remote sensing

MDPI

Article

Spatial Autocorrelation and Uncertainty Associated with Remotely-Sensed Data

Daniel A. Griffith *,† and Yongwan Chun

School of Economic, Political and Policy Sciences, The University of Texas at Dallas, 800 West Campbell Road, Richardson, TX 75080, USA; ywchun@utdallas.edu

* Correspondence: dagriffith@utdallas.edu; Tel.: +1-972-883-4950

† Daniel A. Griffith is an Ashbel Smith Professor.

Academic Editors: Yudong Tian, Ken Harrison, Yoshio Inoue and Prasad S. Thenkabail

Received: 5 March 2016; Accepted: 14 June 2016; Published: 23 June 2016

Abstract: Virtually all remotely sensed data contain spatial autocorrelation, which impacts upon their statistical features of uncertainty through variance inflation, and the compounding of duplicate information. Estimating the nature and degree of this spatial autocorrelation, which is usually positive and very strong, has been hindered by computational intensity associated with the massive number of pixels in realistically-sized remotely-sensed images, a situation that more recently has changed. Recent advances in spatial statistical estimation theory support the extraction of information and the distilling of knowledge from remotely-sensed images in a way that accounts for latent spatial autocorrelation. This paper summarizes an effective methodological approach to achieve this end, illustrating results with a 2002 remotely sensed-image of the Florida Everglades, and simulation experiments. Specifically, uncertainty of spatial autocorrelation parameter in a spatial autoregressive model is modeled with a beta-beta mixture approach and is further investigated with three different sampling strategies: coterminous sampling, random sub-region sampling, and increasing domain sub-regions. The results suggest that uncertainty associated with remotely-sensed data should be cast in consideration of spatial autocorrelation. It emphasizes that one remaining challenge is to better quantify the spatial variability of spatial autocorrelation estimates across geographic landscapes.

Keywords: spatial autocorrelation; spatial variability; NDVI; NBR; Florida Everglades

1. Introduction

Spatial autocorrelation prevails in virtually all georeferenced data, tending to be moderate and positive for socio-economic/demographic data (i.e., correlations between 0.4 and 0.6), and positive and very strong for remotely sensed data (i.e., correlations between 0.85 and 0.95). One well known impact of positive spatial autocorrelation is variance inflation (VIF; e.g., [1]) which, in turn, impacts uncertainty quantification and assessment of remotely sensed data. However, although popular versions of spatial regression techniques have existed since 1972 [2] as spatial statistical tools, and since 1988 [3] as spatial econometric tools, remote sensing researchers continue to shy away from them and use non-spatial regression techniques (e.g., [4]; as of 21 January 2016, this article had 447 Google Scholar citations, and 216 Web of Science citations). Principal drawbacks of these implementations vis-à-vis massively large georeferenced datasets, such as remotely-sensed images, include that they: (1) involve nonlinear regression, which requires multiple iterations, each essentially executing a linear regression, to calculate parameter estimates; and (2) require calculating the eigenvalues of an n-by-n spatial weights matrix in order to compute the normalizing constant for an auto-normal probability model.

One recent spatial statistical advance replaces the nonlinear regression solution with a condensed (i.e., by reducing the number of parameters upon which the function depends from three to one) normal equation solution [5]. This substitution works extremely well for remotely-sensed data because

they contain extremely strong positive spatial autocorrelation; because of the form of the auto-normal likelihood function [6], this new solution still needs some tweaking for negative spatial autocorrelation situations, which are rarely encountered for remotely-sensed data. Meanwhile, because the extreme eigenvalues of a spatial weights matrix define the feasible range of the spatial autocorrelation parameter, a relatively simple implementation can be achieved with only them. They are ± 1 for a row-standardized spatial weights matrix and, hence, do not need to be calculated. However, the remaining n–2 eigenvalues are unknown, although they can be very accurately approximated [5] (p. 2417).

The purpose of this paper is to outline methodology for quantifying uncertainty in remotely-sensed data that relates to spatial autocorrelation latent in these data. Doing so should aid in the extraction of information and distilling of knowledge from such data. An illustration of this methodology uses both simulation experiments and a remotely-sensed image of the Florida Everglades, United States (US).

2. The Florida Everglades Data

The empirical dataset employed for illustrative purposes in this paper is a 1 January 2002 Landsat 7 Enhanced Thematic Mapper Plus (ETM+) image of the Florida Everglades forming a 7649-by-8581 (n = 65,636,069 pixels) rectangular region rotated clockwise on the horizontal axis (Figure 1). This image has been orthorectified and converted to the UTM 17-N projection, and includes spectral bands B1–B7; its spatial resolution is 28.5 m for bands B1–B5 and B7, and 57 m for B6 [7]. Pixels with nonzero spectral reflectance values total 41,611,007 (82.38%), whereas 8,935,349 pixels with a zero value (17.68%) form a white border around the remotely-sensed image.

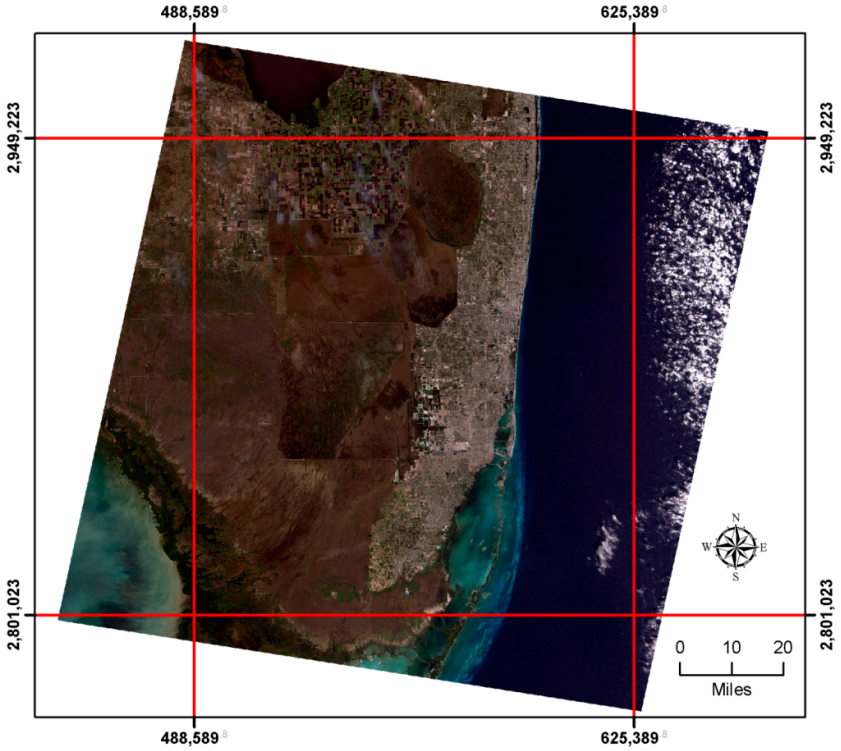


Figure 1. A composite of the Florida Everglades remotely-sensed image respectively using bands B1, B2, and B3 for red, green, and blue colors. The red lines demarcate the study area.

To simplify the sampling experiments whose results are summarized in this paper, a 4800-by-5200 (n = 24,960,000 pixels) rectangular region parallel to the horizontal axis was extracted for analysis purposes (demarcated by red lines in Figure 1). The sampling experiments use two spectral indices, the normalized difference vegetation index (i.e., NDVI = (B4 − B3)/(B4 + B3)) and the normalized burn ratio (i.e., NBR = (B4 − B7)/(B4 + B7)) Both indices range from −1 to 1, and have a variance that,

conceptually, can range from 0 to 1. Large positive NDVI values indicate dense vegetation land cover, whereas large negative values indicate deep water. Large positive NBR values indicate high severity burn, whereas large negative values indicate high post-fire regrowth. The criterion used to select these two indices is their levels of spatial autocorrelation: NBR tends to be around 0.85 (the Everglades image analyzed in this paper has a spatial autocorrelation of 0.874), the common lower bound for many remotely-sensed images, and NDVI tends to be around 0.95 (the Everglades image analyzed in this paper has a spatial autocorrelation of 0.955), the common upper bound for many remotely-sensed images (each of these empirical results is an average of three different approximation method results).

Two of the sampling experiments whose results are summarized in this paper utilized 400-by-400 sub-regions of the Florida Everglades image; the 4800-by-5200 study region can be subdivided into 156 mutually exclusive and collectively exhaustive coterminous sub-regions of this size. Selection of this size is based upon the smaller size remotely sensed images commonly analyzed. For example, Griffith [5] identifies a Yellowstone Park (US) image with dimensions 450-by-350 (i.e., n = 257,500 pixels), and an Adirondack Park (US) image with dimensions 511-by-503 (i.e., n = 257,033 pixels). The 400-by-400 dimensions also were utilized in the simulation experiments.

3. Spatial Regression Model Based Sampling Variability of the Spatial Autocorrelation Parameter

Spatial regression models are widely adapted to model a spatially-autocorrelated response variable. Specifically, the simultaneous spatial autoregressive (SAR) specification and the spatial autoregressive response (AR) specification are popular counterparts of linear regression under the Gaussian assumption. These specifications are utilized in this paper since distributional properties of their spatial autocorrelation parameter have been known in the literature (e.g., [8]). This paper treats the response variable (e.g., NDVI, NBR) case of pure spatial autocorrelation, recognizing that findings can be extended to cases that include covariates (which most likely will reduce the degree of residual spatial autocorrelation). The SAR specification essentially is identical to the AR specification for pure spatial autocorrelation. This simple SAR model specification may be written as:

$$\mathbf{Y} = (\mathbf{1} - \rho)\mu\mathbf{1} + \rho\mathbf{W}\mathbf{Y} + \boldsymbol{\varepsilon} \tag{1}$$

where $\mathbf{Y}$ denotes the n-by-1 vector of response variable (e.g., NDVI, NBR) values, μ is the population mean of variable Y, ρ is the spatial autocorrelation parameter (which ranges between -1 and 1 in this situation, and quantifies a signal in the data), $\mathbf{W}$ is the n-by-n spatial weights matrix (i.e., a quantification of the configuration of pixels), and ε is an n-by-1 vector of random error values (i.e., noise), which jointly are assumed to be normally distributed $N(0,\ \sigma^2\mathbf{I})$. The entries in matrix $\mathbf{W}$ are defined with the row-standardized rook adjacency rule: $w_{ij} = 1/n_i$ if pixels i and j share a non-zero length common boundary, and 0 otherwise; $w_{ii} = 0$, where n_i denotes the number of neighbors pixel i has.

Ord [8] establishes the asymptotic variance of the maximum likelihood estimator of ρ ($\hat{\rho}$), which furnishes the most common standard error used to evaluate estimates of this parameter. His result reduces to:

$$_{ASY}\sigma^2_{\hat{\rho}} = \frac{1}{\mathrm{TR}[(\mathbf{I} - \rho\mathbf{W}^T)^{-1}\mathbf{W}^T\mathbf{W}(\mathbf{I} - \rho\mathbf{W})^{-1}] + \sum_{j=1}^{n} \frac{\lambda_j^2}{(1-\rho\lambda_j)^2} - \frac{2}{n}\{\mathrm{TR}[\mathbf{W}(\mathbf{I} - \rho\mathbf{W})^{-1}]\}^2} \tag{2}$$

where TR denotes the matrix trace operator, superscript T denotes the matrix transpose operator, and λ_j (j = 1, 2, . . . , n) denotes the eigenvalues of matrix $\mathbf{W}$ ($\lambda_1 = 1$, and $\lambda_n = -1$) in descending order. If the null hypothesis is zero spatial autocorrelation, which commonly is posited, then Equation (2) reduces to:

$$_{ASY}\sigma^2_{\hat{\rho}} = \frac{1}{TR[\mathbf{W}^T\mathbf{W}] + \sum_{j=1}^{n}\lambda_j^2} = \frac{1}{2\sum_{j=1}^{n}\lambda_j^2 + \frac{P+Q+12}{72}} = \frac{1}{2\frac{18PQ+11P+11Q+12}{72} + \frac{P+Q+12}{72}} \quad (3)$$

for a regular square tessellation forming a P-by-Q complete rectangular region (e.g., a remotely-sensed image with P rows and Q columns). Equation (3) can be further re-arranged as:

$$_{ASY}\sigma_{\hat{\rho}} = \sqrt{\frac{72}{36PQ + 23(P+Q) + 36}} \quad (4)$$

a quantity that describes the variability of $\hat{\rho}$ regardless of whether exact eigenvalues, approximate eigenvalues, or Monte Carlo simulated eigenvalues (see Appendix A) are used to calculate the Jacobian term or its approximation, but because of its magnitude, this quantity is of little value for quantifying uncertainty for remotely-sensed images whose $\hat{\rho}$ tends to be 0.85 to 0.95, or more. Rather, it suggests that the large sample sizes associated with remotely-sensed data render rather precise $\hat{\rho}$ s: if $\hat{\rho} = 0.95$ for a 400-by-400 image, then 95% of the time ρ should be contained in the interval [0.943, 0.957].

The spatial autocorrelation parameter can be transformed to the range $(1+\rho)/2$. This transformation results in a random variable (RV) that exhibits many properties of an overdispersed beta random variable (BRV), which is defined with two parameters α and β (which appear as exponents in the BRV probability density function, and control the shape of its frequency distribution). Simulation experiments suggest that the parameters α and β of this BRV are such that $\alpha + \beta \approx n$, the number of pixels (i.e., the sample size), and more specifically that $\alpha = pn$ and $\beta = (1-p)n$, $0 < p < 1$, with p (i.e., the percentage of a sample size) increasing as spatial autocorrelation increases from -1 to 1. The parameter p governs the sampling variability for a given nature and degree of spatial autocorrelation, constituting the source of overdispersion. Since it is equivalent to a percentage, p also has many properties of a BRV. The parametric mixture yields a beta-beta random variable (BBRV). Equation (4) allows the beta distributional parameters for p to be calibrated: $\alpha = 80512/\sqrt{1-\rho}$ and $\beta = 80512\sqrt{1-\rho}/(1+\rho)$ for a 400-by-400 image (the value 80,512 was calculated by equating the theoretical variance for ρ = 0 with the BBRV variance, and then solving for the single BBRV parameter). Figure 2 furnishes example sampling distributions for the specimen 400-by-400 image size studied in this paper, for $\rho = -0.85, -0.5, 0, 0.5$, and 0.95. Variation in height in these distributions (Figure 2a) indicates the shrinking variance with increasing $|\rho|$. For example, for ρ = 0.95, a high concentration of $\hat{\rho}$ results in the tallest height of the distribution among the five levels. Figure 2b–d portray zoom-ins of the sample sampling distributions for ρ = 0, 0.5, and 0.95. Table 1 summarizes selected statistics for these types of distributions. These tabulated results corroborate the BBRV specification: the sampling distribution average equals the population parameter; the extra-beta variation is accounted for; negative skewness attributable to the upper bound of 1 exists, and positive skewness attributable to the lower bound of -1 exists; and, peakedness equivalent to that for a normal curve is present. Consequently, sound asymptotic standard errors can be calculated for any size remotely sensed image with this BBRV approximation; all that is needed is n and the value from Equation (4).

Table 1. Beta-beta random variable (BBRV) summary statistics for a 400-by-400 image.

ρ	Source	Mean of Estimates ($\bar{\hat{\rho}}$)	Standard Error	Skewness	Kurtosis
0	BBRV	0.00	0.00353	0	3.00
	simulation	0.00	0.00354	0.02	2.99
0.5	BBRV	0.50	0.00310	−0.01	3.00
	simulation	0.50	0.00289	0.02	2.89
0.95	BBRV	0.95	0.00093	−0.03	3.00
	simulation	0.95	0.00088	−0.10	3.12

NOTE: the simulation experimental design included 1000 replications; NOTE: BBRV parameter values were obtained with the Mathematica 10.2 Parameter Mixture Distribution function, and then verified with SAS simulations involving 1,000,000 random draws.

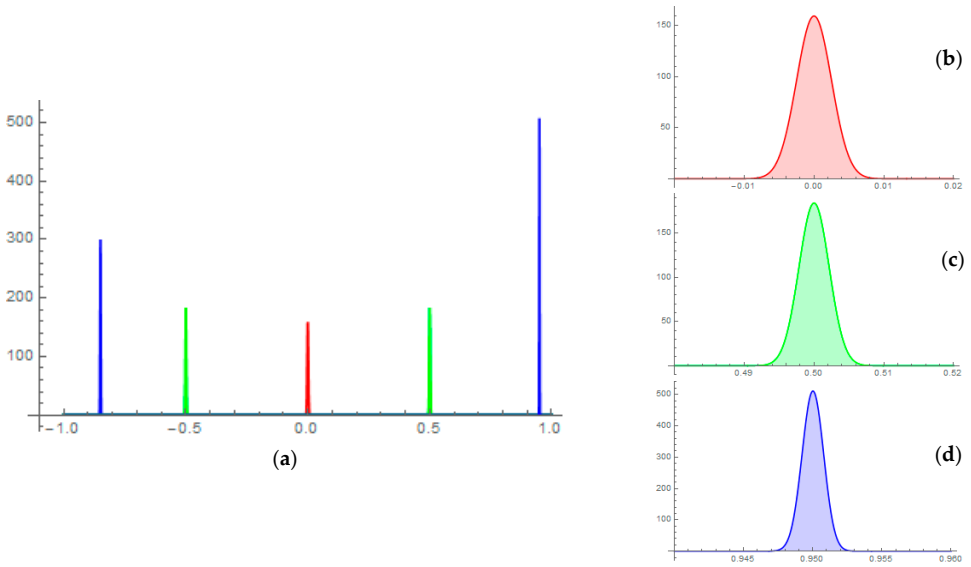


Figure 2. Specimen $\hat{\rho}$ sampling distributions for a 400-by-400 image. (**a**) across the range of ρ values, $[-1, 1]$; (**b**) zoom-in of the $\rho = 0$ sampling distribution; (**c**) zoom-in of the $\rho = 0.5$ sampling distribution; and (**d**) zoom-in of the $\rho = 0.95$ sampling distribution.

Results reported in this section allow statistical significance testing to be undertaken with regard to repeated sampling. This would be equivalent to acquiring simultaneous multiple images of the same section of the Earth's surface. In practice, it would be roughly equivalent to, say, hourly or daily images of the same section of the Earth's surface. Not surprisingly, these standard errors are extremely small for realistically-sized remotely-sensed images, ranging between 0.0009 and 0.0035 (Table 1) for the 400-by-400 dimensions treated in this paper. Most likely, this is not the primary type of uncertainty that is of interest to spatial scientists and other remote sensing researchers.

4. Sampling Experiment Designed Based Sampling Variability of the Spatial Autocorrelation Parameter

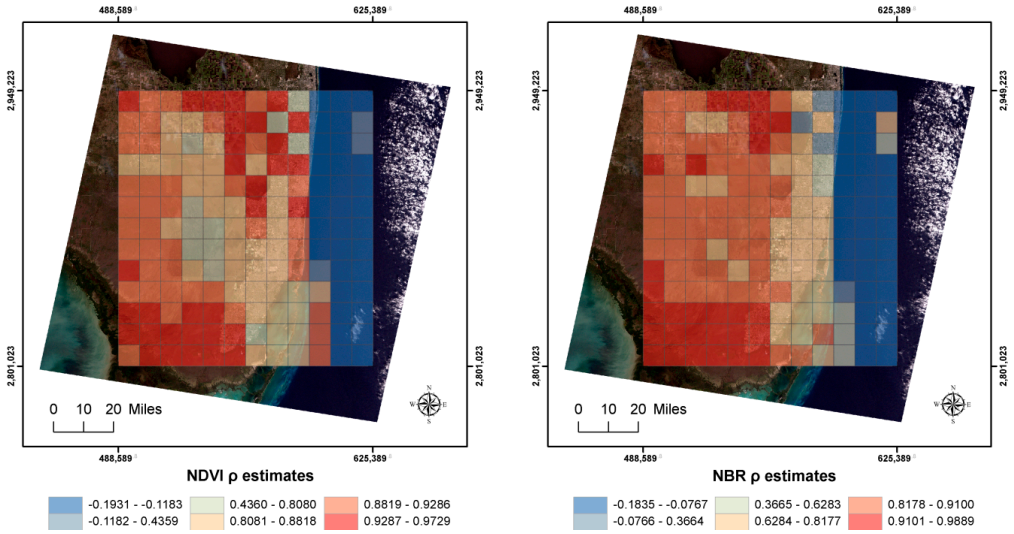
The preceding section addresses model-based inference for $\hat{\rho}$. Its uncertainty quantification weakness is that it furnishes a measure of precision for estimates using repeated images for the same Earth surface region (the population, with a single remotely-sensed image constituting a sample from this population). However, spatial scientists and researchers often are concerned about whether or not the same value of $\hat{\rho}$ would be obtained if its estimation employs different P-by-Q size images, ones that may or may not overlap.

Three sampling experiments were designed to quantify the variability of $\hat{\rho}$ across a geographic landscape, and then implemented with the Florida Everglades remotely-sensed image. The first sampling experiment partitions the 4800-by-5200 pixels image into 156 mutually exclusive and collectively exhaustive 400-by-400 sub-regions. The second experiment involves a randomly selected set of 156 400-by-400 sub-regions that were allowed to overlap. The third experiment involves increasing domain sampling, and began with an 400-by-800 central set of pixels, and successively increased them by 200 pixels in each of the four directions (i.e., 800-by-1200, 1200-by-1600, ... , 4800-by-5200), resulting in 12 $\hat{\rho}$ estimates.

4.1. Coterminous Samples

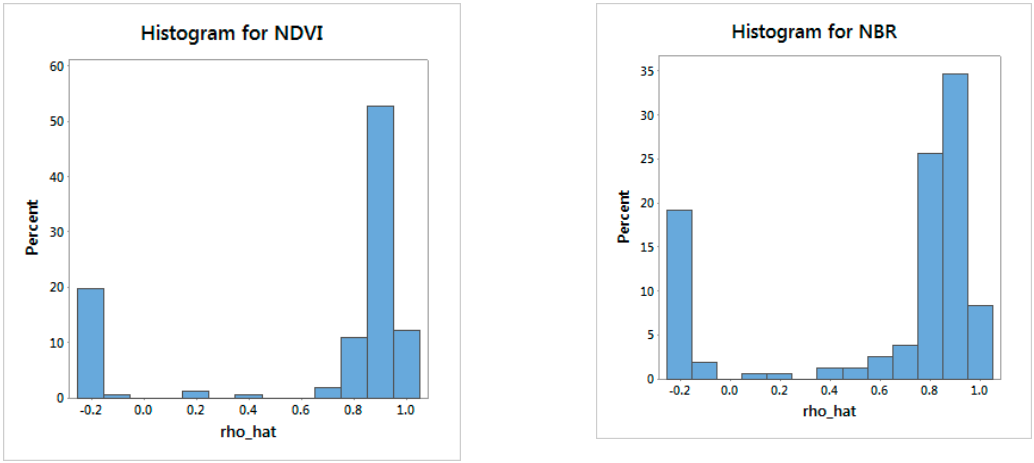
The 4800-by-5200 Florida Everglades geographic landscape was divided into 12-by-13 (i.e., 156) mutually exclusive and collectively exhaustive 400-by-400 sub-regions. Figure 3 depicts the geographic distributions of the $\hat{\rho}$ s across these sub-regions. Figures 4 and 5 indicate that the image contains two different populations: land and ocean (Figure 1). The ocean yields negative spatial autocorrelation:

32 sub-regions for NDVI, and 33 sub-regions for NBR. The respective $\hat{\rho}$ averages are 0.66 and 0.62, substantially less than their image-wide counterparts (Only 0.23% of the rook's adjacency connections are lost by analyzing the 156 sub-regions. The sub-region means display moderate-to-strong positive spatial autocorrelation: for NDVI, Moran Coefficient (MC) = 0.83, and Geary Ratio (GR) = 0.12, which indicate strong positive spatial autocorrelation; for NBR, MC = 0.68, and GR = 0.25, which indicate moderate-to-strong positive spatial autocorrelation). The respective standard errors are 0.03511 ($= 0.43857/\sqrt{156}$) and 0.03422 ($= 0.42744/\sqrt{156}$), substantially greater than their model-based counterparts. These results may be inflated because this sampling experiment is similar to cluster sampling. But each sub-region being the same size removes one source of bias in cluster sampling. Because of the mixture of two populations (landscape wide, this is conspicuous for the NDVI values, but not for the NBR values; Figure 5), this uncertainty quantification is unhelpful: the landscape-wide NDVI $\hat{\rho}$ is not in the 95% confidence interval [0.591, 0.729], and the landscape-wide NBR $\hat{\rho}$ is not in the 95% confidence interval [0.553, 0.687]. This approach also is impractical because so many subregions would need to be analyzed in order to obtain this quantification.



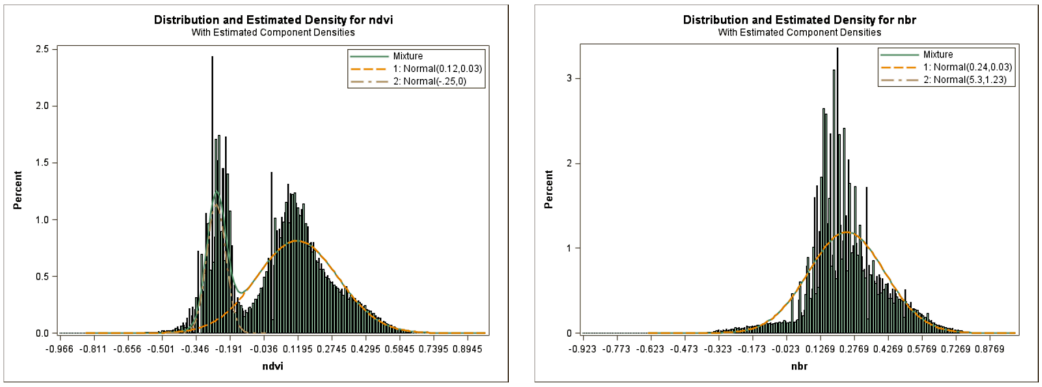
(**a**) (**b**)

Figure 3. Contour map portrayals of the geographic distribution of $\hat{\rho}$. (**a**) NDVI (MC = 0.81, GR = 0.15); and (**b**) NBR (MC = 0.83, GR = 0.14).



(**a**) (**b**)

Figure 4. The frequency distribution of $\hat{\rho}$ for the 156 coterminous quadrats partitioning the Florida Everglades region. (**a**) NDVI; and (**b**) NBR.



(**a**) (**b**)

Figure 5. Histograms of spectral index values for the Florida Everglades 4800-by-5200 pixels image. (**a**) NDVI; and (**b**) NBR.

Restricting attention to the positive $\hat{\rho}$ values, the respective estimates are 0.88 and 0.83, which are much closer to their landscape-wide counterparts. In addition, the sampling variance also decreases, with respective standard errors becoming 0.00976 and 0.01268. These latter values are much closer to their model-based counterparts, but are still noticeably greater than roughly 0.0006. A spatial scientist or other remote sensing researchers may find the 95% confidence intervals, $0.931 < \rho < 0.969$ for NDVI (i.e., markedly strong positive spatial autocorrelation) and $0.805 < \rho < 0.855$ for NBR (i.e., strong positive spatial autocorrelation), helpful because their respective landscape-wide spatial autocorrelation parameter values are such that NDVI's falls into its interval, and NBR's is close to its interval. This uncertainty quantification also allows ocean results to be differentiated from land results, but remains impractical because so many sub-regions would need to be analyzed in order to obtain it. The challenge that remains is to be able to compute these standard errors analytically, perhaps with guidance from cluster samples sampling theory (The $\hat{\rho}$ s are an inverse function of their subregion standard deviations, a specification that offers a way to establish a random variable whose variance may be useful here (see [9])).

4.2. Random Subregion Samples

Next, 156 random sub-regions were selected, such that their centers came from the central 4400-by-4800 sub-region of the Florida Everglades image, which is defined by the southwest coordinate (2101, 1401) and the northeast coordinate (6500, 6200). The UTM coordinates for the southwest and northeast corners are (494,289.8, 2,806,723) and (619,689.8, 2,943,523), respectively. Figure 6a portrays the geographic distribution of the centroids of these sub-regions; their spacing/density distribution is a Poisson RV (Figure 6b). The ocean yields negative spatial autocorrelation: 24 random sub-regions for NDVI, and 27 random sub-regions for NBR. The respective $\hat{\rho}$ averages are 0.69 and 0.64, respectively, very similar to their coterminous quadrat counterparts. The respective standard errors are 0.03203 ($= 0.40003/\sqrt{156}$) and 0.03162 ($= 0.39494/\sqrt{156}$), again very similar to their coterminous quadrat counterparts, and substantially greater than their model-based counterparts.

Restricting attention to the positive $\hat{\rho}$ values, the respective estimates are 0.85 and 0.81, which are much closer than equivalent coterminous quadrat values to their landscape-wide counterparts. In addition, the sampling variance also decreases, with respective standard errors becoming 0.0320 and 0.0316. These means are much closer to their model-based counterparts, but these standard deviations are noticeably greater than their coterminous quadrat counterparts. A spatial scientist or other remote sensing researchers may find the 95% confidence intervals, $0.824 < \rho < 0.876$ for NDVI and $0.787 < \rho < 0.833$ for NBR, less helpful because neither landscape-wide spatial autocorrelation parameter value falls into its respective interval. This uncertainty quantification seems less useful than the preceding coterminous quadrat based one. Therefore, the challenge appears to be designing

an analytical method to compute standard errors relating to mutually exclusive and collectively exhaustive coterminous quadrat sub-regions.

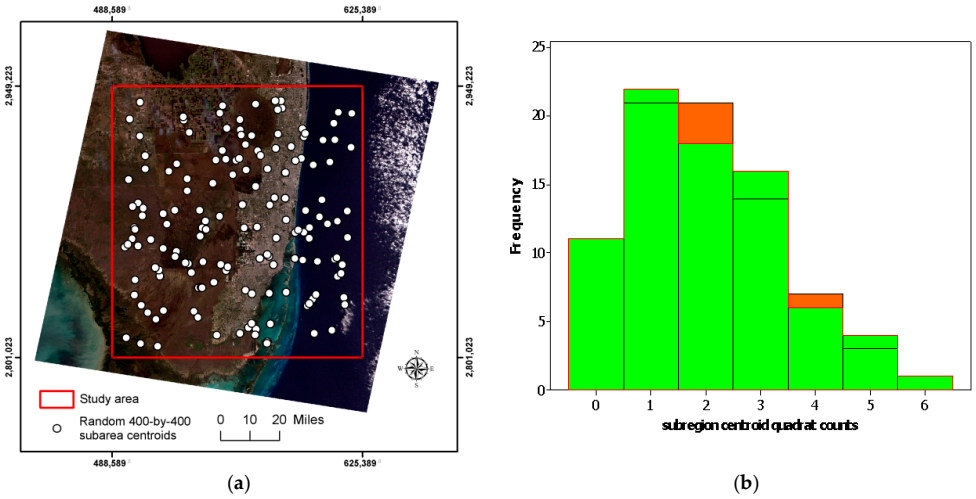


Figure 6. Features of the random sample of sub-regions. (**a**) the geographic distribution of the sub-region centroids; and (**b**) a histogram comparison between observed (green) and theoretical Poisson (red and black lines for hidden bars) quadrat counts.

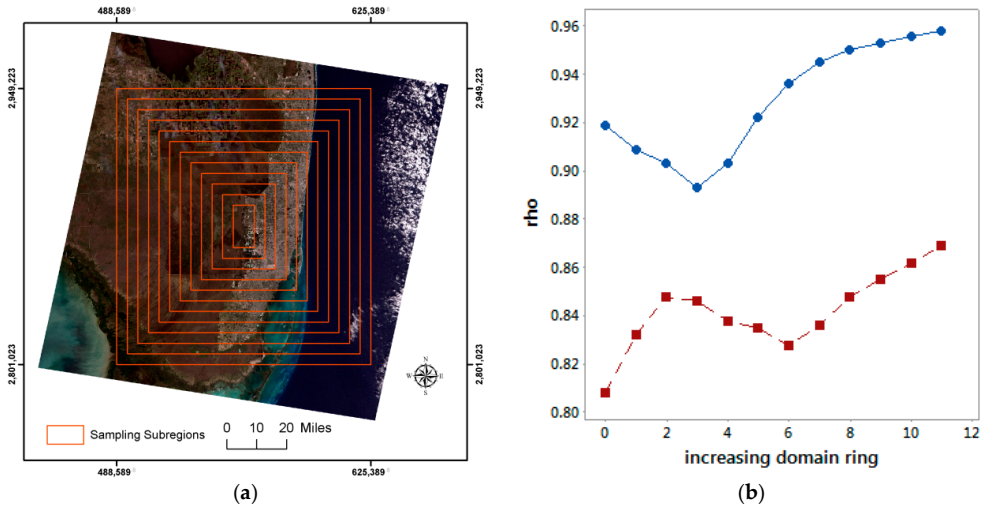
Removing the coterminous constraint on the data, an experiment was performed involving 156 samples of size 160,000, where the sample pixels were randomly selected from the set of 24,960,000 pixels with equal probability but without replacement; the spatial lag term was calculated with the four surrounding pixels, and hence was selected as part of a bivariate sample. The Jaocbian term is not correct here. Nevertheless, the average $\hat{\rho}$ s obtained with ordinary least squares (OLS) are 1.006 and 0.995, respectively, for NDVI and NBR; the first is not a feasible value (not only does this arithmetic average exceed 1, but all 156 estimates exceed 1). Retaining the original Jacobian term yields, respectively, $\hat{\rho}$ s of 0.958 and 0.869 (these equal their respective estimates for the full 4800-by-5200 image). Their corresponding standard errors are 0.00056 and 0.00105, which, again, suggests that repeated sets of 156 random samples of size 160,000 will yield sound complete image parameter estimates.

Consequently, random sub-regions furnish useful measures of uncertainty, but random pixels do not. Unfortunately, in practice, such repeated sampling places a considerable data collection and computational burden on a spatial analyst. Once more, the challenge that remains is to be able to compute these standard errors analytically, again perhaps with guidance from cluster samples sampling theory.

4.3. Increasing Domain Subregions

A third possible approach to quantifying uncertainty of $\hat{\rho}$ is to study how this estimate changes with increasing geographic landscape size (i.e., scale; Figure 7a). Figure 7b portrays the change in the Everglades NDVI and NBR $\hat{\rho}$ s with an increasing domain. Their respective unweighted averages are 0.929 and 0.842 (Table 2), again indicating some bias. Meanwhile, their respective standard errors are 0.02353 and 0.01635, which are smaller than those obtained with coterminous or random cluster sampling. These standard errors are still substantially larger than their counterpart values furnished by Equation (4).

Therefore, an increasing domain furnishes somewhat useful measures of uncertainty. Unfortunately, in practice, this approach also places a considerable data collection and computational burden on a spatial analyst. As before, the challenge that remains is to be able to compute these standard errors analytically.



(**a**) (**b**)

Figure 7. An increasing domain sampling design. (**a**) the increasing size landscapes; and (**b**) $\hat{\rho}$ with an increasing distance from the core region (Ring 0).

Table 2. Increasing domain results.

Ring	Dimension	Jacobian Approximation		$\hat{\rho}$	
		Alpha	Delta	NDVI	NBR
inner core	400-by- 800	0.1618076	1.1697864	0.919	0.808
1	800-by-1200	0.1618659	1.1704490	0.909	0.832
2	1200-by-1600	0.1618877	1.1706978	0.903	0.848
3	1600-by-2000	0.1618994	1.1708305	0.893	0.846
4	2000-by-2400	0.1619066	1.1709135	0.903	0.838
5	2400-by-2800	0.1619116	1.1709704	0.922	0.835
6	2800-by-3200	0.1619152	1.1710118	0.936	0.828
7	3200-by-3600	0.1619182	1.1710438	0.945	0.836
8	3600-by-4000	0.1619204	1.1710687	0.950	0.848
9	4000-by-4400	0.1619221	1.1710888	0.953	0.855
10	4400-by-4800	0.1619236	1.1711054	0.956	0.862
outer	4800-by-5200	0.1619248	1.1711194	0.958	0.869

5. Discussion

Two points merit discussion. First, because remotely-sensed data comprising spectral measures contains considerable redundant information, quantified by the spatial autocorrelation parameter of a spatial regression model, the equivalent amount of independent information is of interest. Second, because direct and indirect effects associated with a spatial process generate layer upon layer of stochastic noise similar to compounding of monetary interest, the degree of variance inflation affiliated with spatial data is of interest.

5.1. Effective Sample Size

One impact of $\hat{\rho}$ on uncertainty measures is the exaggeration of a sample size, n, which needs to be reduced to its effective sample size, n*, the equivalent sample size for the amount of nonredundant information contained in georeferenced data. Griffith [10] (p. 743) furnishes the following estimate of n*:

$$n^* = n\left[1 - \frac{1}{1 - e^{-1.92369}}\frac{n-1}{n}\left(1 - e^{-2.12373\hat{\rho}+0.20042\sqrt{\hat{\rho}}}\right)\right] \quad (5)$$

Equation (5) is only for $\hat{\rho} \geqslant 0$.

Table 3 summarizes selected effective sample size results for the Florida Everglades remotely-sensed image. Summaries are based upon fewer than 156 coterminous quadrat results because estimates of $\hat{\rho}$ for the water areas are negative. NDVI contains a higher level of spatial autocorrelation than NBR, and consequently tends to yield a smaller n*. The random quadrat and increasing domain samples are reasonably consistent, with relatively small sample variation. The general finding here is that n* is roughly 5% of n: the 400-by-400 pixel quadrats contain about as much nonredundant information as 8000 independent pixels, and the entire set of 24,960,000 pixels in the image contains about as much nonredundant information as 1,248,000 independent pixels.

Table 3. Effective sample size percentage, 100 n*/n, for the Florida Everglades remotely-sensed image.

Summary Statistic	Coterminous Quadrats		Random Quadrats		Increasing Domain	
	NDVI	NBR	NDVI	NBR	NDVI	NBR
Mean	5.48	8.21	3.06	8.36	2.67	6.45
Standard deviation	8.33	10.96	0.03	0.06	0.94	0.78
Minimum	0.97	0.40	2.98	8.22	1.52	5.19
Maximum	64.42	85.40	3.14	8.50	4.14	8.11
# sample quadrats	124	123	156	156	12	12

NOTE: coterminous and random quadrats form 400-by-400 (n = 160,000) square regions; increasing domain quadrats begin with 400-by-800, and increase, by increments of 400 in both dimensions, to 4800-by-5200.

5.2. The Variance Inflation Factor (VIF)

A VIF is also of concern when assessing uncertainty, especially with $\hat{\rho}$ values near 0.9. The bias-adjusted maximum likelihood estimate of the variance is given by:

$$(\mathbf{Y} - \hat{\mu}1)^{\mathrm{T}}(\mathbf{I} - \hat{\rho}\mathbf{W})^{\mathrm{T}}(\mathbf{I} - \hat{\rho}\mathbf{W})(\mathbf{Y} - \hat{\mu}1)/(\mathrm{n} - 2) \tag{6}$$

The VIF is given by:

$$\mathrm{VIF} = \frac{\mathrm{n} - 1}{\mathrm{n} - 2}(\mathbf{Y} - \hat{\mu}1)^{\mathrm{T}}(\mathbf{I} - \hat{\rho}\mathbf{W})^{\mathrm{T}}(\mathbf{I} - \hat{\rho}\mathbf{W})(\mathbf{Y} - \hat{\mu}1)/[(\mathbf{Y} - \hat{\mu}1)^{\mathrm{T}}(\mathbf{Y} - \hat{\mu}1)] \tag{7}$$

Table 4 summarizes VIF results for the Everglades data. The NDVI VIF tends to be a factor of about 5, whereas the NBR VIF tends to be a factor of about 4. In other words, confidence intervals calculated with variance estimates from the unadjusted remotely sensed data tend to be about twice as wide as they should be. The implication is considerably more uncertainty attributable to sampling error than appears to exist, if latent spatial autocorrelation is overlooked.

In summary, two features of data commonly used to quantify uncertainty, namely sample size and variability, are distorted by the presence of non-zero spatial autocorrelation. This distortion becomes considerable with the high levels of positive spatial autocorrelation commonly found in remotely-sensed data. Therefore, accounting for spatial autocorrelation in remotely-sensed data contributes to a better extraction of information and distilling of knowledge from such data.

Table 4. The square root of VIF factors for the Florida Everglades remotely-sensed image.

Summary Statistic	Coterminous Quadrats		Random Quadrats		Increasing Domain	
	NDVI	NBR	NDVI	NBR	NDVI	NBR
Mean	2.45	2.20	2.12	1.83	2.90	1.89
Standard deviation	0.77	0.98	0.82	0.81	0.59	0.13
Minimum	1.02	1.00	1.00	1.00	2.21	1.65
Maximum	4.69	5.40	4.59	5.07	3.94	2.13
# sample quadrats	124	123	156	156	12	12

6. Conclusions and Implications

Uncertainty associated with remotely-sensed data should be cast in terms of its latent spatial autocorrelation, denoted here by $\hat{\rho}$. Since this measure tends to be in the range [0.85, 0.95] for most remotely-sensed data, it indicates that these data contain considerable redundant information which, in turn, means the presence of substantial VIFs. The standard method of quantifying any additional uncertainty introduced by estimating ρ appears to be negligible, as indicated by Equation (4) for $\rho = 0$ (which essentially is irrelevant for remotely sensed data), and indicted by both Equation (3) and the BBRV mixture proposed in this paper for $\rho \neq 0$.

Most geographic landscapes suggest that Equation (3) fails to furnish an adequate quantification of uncertainty for $\hat{\rho}$; its variability in a geographic landscape is substantially greater than values produced by Equation (3). The various analyses summarized in this paper for the Florida Everglades remotely-sensed image corroborate this contention. This inconsistency arises from the inferential basis for Equation (3): repeated images of the same region of the Earth's surface. Treating different regions introduces another source of variability into $\hat{\rho}$. This is the uncertainty of interest to most spatial analysts. Would a slightly larger/smaller image, an image shifted slightly to the east and/or north, or an image from a different part of a larger geographic landscape yield essentially the same $\hat{\rho}$? The variability detected in analyses summarized in this paper reveals quantities many times larger than those rendered by Equation (3).

The remaining challenge is to formulate an analytical specification of this type of variability. Evidence presented in this paper suggests that cluster sample sampling theory may furnish insights into what this specification should be. The final formula should be a function of $\hat{\rho}$, the variance of the spectral index being studied, the variance of the spatial autocorrelation filtered data (i.e., what is equivalent to independent and identically distributed values), and the sample size, n. Establishing such a formula will support a better extraction of information and distilling of knowledge from remotely-sensed data by, especially, accounting for their latent spatial autocorrelation.

Finally, spatial regression should be utilized in studies that utilize remote sensing data. Regression is extensively used to model various phenomena such as land use and land cover [4,11], NDVI [12,13], urban heat island [4], and landslide susceptibility [14], but spatial autocorrelation has been barely accommodated in modeling remotely-sensed data. Remotely-sensed data has a strong positive spatial autocorrelation in most cases: even one with a fragmented (e.g., land use) pattern with a coarse resolution (e.g., 250 m of MODIS). Ignorance of spatial autocorrelation can result in unreliable results.

Author Contributions: Both authors contributed equally to the paper.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

A simulation experiment was conducted to examine the uncertainty of the spatial autocorrelation parameter estimates ($\hat{\rho}$) for a series of regular tessellations with an increasing number of spatial units (i.e., the number of pixels comprising a remotely sensed image). A total of 40 different regular square tessellations were used, from a 10-by-10 to a 400-by-400 tessellation, using increments of 10. For each of these tessellations, 1000 sets of random numbers were drawn from the standard normal distribution; i.e., N(0,1). The spatial autocorrelation parameter for the SAR process (ρ) was estimated with the Monte Carlo eigenvalue method developed by Barry and Pace [15] (these results were double-checked with exact eigenvalues through the 100-by-100 tessellation, as well as with Griffith's eigenvalue and approximation Jacobian approximations [3]). This estimation is available in R: errorsarlm() function in the *spdep* package with the "MC" option. Table A1 reports the results of $\hat{\rho}$. The means of the estimates are around zero; all estimates except those for the 10-by-10 tessellation are 0.00 when they are rounded to hundredths. Their standard deviations tend to get smaller as the tessellation size gets larger. The minimum and maximum values of the estimates also get smaller as the tessellation size gets larger. That is, the uncertainty of $\hat{\rho}$ gets smaller as the size of an image gets larger.

Table A1. Estimates of the spatial autocorrelation parameter (ρ) for increasing tessellation size based upon 1000 simulation replications and the standard normal distribution.

Tessellation	Mean	Std. Dev.	Min	Max	Tessellation	Mean	Std. Dev.	Min	Max
10-by-10	−0.0144	0.1299	−0.4329	0.3742	210-by-210	−0.0002	0.0068	−0.0218	0.0215
20-by-20	−0.0018	0.0683	−0.2101	0.2079	220-by-220	−0.0004	0.0063	−0.0191	0.0259
30-by-30	0.0013	0.0448	−0.1570	0.1384	230-by-230	−0.0001	0.0062	−0.0194	0.0233
40-by-40	−0.0017	0.0343	−0.1129	0.1175	240-by-240	0.0002	0.0058	−0.0230	0.0193
50-by-50	−0.0032	0.0263	−0.0898	0.0985	250-by-250	0.0002	0.0056	−0.0183	0.0203
60-by-60	−0.0005	0.0233	−0.0810	0.0655	260-by-260	0.0003	0.0054	−0.0202	0.0170
70-by-70	−0.0012	0.0203	−0.0690	0.0629	270-by-270	−0.0002	0.0052	−0.0150	0.0143
80-by-80	0.0001	0.0178	−0.0610	0.0544	280-by-280	0.0000	0.0050	−0.0173	0.0175
90-by-90	−0.0005	0.0153	−0.0476	0.0678	290-by-290	−0.0003	0.00482	−0.0168	0.0132
100-by-100	0.0011	0.0142	−0.0397	0.0471	300-by-300	0.0000	0.00478	−0.0132	0.0151
110-by-110	−0.0004	0.0126	−0.0432	0.0435	310-by-310	−0.0004	0.0045	−0.0156	0.0136
120-by-120	−0.0002	0.0121	−0.0356	0.0313	320-by-320	−0.0001	0.0043	−0.0146	0.0134
130-by-130	−0.0003	0.0109	−0.0383	0.0360	330-by-330	0.0000	0.0042	−0.0133	0.0150
140-by-140	−0.0004	0.0103	−0.0283	0.0342	340-by-340	0.0001	0.0041	−0.0122	0.0116
150-by-150	−0.0001	0.0093	−0.0284	0.0285	350-by-350	0.0002	0.0040	−0.0130	0.0141
160-by-160	−0.0002	0.0086	−0.0263	0.0230	360-by-360	−0.0002	0.00383	−0.0126	0.0141
170-by-170	−0.0005	0.0084	−0.0260	0.0327	370-by-370	−0.0001	0.00377	−0.0128	0.0124
180-by-180	0.0001	0.0078	−0.0255	0.0280	380-by-380	0.0001	0.0036	−0.0116	0.0118
190-by-190	0.0005	0.0077	−0.0251	0.0259	390-by-390	0.0001	0.0036	−0.0117	0.0132
200-by-200	−0.0003	0.0071	−0.0234	0.0278	400-by-400	−0.0001	0.0035	−0.0113	0.0124

The estimate $\hat{\rho}$ also was examined with higher levels of spatial autocorrelation only for a 400-by-400 tessellation. A pure SAR random process was used to generate spatially autocorrelated random number sets: i.e., with no covariates and a zero intercept. Due to the large size of a spatial weights matrix for the tessellation (i.e., 400^2-by-400^2), the calculation of $(\mathbf{I} - \rho\mathbf{W})^{-1}$ is infeasible. Hence, the matrix expansion of this inverse, the summation of a power series, $(\mathbf{I} + \rho\mathbf{W} + \rho\mathbf{W}^2 + \ldots + \rho\mathbf{W}^k)$, was used with k = 1000 to approximate it. Table A2 reports the estimation results for ρ = 0.0, 0.5, 0.85, 0.90, and 0.95. The means of the $\hat{\rho}$ s are approximately the nominal levels, and their standard deviations get smaller as the level of spatial autocorrelation gets larger.

Table A2. Estimates of the spatial autocorrelation parameter (ρ) for selected high levels of spatial autocorrelation and a 400-by-400 tessellation based upon 1000 simulation replications and the standard normal distribution.

ρ	Mean of Estimates ($\bar{\hat{\rho}}$)	Std. Dev.	Min	Max
0.00	−0.0001	0.0035	−0.0113	0.0124
0.50	0.5000	0.0029	0.4915	0.5096
0.85	0.8501	0.0015	0.8453	0.8545
0.90	0.9007	0.0012	0.8965	0.9044
0.95	0.9533	0.0009	0.9497	0.9560

References

1. Griffith, D.A. Positive spatial autocorrelation impacts on attribute variable frequency distributions. *Chil. J. Stat.* **2011**, *2*, 3–28.
2. Cliff, A.; Ord, K. Testing for spatial autocorrelation among regression residuals. *Geogr. Anal.* **1972**, *4*, 267–284. [CrossRef]
3. Anselin, L. *Spatial Econometrics: Methods and Models*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1988.
4. Chen, X.-L.; Zhao, H.-M.; Li, P.-X.; Yin, Z.-Y. Remote sensing image-based analysis of the relationship between urban heat island and land use/cover changes. *Remote Sens. Environ.* **2006**, *104*, 133–146. [CrossRef]
5. Griffith, D.A. Approximation of Gaussian spatial autoregressive models for massive regular square tessellation data. *Int. J. Geogr. Inf. Sci.* **2015**, *29*, 2143–2173. [CrossRef]

6. Mardia, K.; Watkins, A. On multimodality of the likelihood in the spatial linear model. *Biometrika* **1989**, *76*, 289–295. [CrossRef]
7. Griffith, D.A. Methods: Spatial autocorrelation. In *International Encyclopedia of Human Geography*; Kitchin, R., Thrift, N., Eds.; Elsevier: New York, NY, USA, 2009; pp. 396–402.
8. Ord, J. Estimation methods for models of spatial interactions. *J. Am. Stat. Assoc.* **1975**, *70*, 120–126. [CrossRef]
9. Shen, A. On asymptotic approximation of inverse moments for a class of nonnegative random variables. *Statistics* **2014**, *48*, 1371–1379. [CrossRef]
10. Griffith, D.A. Effective geographic sample size in the presence of spatial autocorrelation. *Ann. Assoc. Am. Geogr.* **2005**, *95*, 740–760. [CrossRef]
11. Brown, D.G.; Pijanowski, B.C.; Duh, J.D. Modeling the relationships between land use and land cover on private lands in the Upper Midwest, USA. *J. Environ. Manag.* **2000**, *59*, 247–263. [CrossRef]
12. Holm, A.M.; Cridland, S.W.; Roderick, M.L. The use of time-integrated NOAA NDVI data and rainfall to assess landscape degradation in the arid shrubland of Western Australia. *Remote Sens. Environ.* **2003**, *85*, 145–158. [CrossRef]
13. Wessels, K.J.; Prince, S.D.; Zambatis, N.; Macfadyen, S.; Frost, P.E.; van Zyl, D. Relationship between herbaceous biomass and 1 km^2 Advanced Very High Resolution Radiometer (AVHRR) NDVI in Kruger National Park, South Africa. *Int. J. Remote Sens.* **2006**, *27*, 951–973. [CrossRef]
14. Lee, S. Application of logistic regression model and its validation for landslide susceptibility mapping using GIS and remote sensing data. *Int. J. Remote Sens.* **2005**, *26*, 1477–1491. [CrossRef]
15. Barry, R.; Pace, R. Monte Carlo estimates of the log determinant of large sparse matrices. *Linear Algebra Appl.* **1999**, *289*, 41–54. [CrossRef]