7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences.
Edited by M. Caetano and M. Painho.

Statistical efficiency of model-informed geographic sampling designs

Daniel A. Griffith

School of Social Sciences,
University of Texas @ Dallas, Richardson, Texas, USA, P.O. Box 830688, GR31, 75083-0688
Tel. : + 001 972 883 4950; Fax : + 001 883 6297
dagriffith@utdallas.edu

Abstract

As spatial autocorrelation latent in georeferenced data increases, the amount of duplicate information contained in these data also increases, whether an entire population or some type of random sample drawn from that population is being analyzed, resulting in incorrect sample size calculations being given by conventional power and sample size calculation formulae. Griffith (2005) exploits this context to formulate equations for estimating the necessary sample size needed to obtain some predetermined level of precision for an analysis of georeferenced data when implementing a tessellation stratified random sampling design, labeling this approach model-informed, since a model of latent spatial autocorrelation is required. Spatial autocorrelation is accounted for in these power and sample size calculation equations by using the following spatial statistical model specifications: (1) simultaneous autoregressive; (2) geostatistical semivariogram; and, (3) spatial filter. Sample size results are somewhat sensitive to which model is employed to capture spatial autocorrelation effects. This paper addresses issues of efficiency associated with each of these models in the presence of spatial autocorrelation effects. It summarizes results from a set of simulation experiments following experimental design guidelines spelled out by Overton and Stehman (1993) that explore continuous linear, quadratic, and sinusoidal response surfaces.

Keywords: autoregressive model, geostatistical model, spatial autocorrelation, spatial filter model, spatial sampling

1 Introduction

Spatial autocorrelation arises from duplicate information materialized in data that share commonalities because of their relative closeness in geographic space. This redundant information causes observations to be dependent, rather than independent, moving data analysis away from the classical statistical independence model. To date, three model specifications have been formulated to account for spatial autocorrelation, namely the autoregressive, the geostatistical semivariogram, and the spatial filter model. Griffith (2005) derives the effective geographic sample size, or equivalent value of n for independent observations, in the presence of spatial autocorrelation. In turn, he presents model-informed geographic sample size determination based upon each of these three model specifications that is necessary for achieving prespecified levels of power and precision.

Denote the sample size by n, the equivalent variance for independent and identically distributed (iid) error values by $\sigma_{e^*}^2$, the Type I error (i.e., rejecting the null hypothesis when it is true) probability for a two-tailed test by $z_{1-\alpha/2}$, the Type II error probability by $z_{1-\beta}$,

and the maximum tolerable precision by $\Delta = |\mu - \mu_o|$, where μ and μ_o respectively denoting the null and the alternate hypothesis means. The problem is to determine a value of n that allows a predetermined, desired level of statistical precision to be obtained for an analysis in the presence of spatial autocorrelation. Accordingly, for the spatial simultaneous autoregressive (SAR) model specification, n is given by

$$\frac{\sigma_{e*}^2 \dfrac{(z_{1-\alpha/2} + z_{1-\beta})^2}{\Delta^2} - \dfrac{1 - e^{-2.12373\hat{\rho} + 0.20024\sqrt{\hat{\rho}}}}{1 - e^{-1.92349}}}{1 - \dfrac{1 - e^{-2.12373\hat{\rho} + 0.20024\sqrt{\hat{\rho}}}}{1 - e^{-1.92349}}} \tag{1}$$

where ρ denotes the spatial autoregressive parameter. For the geostatisticsal semivariogram model specification, n is given by

$$1 + \left[(C_0 + C_1) \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{\Delta^2} - 1 \right] (1 + b \frac{r}{d_{\max}})^c \tag{2}$$

where C_0 and C_1 are parameters of a particular estimated semivariogram model, r is the semivariogram model range, d_{max} is the maximum distance in a geographic landscape, and b and c are coefficients associated with sampling intensity. And, for the eigenfunction-based spatial filter model specification, n is given by

$$\frac{\sigma_{e*}^2 \dfrac{(z_{1-\alpha/2} + z_{1-\beta})^2}{\Delta^2}}{1 - R^2} \tag{3}$$

where the R^2 value is the multiple correlation coefficient obtained by regressing a response variable on a judiciously selected set of eigenvectors.

The primary purpose of this paper is to summarize a set of simulation experiments that assesses efficiency associated with the use of each of these models to inform geographic sampling.

2 The simulation experimental design

Overton and Stehman (1993) outline superpopulation experiments in which they employ a planar, a quadratic, and a sinusoidal response variable surface. These are the three surfaces utilized here to represent spatial continuity, or the manifestation of spatial autocorrelation. The geographic landscape is a unit square that is partitioned into $\sqrt{n} \times \sqrt{n}$ equal-sized squares; this partitioning differs from the Overton and Stehman (1993) and Griffith (2005) hexagonal tessellation in its surface partitioning geometric form, but without loss of generality, and is used to simplify analytical and simulation calculations. A single random point is sampled from each square.

Denote georeferencing coordinates for a sample point by (u,v), and its accompanying iid error term by $\varepsilon(u, v) \sim N(0, \sigma_\varepsilon^2)$; then the geographic distribution of an attribute variable may be written as

$$\underline{\text{planar}}: y(u,v) = \alpha + \beta_u u + \beta_v v + \varepsilon(u, v) \quad (4)$$

$$\underline{\text{quadratic}}: y(u,v) = \alpha + \beta_u u + \beta_v v + \beta_{u^2} u^2 + \beta_{v^2} v^2 + \beta_{uv} uv + \varepsilon(u, v) \quad (5)$$

$$\underline{\text{sinusoidal}}: y(u,v) = \alpha + \sqrt{3}\,[\text{SIN}(2\pi\beta_u d\, u) + \text{SIN}(2\pi\beta_v d\, v)] + \varepsilon(u, v) \quad (6)$$

where

α is an intercept term,

β_j are slope coefficients.

The analytical mean (μ) and variance (σ^2) derived for these equations is as follows:

$\underline{\text{planar}}$: $\mu = \alpha + \dfrac{\beta_u + \beta_v}{2}$ and $\sigma^2 = \sigma_\varepsilon^2 + \dfrac{\beta_u^2 + \beta_v^2}{12}$,

$\underline{\text{quadratic}}$: $\mu = \alpha + \dfrac{\beta_u + \beta_v}{2} + \dfrac{\beta_{u^2} + \beta_{v^2}}{3} + \dfrac{\beta_{uv}}{4}$ and

$$\sigma^2 = \sigma_\varepsilon^2 + \frac{\beta_u^2 + \beta_v^2 + \beta_{uv}(\beta_u + \beta_{u^2} + \beta_v + \beta_{v^2})}{12} + \frac{\beta_u\beta_{u^2} + \beta_v\beta_{v^2}}{6} + \frac{4(\beta_{u^2}^2 + \beta_{v^2}^2)}{45} + \frac{7\beta_{uv}^2}{144},$$

and

$\underline{\text{sinusoidal}}$: $\mu = \alpha + \sqrt{3}\left[\dfrac{1 - \text{COS}^2(\pi\beta_u d)}{\pi\beta_u d} + \dfrac{1 - \text{COS}^2(\pi\beta_v d)}{\pi\beta_v d}\right]$ and, for $\beta_u = \beta_v$,

$$\sigma^2 = \sigma_\varepsilon^2 + 3\frac{\pi\beta_u d\,\text{SIN}(\pi\beta_u d)\,\text{COS}(\pi\beta_u d)[1 - 2\text{COS}^2(\pi\beta_u d)] - 2[1 - \text{COS}^2(\pi\beta_u d)]^2}{(\pi\beta_u d)^2},$$

where

$d = 1/\sqrt{n}$.

Variance inflation induced by spatial correlation equals $\sigma^2 - \sigma_\varepsilon^2$.

The simulation experiments involve 10-by-10 and 20-by-20, regular square tessellations. Spatial correlation variance components are 25%, 50% and 90% (i.e., increasing degrees of positive spatial autocorrelation), with σ_ε^2 = 1. A row standardized version of the n-by-n geographic connectivity matrix is coupled with the SAR model. Maximum distance for

semivariogram modeling is set to $\sqrt{n}/2$ (i.e., half of the maximum distance in the unit square). Only eigenvectors whose corresponding Moran Coefficient divided by the maximum possible Moran Coefficient [i.e., $2COS(\frac{\pi}{1+\sqrt{n}}) + 2COS(\frac{2\pi}{1+\sqrt{n}})$] exceeds 0.25 are candidates for stepwise regression selection. The same random variable value realizations are analyzed with each model specification. And, each simulation in the experiment is replicated 500 times; the number of replications is constrained by the need to execute a massive number of matrix inversions.

3 Numerical intensities

Table 1 SAR model estimation results for the simulation experiments, r = 500.

Equation	SA[a]	With spatial		Without spatial		$\hat{\rho}$	$\hat{\sigma}_{\hat{\rho}}$
		S-W[b]	P(S-W)[c]	S-W[b]	P(S-W)[c]		
10-by-10 regular square tessellation sampling grid							
linear	0.25	0.98670	0.50275	0.98659	0.50071	0.32373	0.11591
	0.50	0.98728	0.52734	0.98680	0.50957	0.57193	0.07410
	0.90	0.98791	0.53685	0.98679	0.50935	0.89244	0.01948
quadratic	0.25	0.98668	0.50634	0.98667	0.50894	0.31645	0.12133
	0.50	0.98605	0.47030	0.98640	0.48977	0.56486	0.07827
	0.90	0.97825	0.16954	0.98599	0.48296	0.89153	0.01913
sinusoidal	0.25	0.98573	0.46824	0.98600	0.48405	0.31631	0.12221
	0.50	0.98668	0.51128	0.98639	0.48994	0.55866	0.07919
	0.90	0.97976	0.20227	0.98637	0.48254	0.88573	0.02188
20-by-20 regular square tessellation sampling grid							
linear	0.25	0.99594	0.47809	0.99591	0.47691	0.34912	0.059059
	0.50	0.99601	0.47938	0.99601	0.49020	0.58619	0.036776
	0.90	0.99372	0.15871	0.99611	0.51125	0.89537	0.009236
quadratic	0.25	0.99589	0.47992	0.99599	0.48339	0.34967	0.058248
	0.50	0.99469	0.32250	0.99605	0.50946	0.58636	0.036831
	0.90	0.98417	0.00163	0.99599	0.49276	0.89615	0.009344
sinusoidal	0.25	0.99596	0.48924	0.99589	0.47928	0.34809	0.056283
	0.50	0.99568	0.43244	0.99601	0.49452	0.58858	0.034564
	0.90	0.98524	0.00129	0.99594	0.48577	0.89594	0.008771

[a] The percentage of variance attributable to spatial autocorrelation.

[b] The Shapiro-Wilk normality diagnostic statistic.

[c] Probability of the S-W statistic under the null hypothesis of normality.

All three model specifications involve numerical intensities. Estimation of the SAR model requires nonlinear regression coupled with the calculation of a numerically demanding Jacobian term; the Jacobian approximation derived by Griffith (2004) was employed here. Estimation of the semivariogram model, with the spherical specification employed here, requires nonlinear weighted least squares coupled with the calculation of $(n-1)^2/2$ paired comparisons, followed by the calculation and then inversion of an n-by-n matrix. And, estimation of the spatial filter model requires stepwise linear regression to select a subset of geographic connectivity matrix eigenvectors from a relatively large candidate set.

Simulation experiment estimates of the spatial autoregressive parameter—ρ—appearing in equation (1) are summarized in Table 1. The original normality of the data only appears to be somewhat compromised with relatively high levels of positive spatial autocorrelation. The spatial autoregressive parameter estimates index well their corresponding levels of positive spatial autocorrelation. In addition, as with a Pearson product moment correlation coefficient, the standard error of these estimates decreases as the boundary of the feasible parameter space, namely 1, is approached.

A number of semivariogram models could be explored to see which best describe the data. Griffith (2005) suggests that the spherical model furnishes a good description of weak positive spatial autocorrelation, the exponential model furnishes a good description of moderate positive spatial autocorrelation, and the Bessel function model furnishes a good description of strong positive spatial autocorrelation. Exploratory analysis of these particular models cross a distance of 0.20 units revealed that the spherical model performs acceptably well here, and for simplicity is used in all cases here. Distances were grouped into 30 equal-sized bins for estimation purposes. Simulation experiment estimates of the spherical semivariogram model parameters—C_0, C_1 and r —appearing in equation (2) are summarized in Table 2. As is expected, the intercept term, C_0, is approximately 0. This geostatistical analysis furnishes the sill as a measure of the variance, or $\sqrt{C_0 + C_1}$ as a measure of the standard deviation. The estimated range approaches two-thirds of the maximum distance in the most extreme case.

Table 2 Seimvariogram model estimation results for the simulation experiments, r = 500.

Equation	SA[a]	C_0	$\hat{\sigma}_{C_0}$	C_1	$\hat{\sigma}_{C_1}$	r	$\hat{\sigma}_r$
10-by-10 regular square tessellation sampling grid							
linear	0.25	0.00357	0.01079	1.0251	0.15796	0.04467	0.03579
	0.50	0.00511	0.01724	1.0761	0.30785	0.05295	0.09692
	0.90	0.06334	0.07327	3.4696	4.08935	0.41060	0.64376
quadratic	0.25	0.00335	0.00954	1.0289	0.14779	0.04411	0.03484
	0.50	0.00558	0.01765	1.0699	0.15929	0.05002	0.04244
	0.90	0.07142	0.07612	3.5763	4.18647	0.42267	0.63725
sinusoidal	0.25	0.00335	0.01003	1.0326	0.15182	0.04531	0.03668
	0.50	0.00485	0.01749	1.0814	0.16069	0.04954	0.04609
	0.90	0.07303	0.07788	10.8253	8.20114	1.09143	0.87032
20-by-20 regular square tessellation sampling grid							
linear	0.25	0.00396	0.009569	1.0144	0.07502	0.01770	0.01395
	0.50	0.00676	0.014655	1.0596	0.07905	0.02171	0.01555
	0.90	0.50740	0.061537	6.0520	3.85521	1.17680	0.76528
quadratic	0.25	0.00432	0.009826	1.0189	0.07451	0.01842	0.01395
	0.50	0.00714	0.014325	1.0584	0.07364	0.02198	0.01603
	0.90	0.49936	0.070000	6.1515	3.90050	1.17306	0.76092
sinusoidal	0.25	0.00426	0.010383	1.0146	0.07496	0.01766	0.01409
	0.50	0.00697	0.015585	1.0500	0.07629	0.02155	0.01556
	0.90	0.30857	0.071322	16.6784	1.59804	1.97968	0.16055

[a] The percentage of variance attributable to spatial autocorrelation.

Simulation experiment estimates of the spatial filter model are summarized in Table 3, and reveal that, for the example of n = 100, only 23 of 31 candidate eigenvectors with a relative Moran Coefficient exceeding 0.25 were ever selected to describe at least one of the simulated data sets. The corresponding R^2 value appearing in equation (3) furnishes a good index of spatial autocorrelation. The n = 100 example also illustrates that the maximum spatial autocorrelation represented by the first eigenvector virtually always plays an important role in describing underlying spatial continuity, which is expected here because of the gradient models used to embed spatial autocorrelation.

Table 3 Spatial filter model estimation results for the simulation experiments, r = 500.

Equation	SAa	S-W b	P(S-W) c	R2	Frequently selected eigenvectors					
					E1	E7	E19	E3	E14	E9
10-by-10 regular square tessellation sampling grid										
linear	0.25	0.98666	0.50316	0.390	486	491	281	297	190	189
	0.50	0.98656	0.50152	0.591	500	500	447	446	339	328
	0.90	0.98548	0.46483	0.891	500	500	500	500	500	500
quadratic	0.25	0.98646	0.49298	0.386	485	480	297	277	193	193
	0.50	0.98673	0.50319	0.585	500	500	448	444	326	331
	0.90	0.98259	0.39754	0.889	500	500	500	500	500	500
sinusoidal	0.25	0.98672	0.50446	0.387	492	485	300	300	176	182
	0.50	0.98671	0.50011	0.578	500	500	450	436	337	318
	0.90	0.98526	0.46467	0.891	500	500	362	361	500	500
20-by-20 regular square tessellation sampling grid										
					E1	E14	E3	E40	E16	E28
linear	0.25	0.99613	0.50948	0.393	500	500	477	484	365	374
	0.50	0.99590	0.48254	0.593	500	500	500	500	492	492
	0.90	0.99546	0.44548	0.913	500	500	500	500	500	500
quadratic	0.25	0.99592	0.48214	0.391	500	500	469	466	380	371
	0.50	0.99596	0.49945	0.592	500	500	500	500	491	490
	0.90	0.99406	0.32749	0.913	500	500	500	500	500	500
sinusoidal	0.25	0.99594	0.48412	0.392	500	500	476	472	375	371
	0.50	0.99599	0.49827	0.595	500	500	500	500	494	491
	0.90	0.99545	0.45417	0.912	500	500	500	500	500	500

[a] The percentage of variance attributable to spatial autocorrelation.
[b] The Shapiro-Wilk normality diagnostic statistic.
[c] Probability of the S-W statistic under the null hypothesis of normality.

Table 4 Summary of simulation experiment results.

Equa-tion	SAa	$\bar{y}$	$\bar{y}^*$ (μ=5)			s	s_ε (σ_ε = 1)		
			SARb	GSc	SFd		SARb	GSc	SFd
10-by-10 regular square tessellation sampling grid									
linear	0.25	4.9945	4.9945	4.6125	4.9945	1.15460	1.08378	1.01423	0.94115
	0.50	5.0096	5.0086	4.4852	5.0096	1.41954	1.13727	1.03981	0.95807
	0.90	4.9972	4.9981	0.6008	4.9972	3.17349	1.23828	1.87961	1.12877
quad-ratic	0.25	4.9975	4.9987	4.6296	4.9975	1.15571	1.08665	1.01600	0.94422

	0.50	5.0083	5.0130	4.4787	5.0083	1.41619	1.14173	1.03705	0.96345
	0.90	4.9984	5.0816	0.2261	4.9984	3.18135	1.25102	1.90990	1.15549
sinu-soidal	0.25	5.0054	5.0053	4.6110	5.0054	1.15821	1.08940	1.01782	0.94596
	0.50	5.0021	4.9990	4.4685	5.0021	1.41012	1.14341	1.04223	0.96596
	0.90	5.0101	4.7521	0.8410	5.0101	3.17628	1.31971	3.30126	1.15223
20-by-20 regular square tessellation sampling grid									
linear	0.25	4.9982	4.9983	4.7813	4.9982	1.15350	1.08226	1.00914	0.93499
	0.50	5.0017	5.0018	4.6570	5.0017	1.41736	1.13043	1.03265	0.94844
	0.90	4.9981	4.9972	4.2843	4.9981	3.16628	1.17016	2.56113	1.00431
Quad-ratic	0.25	5.0071	5.0075	4.7779	5.0071	1.15330	1.08204	1.01154	0.93683
	0.50	5.0007	5.0018	4.6307	5.0007	1.41365	1.12786	1.03225	0.94791
	0.90	4.9982	5.0179	4.3396	4.9982	3.16627	1.16709	2.57893	1.01661
sinusoidal	0.25	4.9992	4.9992	4.7791	4.9992	1.15386	1.08315	1.00939	0.93592
	0.50	4.9927	4.9918	4.6543	4.9927	1.41128	1.12345	1.02809	0.94275
	0.90	4.9985	4.9419	0.1274	4.9985	3.16371	1.18065	4.12153	1.01715

[a] The percentage of variance attributable to spatial autocorrelation.
[b] Denotes the spatial simultaneous autoregressive model.
[c] Denotes the geostatistical semivariogram model.
[d] Denotes the spatial filter model.

4 Results

Simulation experimental results are summarized in Table 4. The SAR and spatial filter models yield correct mean estimates; the geostatistical semivariogram model struggles with this task. Although spatial autocorrelation is found to corrupt the normality of data as it becomes stronger, accounting for spatial autocorrelation effects tends to correct for much of this corruption. Weak-to-moderate levels of positive spatial autocorrelation are well captured by all three models, which to various degrees have some difficulty fully capturing very high levels. The poorest performance is by the geostatistical semivariogram model for very high levels of positive spatial autocorrelation, and for the cases of a sinusoidal random variable across all modeling efforts. This former outcome may well be attributable to a failure to customize the semivariogram model range when the autocorrelation is very high. One advantage of the semivariogram model is that it directly renders the sill as the variance estimate for the underlying independence-equivalent data. Correcting for weak-to-moderate levels of positive spatial autocorrelation yields much improved variance estimates; for strong levels, variance estimates are better but still tend to suffer from geographic dependency effects, with this effect tending to diminish as sample size increases for the SAR and spatial filter models.

5 Discussion and implications

Considerable gains in efficiency can be gleaned from model-informed spatial statistical analysis, as is illustrated by the following relative efficiencies calculated for the simulation experiments (computations from values appearing in Table 4):

		n = 100			n = 400		
	SA[a]	SAR[b]	GS[c]	SF[d]	SAR[b]	GS[c]	SF[d]
linear	0.25	0.938663	0.87843	0.815131	0.938240	0.87485	0.810568
	0.50	0.801154	0.73250	0.674916	0.797560	0.72857	0.669160
	0.90	0.390195	0.59228	0.355687	0.369569	0.80888	0.317189
quadratic	0.25	0.940245	0.87911	0.817004	0.938212	0.87708	0.812304
	0.50	0.806198	0.73228	0.680311	0.797835	0.73020	0.670541
	0.90	0.393236	0.60034	0.363207	0.368601	0.81450	0.321075
sinusoidal	0.25	0.940589	0.87879	0.816743	0.938719	0.87479	0.811121
	0.50	0.810860	0.73911	0.685020	0.796050	0.72848	0.668011
	0.90	0.415489	1.03935	0.362761	0.373185	1.30275	0.321505

[a] The percentage of variance attributable to spatial autocorrelation.

[b] Denotes the spatial simultaneous autoregressive model.

[c] Denotes the geostatistical semivariogram model.

[d] Denotes the spatial filter model.

The spatial filter model outperforms the SAR model, but not by a sizeable amount. The geostatistical semivariogram model furnishes inferior results. Modest gains can be attained when spatial autocorrelation is very large by increasing sample size.

References

Griffith, D. 2004. Extreme eigenfunctions of adjacency matrices for planar graphs employed in spatial analyses, Linear Algebra & Its Applications, 388: 201-219.

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

Overton, W. and S. Stehman. 1993. Properti4es of designs for sampling conti8nuous spatial resources from a triangular grid, *Communications in Statistics─Theory and Methods*. 22: 2641-2660.