95$ for PSA (Figures 2b, 2d, and 2e). Note that remotely sensed images often contain strong PSA (e.g., Li et al., 2016; Griffith and Chun, 2016), whereas empirical data with strong NSA rarely are recognized in the literature, as noted in the preceding background section. Figures 2a and 2b portray a simulated normal distribution following equation (4). The range of this simulated normal distribution is larger than that of a typical normal random set of values without SA, which indicates an increase in variance. In addition, these simulated sets do not show a big difference between PSA and NSA impacts. Similarly, Figures 2c and 2d portray outcomes for a simulated binomial distribution, and Figures 2e and 2d portray outcomes for a simulated Poisson distribution.

4.1 The normal distribution

Table 1 reports the distribution properties of spatially autocorrelated random values for extreme positive and negative SA cases (i.e., $\rho = 0.95$ and $\rho = -0.95$). Here, the nominal level of μ is zero and the variance is one; that is, ϵ was drawn from $\mathcal{N}(0, 1)$. These results show that while the means are around the nominal level (i.e., zero), the variance is substantially inflated for both the PSA and the NSA cases. Overall, skewness approximately equals its theoretical value (i.e., zero), whereas excess kurtosis overall tends to be less than its theoretical value (i.e., zero); more specifically, negative excess kurtosis occurs in 9 out of 10 sets. This outcome means that the normal distribution with a considerable level of SA is flatter than its iid counterpart. In general, no distinctive difference exists across the 10

map pattern results. Hence, hereafter, reported distributional properties are based upon a merging of results for the 10 map patterns.

Table 1. Distributional properties for spatially autocorrelated normal random values for the extreme PSA and NSA cases.

map	PSA $\rho = 0.95$				NSA $\rho = -0.95$			
	mean	var	skewness	kurtosis	mean	var	skewness	kurtosis
1	-0.3215	10.7001	0.1108	-0.0255	-0.0089	7.9181	0.0057	-0.0976
2	0.2102	8.4208	-0.0839	-0.1343	0.0039	6.4695	-0.0449	-0.1171
3	-0.0894	7.8912	-0.0844	-0.2630	0.0012	7.7540	0.0406	-0.1700
4	-0.0416	8.0904	-0.1739	0.2121	0.0000	6.8903	0.0164	-0.0599
6	-0.2472	7.9979	0.0022	-0.1332	-0.0032	8.6540	0.0225	-0.1194
7	-0.0632	7.3826	-0.0495	-0.0702	0.0001	7.1643	0.0001	-0.1390
8	0.2743	7.2534	0.0389	-0.0678	0.0068	7.3301	0.0061	-0.1594
9	0.6511	6.7562	0.0497	0.0168	0.0172	7.2461	0.0131	0.0615
10	-0.0422	7.0616	-0.0651	-0.1648	-0.0010	8.4085	-0.0022	0.2237

Note: the reported values are averages from 10,000 random sets; var denotes variance.

Figure 3 presents variation in the four distributional properties of the simulated random values as PSA and NSA change. Figure 3a confirms the unbiasedness of the mean for a normal distribution: it closely tracks zero across the ranges of PSA and NSA. In contrast, varince is consdierably affected by the level of SA; it is close to one, the nominal level, when ρ is 0, but increases exponentially as the absolute value of ρ increases. Figure 3b reveals that skewness is around zero across all levels of ρ, but that excess kurtosis decreases as the absolute value of ρ increases. These findings are consistent with Griffith (2011), who finds that PSA has an influence on variance and kurtosis. Of note is that these patterns do not differ much between PSA and NSA, at least for this simulation experiment.

4.2 The binomial distribution

Employing mathematical spatial statistical theory, Griffith (2010) extends PSA results for a normal RV to a suite of non-normal RVs. He also establishes how PSA impacts upon histograms (Griffith, 2011), relating normal, Poisson/negative binomial, and Bernoulli/binomial RVs the most commonly employed ones in the spatial sciences-containing PSA-to particular mixture model specifications.

The expectation is that histogram impacts from PSA and NSA differ. Consider a RV distributed across a 70-by-70 regular square tessellation. Its extreme MC values are 1.012 and -1.014 (rook's adjacency definition), which already suggests a difference. These values dictate the maximum level of SA that can materialize on this surface. A simulation experiment was conducted to study properties of spatially autocorrelated binomial and Poisson RVs. The curved lines in Figure 4a demonstrate for a binomial RV that, when $p = 1/2$ and SA is weighted equally with stochastic noise, this maximum level of SA is a function of the number of trials, n_{tr}. As $n_{tr} \rightarrow \infty$, a binomial distribution converges on a normal distribution, and its ability to capture maximum SA naturally increases. The parallel horizontal lines in Figure 4a demonstrate that when SA is weighted far more than stochastic noise, maximum SA still can materialize when n_{tr} is small. The most important aspect of this graphic is that both PSA and NSA behave in the same way. Figure 4b shows the maximum SA achieved with increasing weights, c_1 in equation (7), that is explored with a grid search of weights from 1 to 200. It shows that the extreme SA converges to that of a normal distribution with a large weight. Figure 4b shows the average weight values when extreme SA is achieved in Figure 4c with 100 repeated results. The weight values

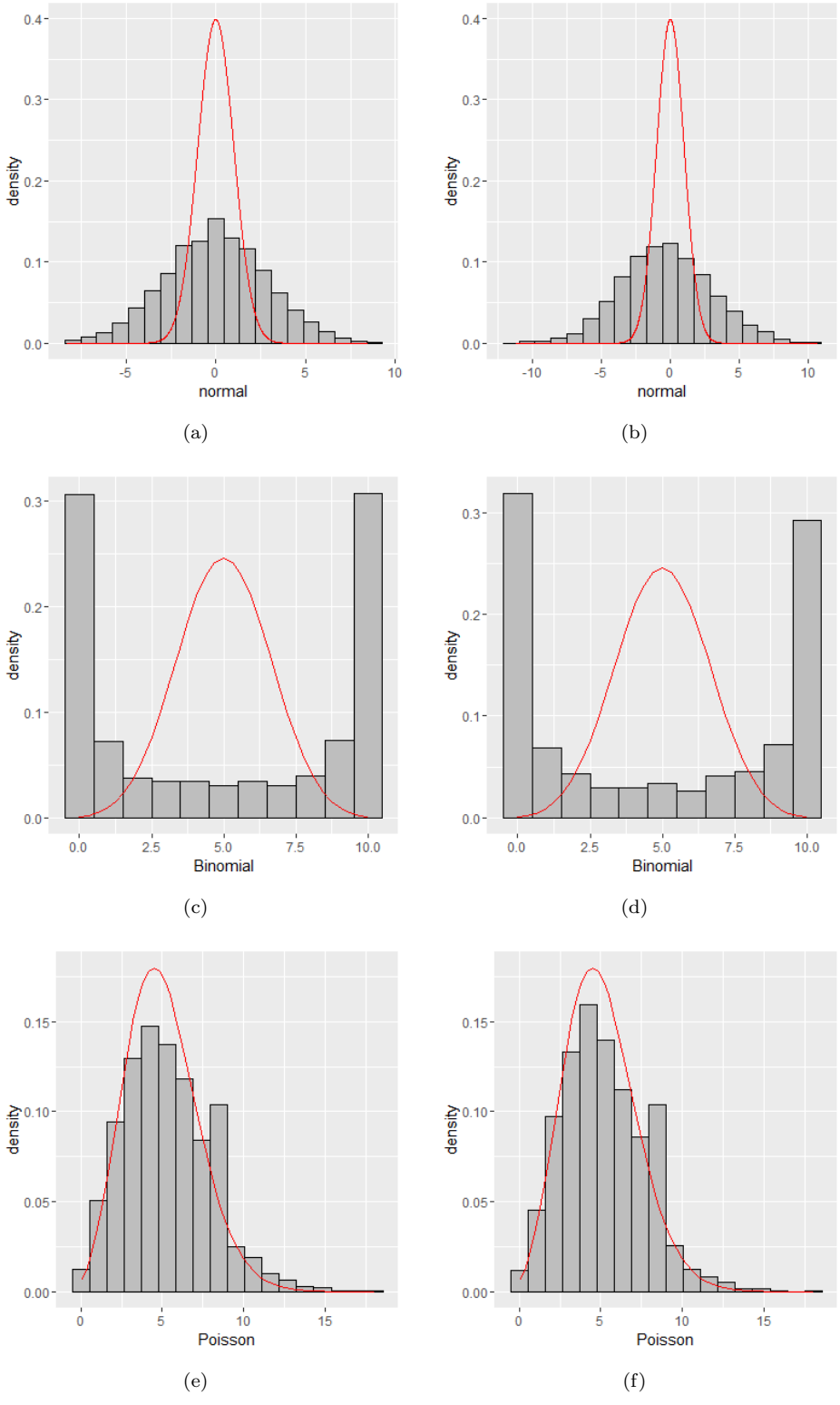
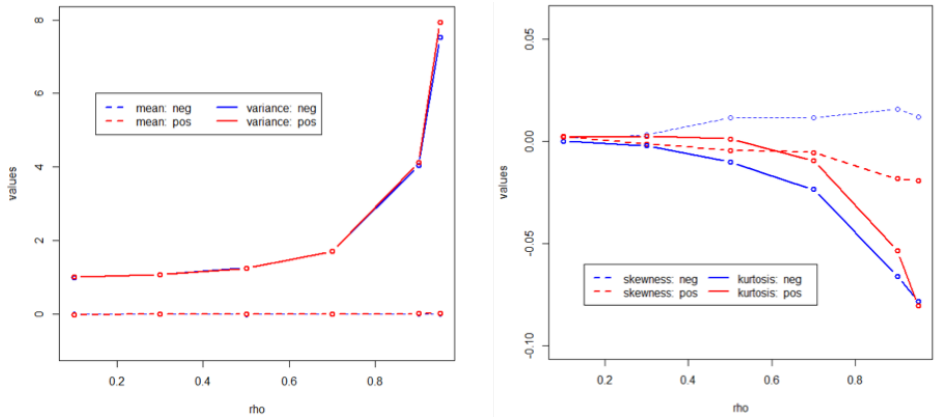
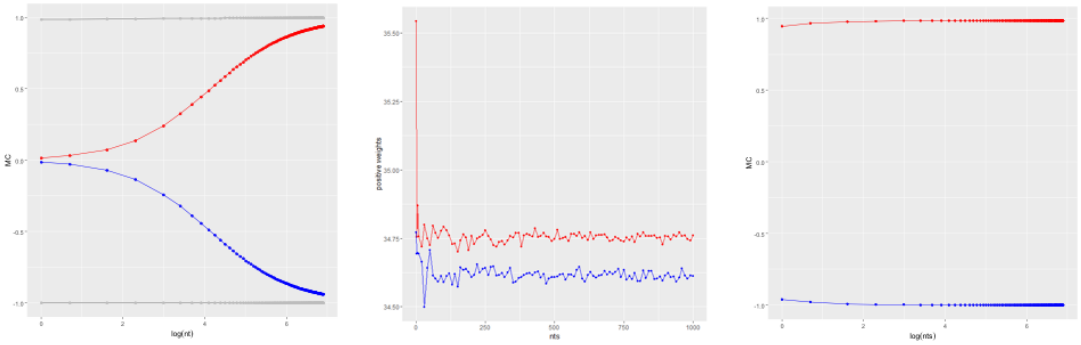


Figure 2. Simulated sets of random values with embedded spatial autocorrelation for a normal (a)-(b), a binomial (c)-(d), and a Poisson distribution (e)-(f). Note: (a), (c), and (e) are for $\rho = -0.95$ (an NSA case), and (b), (d), and (e) are for $\rho = 0.95$ (a PSA case).



(a) (b)

Figure 3. Distributional properities of simulated spatially autocorrelated normal random values.



(a) (b) (c)

Figure 4. The range of spatial autocorrelation for a binomial distribution; blue denotes NSA, and red denotes PSA. (a) the maximum SA by number of trials, with $c_1 = 1$ and $p = 1/2$. (b) the weight values, c_1, needed to achieve extreme SA. (c) the extreme SA levels acheived with c_1 values in (b).

are stable across the number of trials other than $n_{tr} = 1$ or 2, which are about 34.75 for PSA and 34.62 for NSA. Hence, binomial random values are generated with these weight values.

Figure 5 portrays the results for a binomial distribution. Figure 5a shows that the mean is generally not affected by the level of SA, with the dotted lines approximating the nominal levels across all ρ values (i.e., $n_{tr}p = 5$). In contrast, the variance is considerably inflated from the nominal level, $2.5 = n_{tr}p(1-p)$. The variance is larger even when $\rho = 0.1$ and -0.1, and increases as the absolute value of ρ increases. In Figure 5b, the skewness is constant (approximately zero, the nomial level) across all ρ values. That is, skewness is not affected by SA. In constrast, excess kurtosis, $[1-6p(1-p)]/[n_{tr}p(1-p)]$, is much smaller than the nominal level of -0.2, which indicates a flattened distribution. Excess kurtosis further descreases as ρ increases. The same patterns are observed in Figures 5c and 5d, for which $n_{tr} = 200$. That is, while the mean and skewness are not affected by SA, the variance and the kurtosis experience a considerable impact. Specically, the variance is considerably larger than its theoretical value of 50. This overall pattern for a binomial distribtuion is similar to that for a normal distribution. However, although it shows excessive variance, it also shows little excess kurtosis, for small ρ values (i.e., $\rho = 0.1$ and -0.1). This difference between a binomial and a normal RV may arise with the relative weight attached to the spatial component here, which is shown in Figure 4b.

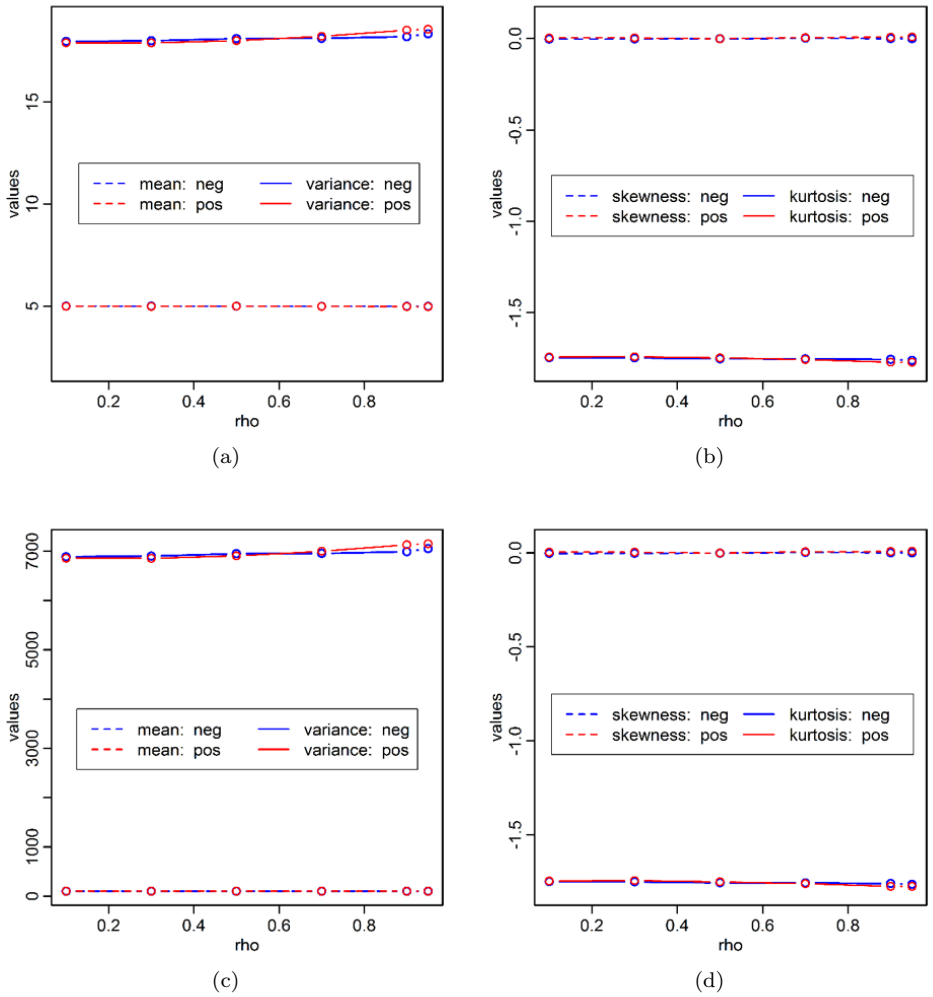


Figure 5. Distributional properities of simulated spatially autocorrelated binomial random numbers with $n_{tr} = 10$ (a, b), and $n_{tr} = 200$ (c, d)

4.3 The Poisson distribution

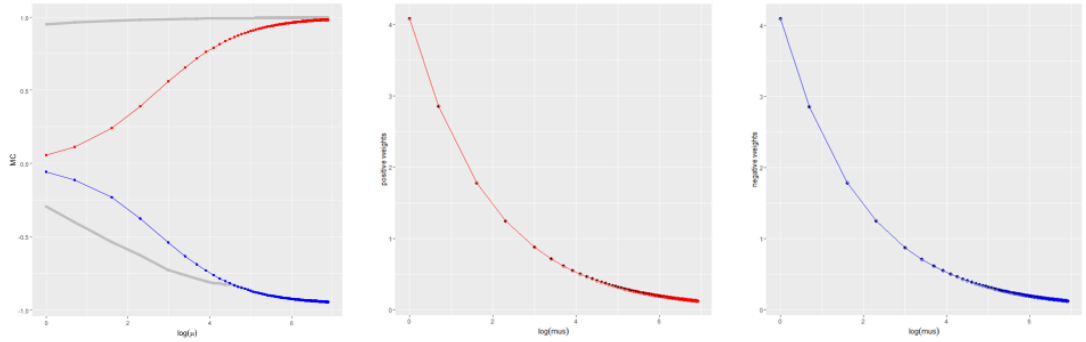
The curved lines in Figure 6a demonstrate for a Poisson RV that, when SA and stochastic noise are weighted equally, this maximum level of SA is a function of the Poisson mean, μ. As $\mu \to \infty$, a Poisson distribution converges on a normal distribution, and its ability to capture maximum SA naturally increases. The dotted lines with small filled circles in Figure 6a demonstrate that when SA is weighted far more than stochastic noise with a high value for c_1 in Equation (6), a greater degree of SA can materialize. But, unlike the preceding binomial RV (which is symmetric because $p = 1/2$), maximum PSA or NSA cannot necessarily materialize. Furthermore, a substantial difference exists between PSA and NSA outcomes here. This finding suggests that the behavior of a binomial RV with $p \neq 1/2$ may deviate from that portrayed in Figure 4a. Figures 6b and 6c show that the weights to achieve extreme SA are negative exponentially related to μ. The non-linear model fit in Figure 6b suggests the following equation for PSA weights:

$$\widehat{c_1} = 4.07100 \exp(-0.51662 \log(\mu)) + 0.01067,$$

with $R^2 = 0.99998$. The non-linear model fit in Figure 6c suggests the following equation for NSA weights:

$$\widehat{c}_1 = 4.07674 \exp(-0.51811 \log(\mu)) + 0.01175,$$

with $R^2 = 0.99997$. Hence, Poisson random values are generated with $\widehat{c}_1$ from these equations, which appear to differ only by sampling error.



(a) (b) (c)

Figure 6. The range of spatial autocorrelation for a Poisson distribution; blue denotes NSA, and red denotes PSA. (a) the maximum SA as a function of μ, with $c_1 = 1$ (inner curves) and c_1 at its maximum (outer curves). (b) weights to acheive maximum PSA, as a function of μ. (c) weights to acheive maximum NSA, as a function of mu.

Figure 7 portrays selected properties of a set of Poisson random numbers. The overall patterns are similar to those for a binomial distribution. Figures 7a and 7b show that the mean of a Poisson distribution ($\mu = 5$ here) is not affected by SA, whereas the variance is substantially inflated. The mean and the variance of a classical Poisson distribution are the same, μ. In this example, the variance is greater than 6.5, even for the smallest magnitudes of ρ (i.e., $\rho = 0.1$ and -0.1); it increases as ρ increases. The observed skewness is greater than its theoretical value of 0.4472 ($= \mu^{-1/2}$) for independent values, and the observed excess kurtosis also is greater than its theoretical counterpart of 0.2 ($= \mu^{-1}$). But the observed skewness values fall within a small range across the ρ levels. The results for $\mu = 100$ produce the same pattern. Figure 7c shows that the mean is around 100, the nomial level, while the variance is greater than 100, and increases as ρ increases. Figure 7d shows that the skewness and excess kurtosis constantly are greater than their theoretical values, which is 0.1 for skewness and 0.01 for kurtosis. Except for the difference in the maximum possible SA, depicted in Figure 6a, these results demonstrate that there is no distictive difference between the impacts of PSA and NSA. Also, the parameter value, μ, does not make a substantial difference in the patterns of the distributional properties for Poisson random values.

5. Summary and Conclusions

This paper summarizes an investigation of the impact of NSA on the frequency distribution of three RVs commonly employed to describe georeferenced data: the normal, binomial, and Poisson. In this paper, random values with SA embedded are generated using MESF, and the distributional properties of the resulting spatially autocorrelated random values are examined.

Simulation experiment results indicate that SA has virtually no impact on the means of the three selected probability distribution models. In contrast, SA inflates the variance,

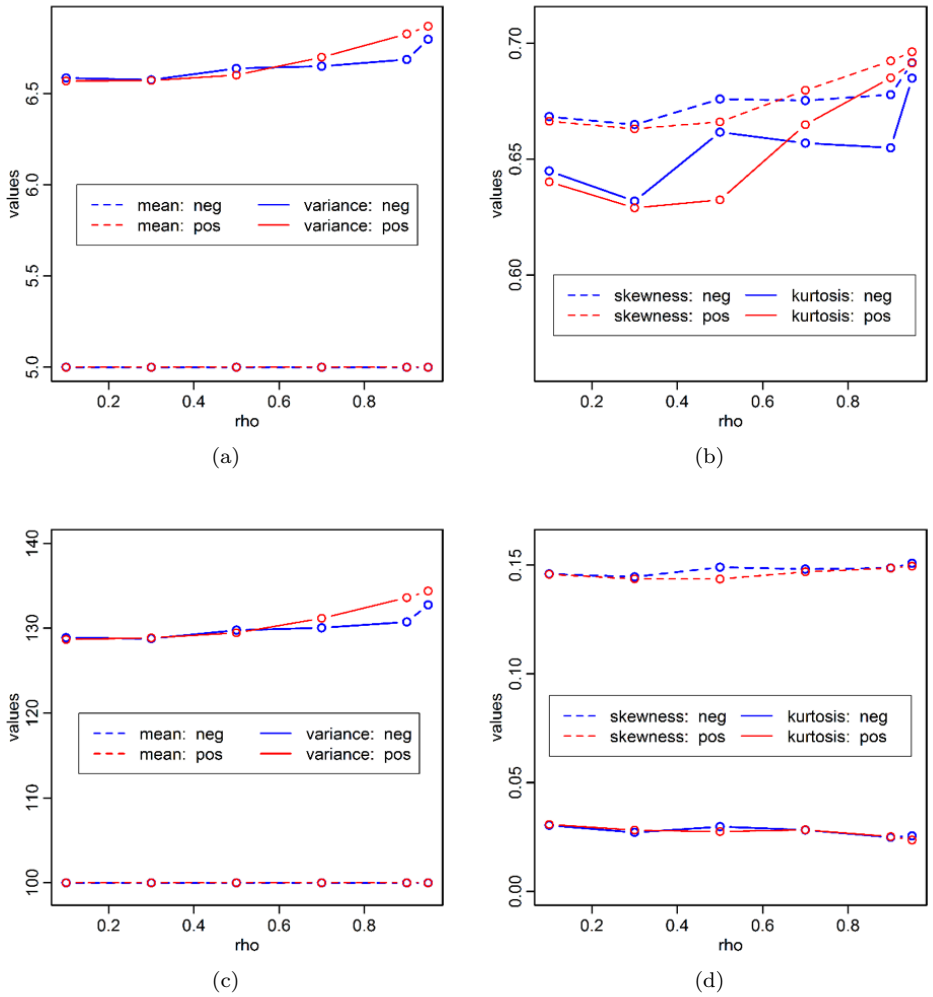


Figure 7. Distributional properities of simulated spatially autocorrelated Poisson random numbers with $\mu = 5$ (a, b) and $\mu = 100$ (c, d)

with this inflation increasing as the level of SA increases. This increasing pattern is conspicuous for the normal distribution, which exhibits little inflation when ρ is small (i.e., $\rho = 0.1$ and -0.1). In contrast, binomial and Poisson distributions have a markedly inflated variance for $\rho = 0.1$ and -0.1. These findings confirm that spatially autocorrelated binomial and Poisson values tend to have overdispersion. Meanwhile, skewness essentially is unaffected by SA for a normal distributions, while its kurtosis decreases as the level of SA increases. This change in kurtosis can be expected with the occurrence of variance inflation. In contrast, skewness and kurtosis are greater than their theoretical counterparts for binomial and Poisson distributions. These statistics also are stable across SA levels, unlike their behavior for a normal distribution. Interestingly, impacts of NSA are not substantially different from those of PSA in these simulation experiments.

This paper further contributes to the literature in two ways. First, it demonstrates that the feasible range of SA is dependent on a mean level for non-normal RVs. When a mean is small, the feasible range of SA is small for both binomial and Poisson distributions. This paper also shows that this feasible range can be expanded by increasing the relative weighting of SA to random noise components. In addition, the PSA and NSA feasible ranges for a binomial distribution with $p = 0.5$ are symmetric, whereas they are asymmetric for a Poisson distribution with a small μ. This outcome implies that the feasible range

might become asymmetric as p deviates from 0.5. Second, this paper shows how spatially autocorrelated random values can be generally based on MESF following Griffith (2011). This method provides a novel approach that can be used to generate non-normal as well as normal random values.

The study upon which this paper is based can be further extended in future research. First, the regular surface partitioning employed can be replaced with an irregular partitioning, which better characterizes most empirical data, supporting a more comprehensive exploration of PSA and NSA impacts on statistical frequency distributions. Second, rather than studying pure PSA and NSA cases, an investigation of mixtures of PSA and NSA should prove illuminating. Although NSA is one of the most ignored topics in spatial statistics, it has received more attention than the mixture topic. This notion was first introduced in a formal way by Griffith and Arbia (2010), who acknowledge that cases of zero SA can result from these two SA components cancelling each other. A propensity for PSA to dominate is one potential reason spatial scientists rarely detect NSA. One of the first encountered examples of this combination was the behavior of the Anopheles arabiensis mosquito (Jacob et al., 2009). Third, impacts of NSA or a mixture of PSA and NSA on other statistical features need to be investigated further. These statistical features include, especially, biasness, consistency, and robustness for regression coefficients as well as effective sample size (Griffith, 2005).

Acknowledgement

This research was supported by the National Institutes of Health, grant 1R01HD076020-01A1. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors, and do not necessarily reflect the views of the National Institutes of Health.

References

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

Anselin, L. (1995). Local Indicators of Spatial Association: LISA. Geographical Analysis 27, 93-115.

Basdas, U. (2009). Spatial econometric analysis of the determinants of location in Turkish manufacturing industry. doi: 10.2139/ssrn.1506888

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society, Series B 36, 192-236.

Boarnet, M., and Glazer, A. (2002). Federal grants and yardstick competition. Journal of Urban Economics 52, 53-64.

Cliff, A., and Ord, J. (1973). Spatial Autocorrelation. Pion, London.

Conley, T., Flyer, F., and Tsiang, G. (2003). Spillovers from local market human capital and the spatial distribution of productivity in Malaysia. Advances in Economic Analysis & Policy 3 (1), Article 5. doi: 10.2202/1538-0637.1229

Chun, Y., Griffith, D.A., Lee, M., and Sinha., P. (2016). Eigenvector selection with stepwise regression techniques to construct eigenvector spatial filters. Journal of Geographical Systems 18, 67-85.

Elhorst, J. and Zigova, K. (2014). Competition in research activity among economic departments: evidence by negative spatial autocorrelation. Geographical Analysis 46, 104-125.

Filiztekin, A. (2009). Regional unemployment in Turkey. Papers in Regional Science 88, 863-879.

Garretsen, H., and Peeters, J. (2009). FDI and the relevance of spatial linkages: do third-country effects matter for Dutch FDI? Review of World Economics 145, 319-338.

Garrett, T., and Marsh, T. (2002). The revenue impacts of cross-border lottery shopping in the presence of spatial autocorrelation. Regional Science and Urban Economics 32, 501-519.

Gray, W.B., and Shadbegian, R.J. (2007). The environmental performance of polluting plants: A spatial analysis. Journal of Regional Science 47, 63-84.

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

Griffith, D.A. (2003). Spatial Autocorrelation and Spatial Filtering: Gaining Understanding through Theory and Scientific Visualization. Springer-Verlag, Berlin.

Griffith, D.A. (2005). Effective geographic sample size in the presence of spatial autocorrelation. Annals of the Association of American Geographers 95, 740-760.

Griffith, D.A. (2010). The Moran Coefficient for non-normal data. Journal of Statistical Planning and Inference 140, 2980-2990.

Griffith, D.A. (2011). Positive spatial autocorrelation impacts on attribute variable frequency distributions. Chilean Journal of Statistics 2, 3-28.

Griffith, D.A. (2017). Some robustness assessments of Moran eigenvector spatial filtering. Spatial Statistics 22, 155-179.

Griffith, D.A., and Arbia, G. (2010). Detecting negative spatial autocorrelation in geo-referenced random variables. International Journal of Geographical Information Science 24, 417-437.

Griffith, D.A., and Chun. Y. (2016). Spatial autocorrelation and uncertainty associated with remotely-sensed data. Remote Sensing 8, 535.

Griffith, D.A., Wong, D., and Whitfield, T. (2003). Exploring relationships between the global and regional measures of spatial autocorrelation. Journal of Regional Science 43, 683-710.

Irwin, G., and Geoghegan, J. (2001). Theory, data, methods: developing spatially-explicit economic models of land use change. Agriculture Ecosystems and Environment 85, 7-24.

Jacob, B., Griffith, D.A., Muturi, E., Caamano, E., Githure, J., and Novak, R. (2009). A heteroskedastic error covariance matrix estimator using a first-order conditional autoregressive Markov simulation of ecological sampled Anopheles arabiensis aquatic habitat covariates. Malaria Journal 8, 216.

Jong, P.D., Sprenger, C., and Veen, F.V. (1984). On extreme values of Moran's I and Geary's c. Geographical Analysis 16, 17-24.

Li, B., Griffith, D.A., and Becker, B. (2016). Spatially simplified scatterplots for large raster datasets. Geo-spatial Information Science 19, 81-93.

Montgomery, R., and Chazdon, R. (2001). Forest structure, canopy architecture, and light transmittance in old-growth and second-growth tropical rain forests. Ecology 82, 2707-2718.

Pavlyuk, D. (2011). Spatial analysis of regional employment rates in Latvia. Scientific Journal of Riga Technical University 2, 56-62.

Saavedra, L. (2000). A model of welfare competition with evidence from AFDC. Journal of Urban Economics 47, 248-279.

Spielman, S. (2012). Exceptions to the law: negative spatial autocorrelation in egocentric spatial analysis. paper presented at GIScience 2012: 7th International Conference on Geographic Information Science, Columbus, Ohio, September 18-21.

Stirböck, C. (2002). Explaining the level of relative investment specialisation: A spatial econometric analysis of EU regions. ZEW Discussion Paper No. 02-49.
doi: 10.2139/ssrn.336501

Tiefelsdorf, M., and Boots, B. (1995). The exact distribution of Morans I. Environment and Planning A: Economy and Space 27, 985-999.

Wu, T., Mather, D., and Dutilleul, P. (1998). Application of geostatistical and neighbor analyses to data from plant breeding trials. Crop Science 38, 15451553.

Appendix A. Map patterns for PSA with $\rho = 0.9$ and NSA with $\rho = -0.95$

The following maps present the 10 distinct PSA and NSA patterns that were used in the simulation experiments.