Conference on Applied Statistics in Agriculture   2004 - 16th Annual Conference Proceedings

A COMPARISON OF GEOSTATISTICAL AND SPATIAL AUTOREGRESSIVE APPROACHES FOR DEALING WITH SPATIALLY CORRELATED RESIDUALS IN REGRESSION ANALYSIS FOR PRECISION AGRICULTURE APPLICATIONS

Ignacio Colonna

Matías Ruffo

Germán Bollero

Don Bullock

See next page for additional authors

Recommended Citation

Colonna, Ignacio; Ruffo, Matías; Bollero, Germán; and Bullock, Don (2004). "A COMPARISON OF GEOSTATISTICAL AND SPATIAL AUTOREGRESSIVE APPROACHES FOR DEALING WITH SPATIALLY CORRELATED RESIDUALS IN REGRESSION ANALYSIS FOR PRECISION AGRICULTURE APPLICATIONS," *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1169

Author Information

Ignacio Colonna, Matías Ruffo, Germán Bollero, and Don Bullock

A COMPARISON OF GEOSTATISTICAL AND SPATIAL AUTOREGRESSIVE APPROACHES FOR DEALING WITH SPATIALLY CORRELATED RESIDUALS IN REGRESSION ANALYSIS FOR PRECISION AGRICULTURE APPLICATIONS

Ignacio Colonna, Matías Ruffo, Germán Bollero and Don Bullock
Dept. of Crop Sciences. University of Illinois at Urbana – Champaign.

Abstract

Regressions such as Grain yield=*f*(soil,landscape) are frequently reported in precision agriculture research, and are typically computed using conventional OLS methods, implicitly ignoring spatial correlation of the residuals. This oversight can have a marked effect on the final conclusions derived from these regressions. A further issue is, which approach should be used to account for this problem? We investigated this question using a 2 year data set that includes site-specific soil and topographic information and soybean yields and compare regression results from direct covariance representation and spatial autoregressive approaches. Our results show that the coefficients from both spatial approaches are in many cases significantly different to those from OLS, but the estimates from both spatial approaches appear to show little differences. To provide further insight into the comparison among these approaches we use a simulation of spatial random fields, with a model containing 2 independent explanatory variables and a spatially structured residual term. We then estimated the coefficients for 1000 simulations of this field and assessed their distributional properties. All methods yielded overall unbiased estimates and OLS showed the largest standard errors, while the 'spatial' approaches proved to be relatively consistent, although a certain neighborhood specification within the spatial autoregressive model had an evidently lower performance than the rest.

Keywords: spatial regression, mixed models, spatial autoregressive model, precision agriculture, simulation.

1. Introduction

During the early 90's, both the users and research agricultural communities have experienced a rapid increase in their interest in technologies related with what has been called precision agriculture (Pierce and Nowak, 1999). In general., the main objectives of precision agriculture are to increase farmers' profitability while keeping environmental contamination to reduced levels compared to conventional management. Within this new approach, the achievement of these objectives is attempted through the incorporation of management strategies at, when convenient, a finer spatial scale than has been traditionally used within conventional field crops farming. Such strategies need to be based then on information collected at scales smaller than an entire field. This information should provide the basic input for decisions such as division of the original field into areas that will potentially receive different management, or the detection of useful field properties that can act as predictors of crop response to the application of inputs. Examples of this kind of 'site-specific' information are yield monitor data, intensive soil samples, remotely sensed imagery, or data from automated soil or plant sensors. While precision agriculture is still a technology under development there exists an important amount of published research aimed at quantifying the relationship between variables of interest such as crop production, economic indices of profitability or residual soil nutrients with environmental variables at within-field spatial scales. One approach that has been popular among these publications is the use of several forms of regression analysis in an attempt to provide a descriptive or functional characterization of this relationship for a particular region (Lark, 1997; Coelho et al., 1999; Kravchenko and Bullock, 2000; Kitchen et al., 2003; Machado et al., 2000; Kaspar et al., 2003 & 2004). Taking for example crop yield as a response variable and soil properties as potentially explanatory variables, these analyses are frequently based on a linear model that can be expressed as in

$$\boldsymbol{yield}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon} \qquad \textbf{(1)}$$

Where ***yield*** represents a vector of n observations taken at different locations within a field, $\boldsymbol{X}$ is a matrix containing information for the n points from v concomitant variables that may be either the originally measured field properties (e.g. Plant et al., 1999; Kaspar et al., 2003) or some transformation frequently based on multivariate methods (Chang et al., 2003; Mallarino et al., 1999; Machado et al., 2000), $\boldsymbol{\beta}$ is the vector of coefficients for each variable in $\boldsymbol{X}$ and $\boldsymbol{\varepsilon}$ is a vector of n residuals, containing the fraction of ***yield*** not explained by the variables in $\boldsymbol{X}$. Even though these studies usually involve the use of numerous concomitant variables, the relative size of the 'non-explained' part of ***yield*** is still typically large, the R^2 of the model being frequently in the order of 0.1-0.6 (e.g. Perez-Quezada et al., 2003; Kaspar et al., 2003; Kitchen et al., 2003). A considerable part of the residual values for these models is thus contributed by field properties omitted from the model, but that are relevant somehow to the explanation of variation in the response variable. These variables often show some degree of spatial structure or clustering, in the sense that similar values tend to appear close to each other in the field (Cambardella et al., 1994, Robertson et al., 1996), thus generating a vector $\boldsymbol{\varepsilon}$ of spatially correlated residuals.

One of the central points in the studies under consideration relies on the interpretation of the significance levels, sign and magnitude of the regression parameters in $\hat{\beta}$, the corresponding estimate for the unknown $\boldsymbol{\beta}$. It is well known that the ordinary least squares (OLS) estimator for $\hat{\boldsymbol{\beta}}=\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y}$ is not the best linear unbiased estimator (BLUE) under situations with non-independent residuals (see e.g. Green, 1991 ch. 13). Although the spatial variability in yield and in the explanatory variables in $\boldsymbol{X}$ is usually explicitly recognized and central to the subject in precision agriculture research, it is paradoxical that the majority of these studies still base the computation of $\hat{\beta}$ in the OLS estimator, ignoring the aforementioned correlation in $\boldsymbol{\varepsilon}$ by assuming their distribution as $N(0,\sigma^2\boldsymbol{I})$. Furthermore, the report of conventional correlation matrices and their corresponding significance tests for a set of field properties is common practice, although these are also likely incorrect for spatially dependent variables, as shown in Bivand (1980), Clifford et al. (1989), and Dutilleul (1993). In research studies like those exemplified above, the use of OLS can lead, specifically, to underestimated standard errors for regression parameter estimates relative to the true OLS standard errors for these estimates and inflated significance levels of the corresponding statistical tests, although $\hat{\beta}$ may be asymptotically unbiased (Greene, 1993 ch. 13, Upton and Fingleton, 1985 p. 285; Haining, 1990 p. 161). Under the general case of non-spherical $\boldsymbol{\varepsilon}$ the BLUE corresponds to the generalized least squares estimator (GLS) of the form

$$\hat{\boldsymbol{\beta}}=\left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Y} \qquad (2)$$

where $\boldsymbol{V}$ is the n x n covariance matrix of the residuals in $\boldsymbol{\varepsilon}$, with an expected value not equal to $\sigma^2 \boldsymbol{I}$ (Greene, 1993, Ch. 13 ; Upton and Fingleton, 1985 p. 277).

The effect of correlated residuals on parameter estimates has been recognized in agricultural research for a long time, but in practice this has been almost exclusively addressed in experimental studies involving variety trials, where the use of some variation of the nearest neighbors approach has become common, (Wilkinson et al., 1983, Bhatti et al., 1991) and other recently adopted methods based on geostatistics (Ball et al., 1993, Brownie et al., 1993, Stroup et al.,1994). In a reduced number of regression analyses in the area of precision agriculture, spatial correlation was accounted for in the estimation of $\boldsymbol{\beta}$ by the use of basically three different approaches: nearest-neighbors analysis (Mamo et al., 2003, Bermudez and Mallarino, 2003), direct covariance representation (Lark and Wheeler, 2003; Kaspar et al., 2004) and spatial autoregression (SAR, Long, 1998; Florax et al., 2002; Anselin et al., 2002). In this paper, for purposes of accounting for spatial correlation in regression residuals, we deal only with the direct covariance representation and SAR methods. The approach we call 'direct covariance representation' corresponds to the 'random field linear model' in Zimmerman and Harville (1991) and is of more common use among researchers in the biological and agricultural sciences, whereas the SAR approach is widely applied in the area of econometrics and economic geography. Both methods are explained in some detail in the next two sections.

While theoretical and applied aspects of both 'spatial' methods have been separately addressed in numerous publications (e.g. Anselin and Bera, 1998; Florax et al., 2002, Zimmerman and Harville, 1991; Grondona and Cressie, 1991; Lark, 2000) we are only aware of one paper in which outcomes from the direct covariance representation approach have been directly compared against those from SAR-error (Lambert et al., 2002), although this comparison was limited to a specific experimental study, with no attempt to describe more general properties of both estimators. Christman and Jernigan (1997) have performed a similar comparison of both spatial models in an application to a problem in evolutionary biology.

In this paper our main purpose is to emphasize the possible consequences of ignoring spatial correlation in $\boldsymbol{\varepsilon}$ in regression analysis used within the context of precision agriculture research and to contrast the possible outcomes obtained from the use of two methods that account for this condition. To achieve this, we first illustrate the actual differences in the estimation outcomes from OLS, direct covariance representation and SAR approaches, which motivates our further attempt to provide practical insight into somewhat more general empirical properties of the three methods. We begin by briefly describing in the next two sections some basic aspects of the direct covariance representation and SAR approaches, continue by presenting the results from two examples based on field data and end with a spatial random field simulation exercise that yields empirical distributions for $\hat{\boldsymbol{\beta}}$ under the specified conditions for the simulated fields. Our presentation of both spatial approaches in the following two sections is of course rather succinct. A good presentation of theoretical and practical aspects of the first method with several examples can be found in Schabenberger and Pierce (2002, ch.7 and 9), and equivalent coverage for the second method in the books by Cliff and Ord (1981), Upton and Fingleton (1985) and its application to econometrics in Anselin (1988) and Anselin and Bera (1998).

2. Direct covariance representation (‘*geostatistical’* approach)

While most references to this method seem to converge to the paper by Zimmerman and Harville (1991), its use in the literature dates back at least to the study by Cook and Pocock (1983), to regressions based on non-experimental data. The case that concerns us in this paper is that of a fully fixed model such as (1), where $\boldsymbol{\varepsilon}$ is distributed as $N(0,\boldsymbol{V})$, and the elements of $\boldsymbol{V}$ are obtained by imposing a certain structure in their covariances that depends on only a few parameters. For the examples given in this paper, we use the spherical and exponential covariance models, two structures frequently applied in spatial analyses for agricultural or ecological applications (e.g. Cambardella and Karlen, 1999; Goovaerts, 1998). For the first one, the covariance between any two residuals ε_i and ε_j is a nonlinear isotropic function of their physical separation distance d and two other parameters as

$$Cov(\varepsilon_i,\varepsilon_j) = V(d,\sigma_s^2,a) = \sigma_s^2\left(1-\frac{3d}{2a}+\frac{d^3}{2a^3}\right) if\ 0 \le d \le a \quad \textbf{(3)}$$

$$V = 0\ otherwise,$$

And for the exponential model, the function is as

$$V(d,\sigma_s^2,a) = \sigma_s^2 \exp(-da) \quad \textbf{(4)}$$

In (3) and (4) $\boldsymbol{V}$ has the same meaning as in (2), the parameter a is usually called ‘correlation range’ and in this context, σ_s^2 is called the ‘sill’ when the counterpart of this function is represented in a semivariogram. A further variance component equal to $\sigma_o^2\,\boldsymbol{I}$, usually called ‘nugget’ in geostatistics, can be added to both models to represent non-spatial measurement error or variation at very small distance lags. Although we have expressed $\boldsymbol{V}$ as a function of only 4 possible parameters, in reality one need not assume constant values of σ_s^2 and a throughout the field, and different structures can be fitted if this appears to better represent the observed covariance patterns in the data (e.g. Schabenberger and Pierce, 2002 p.676). In our regression examples, all parameters in $\boldsymbol{V}$ are either estimated externally and fixed or estimated by restricted maximum likelihood (REML) and $\boldsymbol{\beta}$ further estimated as a function of these covariance estimates by replacing $\boldsymbol{V}$ in (2) by its estimate $\hat{\boldsymbol{V}}$, a method called empirical or estimated GLS (EGLS).

3. Spatial autoregressive approach (SAR)

In this case, correlation of the residuals is not accounted for by a direct modeling of the elements in $\boldsymbol{V}$ but rather indirectly through a specification of the linear model according to the nature of the process generating the correlation pattern. Although many possible model specifications exist (see for example Anselin, 2002 & 2003 or Upton and Fingleton, 1985 ch. 5), the two most widely used in spatial regression applications, are the ‘spatial lag’ and ‘spatial error’ models. Conceptually, the first type corresponds to either a situation whereby the value of the response variable at a certain location is in part the result of a contagion, or diffusion, effect from the same variable at its neighboring locations (e.g. adoption of a certain technology), or to a case where there is a mismatch between the scale at which a variable is measured and the true spatial scale of the process represented by that variable (Anselin and Bera, 1998). The second type, instead, corresponds to the effect observed in the OLS residuals due to the omission of spatially structured explanatory variables in $\boldsymbol{X}$ that are in fact related to the dependent variable, such that the residuals ‘absorb’ the spatial structure from these (Anselin and Bera, 1998). In our interpretation, it is this second specification that applies better to the common case found in most non-experimental studies within the area of precision agriculture, where one attempts to 'explain' yield variations by including a limited number of field properties as concomitant variables in the model, but this group is almost invariably incomplete in terms of the true set of variables that cause the observed spatial variation in grain yield, or whichever other response variable is investigated. For this paper we choose to use the spatial-error model specification and do not pursue further explanations on the spatial lag specification. We note in passing that the use of the spatial lag model in exploratory analyses of this sort, where the explanatory variables have a strong spatial structure themselves as is the case in Long (1998) and Florax et al. (2002), is questionable and can obscure the true influence of the field properties under investigation on the dependent variable (see for example comments in Upton and Fingleton, p. 373 about a similar situation in a study by Cook and Pocock).

Based on (1) the SAR-spatial error model can be expressed by defining ε as

$$\varepsilon = \lambda \boldsymbol{W} \varepsilon + \xi = (\boldsymbol{I} - \lambda \boldsymbol{W})^{-1} \xi \qquad \textbf{(5)}$$

Here, $\boldsymbol{W}$ is a binary weights matrix that defines the neighborhood structure in a discrete way (i.e. 1=neighbors, 0 otherwise) and for practical purposes it is typically used in a row-standardized form such that $\sum_{j=1}^{n} w_{ij} = 1$. Figure 1 gives a physical representation of two commonly used neighbor structures for a regular grid of points, one based on vicinity and called 'queen' criterion, in an analogy to the movement of this piece in a chess game, and the other based on a fixed radius from each point. The coefficient λ ('spatial autocorrelation parameter') represents the magnitude and direction (positive or negative) of correlation in the residuals, with the constraint $|\lambda| < 1$. The first term in (5) can be intuitively seen as an average of the residuals from the neighbors of each point, multiplied by λ. In this sense, then, a value of λ=0 corresponds to a case with spatially independent residuals. The error term ξ , for our purposes, is taken as being $N(0,\sigma^2 \boldsymbol{I})$. From model (5), the resulting covariance matrix for ε is:

$$\boldsymbol{V} = \sigma^2 \left[(\boldsymbol{I} - \lambda \boldsymbol{W})' (\boldsymbol{I} - \lambda \boldsymbol{W}) \right]^{-1} \qquad \textbf{(6)}$$

where λ and σ^2 can be estimated by maximum likelihood, and $\boldsymbol{\beta}$ is again estimated by EGLS based on (2), although other alternatives also exist (Anselin and Bera, 1998). Notice that even if the neighboring structure in $\boldsymbol{W}$ is defined in a discrete way, the final form of $\boldsymbol{V}$ in a SAR-error model is that of a *full* matrix, where all pairs of neighbors have a nonzero covariance. This contrasts with, for example, the case of the spherical model in the previous approach whereby any pair of neighbors separated by a distance greater than *a* have a zero covariance. Also as a consequence of specification (5), $\boldsymbol{V}$ is intrinsically heteroscedastic, unless each point in the dataset has an identical number of neighbors (Anselin and Bera, 1998; Anselin, 2003). This property applies to the SAR-error case regardless of the properties of the *true* covariance matrix in the data, and thus σ^2 in (5) cannot be interpreted in the typical way as a homoscedastic error variance. Figure 2 displays the resulting covariance structure for a particular point in the field as a function of distance to its neighbors, for two geostatistical models and the same two neighborhood structures used in Figure 1.

4. Application to field data: Regression of grain yield to field characteristics

4.1. Data and methods

The area under analysis is a 20-acre production field, located in East-Central Illinois, for which soybean grain yield data was available for the cropping seasons of 1999 and 2001 (Figure 3). These values were originally collected with a combine yield monitor in transects separated 9 m apart, with a density of about one point every 2 m along each transect. From a larger set of field characteristics, we selected two soil chemical properties and five topographical indices as explanatory variables for yield spatial variation (Figure 3). The first two variables are soil phosphorus and potassium, expressed in ppm, which were measured in a grid of 110 points up to a depth of 8". The topographical indices were computed from spatially intensive sampling conducted using a real-time kinematic differential global positioning system (RTK-DGPS). These variables are profile curvature (***Prof***), plan curvature (***Plan***), stream power index (**SPI**), compound topographic index (***CTI***), and aspect (***Ang***). These are defined or computed as follows: ***Prof*** and ***Plan*** are the second derivatives of the elevation surface in the principal direction of the slope and in a direction perpendicular to this one, respectively, and are expressed in degrees. ***SPI*** is computed as ***SCA*** x ***slope*** (%), where ***SCA*** is specific catchment area, a measure of number of grids draining to a given grid in an elevation model, and slope is the change in elevation in the direction of maximum elevation difference. ***CTI*** (or wetness index) is calculated as ln (***SCA*** / slope (%)) and ***Ang*** is the direction of slope and is expressed in radians counter clockwise from the east direction. These landscape attributes have been shown to be good descriptors of the effect of elevation on water dynamics within a field (Moore et al., 1991; Wilson and Gallant, 2000; Western et al., 1999; Chamran et al., 2002), thus influencing grain yield spatial variation by modulating plant water availability during the crop season (Halvorson and Doll, 1991; Kravchenko and Bullock, 2000).

Previous to performing the regression analysis the response and potential explanatory variables were kriged (***yield***, ***P*** and ***K***) or locally averaged (topographical variables) to a common regular grid at 20 m of separation between points, resulting in a total of 371 points out of which 3 were eliminated due to being considered outliers in their values for the topographical variables. Our final analysis was then based on 368 data values (Figure 3).

The grain yield for each year was related to the field characteristics using a linear model with no interactions as

$$\boldsymbol{Yield}_j = \beta_{0j} + \beta_{1j}\boldsymbol{P} + \beta_{2j}\boldsymbol{K} + \beta_{3j}\boldsymbol{Prof} + \beta_{4j}\boldsymbol{Plan} + \beta_{5j}\boldsymbol{Ang} + \beta_{6j}\boldsymbol{SPI} + \beta_{7j}\boldsymbol{CTI} + \boldsymbol{\varepsilon}_j \qquad \textbf{(7)}$$

where j=1999, 2001, and each variable in the model is a vector with n=368, and β_0-β_7 are the elements of vector $\boldsymbol{\beta}$ in (1). More sophisticated and possible nonlinear models including interactions may be more appropriate for understanding the sources of grain yield variability, but a rather simple one is well suited for the purpose of our example.

As a first step, the β_{ij} in model (7) were estimated by OLS, and the spatial structure of the residuals analyzed using experimental semivariograms to which alternative models were fitted using a routine based in PROC NLIN in SAS®. This revealed marked differences in nugget, sill and correlation range between years, justifying the decision of fitting a separate covariance structure to each year. The β_{ij} for the direct covariance representation approach was then estimated in SAS® using a spherical covariance structure (Figure 4), using the following code (example for year 1999):

```
proc mixed data=soiltopyldind noprofile scoring=16;
```

```
title 'spatial regression - 20m - year 99';
model yield= P K  Prof  Plan Spiavg Cti Ang/ddfm=kr solution outp=unadjresid99;
parms /*sill*/ (1000) /*range*/(70) /*nugget*/ (600)/noiter;
repeated /subject=intercept local type=sp(sph) (x y) r rci;
ods output InvCholR= choleskyinv; run;
```

The values specified in parms correspond to the models obtained from the external fitting of semivariogram model (3) as explained above. The decision of fixing the covariance parameter (option noiter) rather than estimating them by restricted maximum likehood (REML) was based on the unrealistic parameters obtained by this method when compared against the experimental semivariograms in Figure 4. This is due to the influence of covariance values at distances beyond half the field length or width (200 m), a threshold after which the points contributing to each value in the covariogram correspond to ever smaller subsets of the entire dataset, and thus are not reliable representations of the overall covariance structure (Littel et al., 1996). The results for the REML estimates are nonetheless reported, to compare their standard errors with those from the noiter case. The option rci requests the inverse of the Cholesky decomposition matrix for the covariance matrix of the residuals. This matrix satisfies the relation r=rci^{-1}(rci^{-1})', r being equivalent in this case to ***V*** in (1). It is thanks to this property that this matrix can be used to check the appropriateness of the fitted covariance model in the representation of the observed structure. For this purpose, an experimental variogram was computed for a set of residuals, adjresid, adjusted by the imposed covariance structure. This adjusted set was obtained using the code:

```
proc iml;
use choleskyinv; read all into choleskyinv99;
choleskyinv99=choleskyinv99[,3:370];            *cholesky inverse matrix;
use unadjresid99;         *dataset containing unadjusted residuals from proc mixed;
read all var{x y resid};
adjresid=choleskyinv99 *resid;  *'adjust' residuals;
create residadjust var{x y adjresid};
append; quit;
```

Two models were obtained with the SAR-error approach corresponding to the neighborhood structures shown in Figures 1 and 2, queen and distance-based with a radius of 50 m (‘d50’). The weights matrices for both schemes were obtained in Geoda™ (Anselin, 2003) and the regression analysis was performed using the spdep library (Bivand, 2003) in R statistical package (http://cran.r-project.org). The basic code in R is:

```
attach(north99) #dataset for regression
#fit the model:
yldyear.err.99<-errorsarlm(yield~P+K+Prof+Plan+Spi+Cti+Ang,queen99.wt)
#get adjusted residuals for variogram computation:
res.err.99<-resid(yldyear.err.99)
```

The residuals obtained from spdep correspond to the error term ξ in (5) and are thus already adjusted for the resulting spatial structure imposed by ***W***.

4.2. Results and discussion

The consistency in the parameter estimates of (7) obtained from OLS, SAR-error or direct covariance representation was largely dependent on the year under analysis (Table 1).

Noticeably, the OLS estimates did not differ significantly from the estimates by the other methods for any of the explanatory variables in year 1999, but considerable differences were observed among them for 2001. Specifically, **P 01** under OLS is significant at p<0.001 but non-significant under all of the other spatial approaches, while both curvature variables change from non-significant to significant and negative, and the opposite occurs with ***SPI*** and ***CTI***. Our interpretation of this result relies in the fact that the spatial structure for the OLS residuals (Figure 3) for 1999 is much lower than that for 2001, as evidenced by a higher nugget/sill ratio (0.38 vs. 0.14) and a shorter correlation range (70 vs. 120 m) for the first year. Consequently, the departure of the structure in $\boldsymbol{V}$ from $\sigma^2\boldsymbol{I}$ is larger for 2001 than for 1999, thus $\hat{V}^{-1}$ has a stronger influence in the computed $\hat{\beta}$ for 2001 when these are based in EGLS of the form (3). The distinct spatial structure of residuals between years can be explained at least in part by the contrasting rainfall amounts between years. During the 1999 cropping season, rainfall was considerably higher than for 2001 (600 vs 411 mm for the period April-August), most of this difference being concentrated early in the season where water excess has a strong negative effect on plant establishment and growth. This might explain the higher influence of topographical variables in yield variation in 1999 versus that in 2001 as expressed by the significantly larger (p<0.01) absolute values in the regression coefficients for the first year. Thus, a larger part of the spatially structured variation in grain yield was explained by the variables $\boldsymbol{X}$ in 1999 than in 2001, implying that for this last year the influence of spatially dependent explanatory variables omitted from the model, likely unrelated to topography, was included in the residual term.

Both spatial approaches appear to have eliminated most of the spatial structure in $\boldsymbol{\varepsilon}$ compared to that in the unadjusted values, as shown by the experimental semivariograms in Figure 5. The residuals for the SAR-error methods seem to show a slight spatial structure with a short range, but this is still considerably lower than that for the OLS residuals. As a result, the parameter estimates obtained in all spatial approaches to regression are somewhat consistent, and do not show appreciable differences among them.

It is conceivable that the high degree of multicollinearity present in the variables in $\boldsymbol{X}$ might have an influence on the observed outcome from our comparison. To verify this, we ran a similar analysis using a set of orthogonal variables obtained by a principal components transformation on the original set. While these results are too extensive to be included in this paper, we do confirm that the degree of consistency among methods and years we found for this case was equivalent to that reported for the original variables. While for this example the results from different spatial methods appeared to agree, a more general and informative approach to compare them under conditions similar to those in this analysis was pursued.

5. A more general empirical comparison: Monte Carlo simulation

5.1. Generation of spatial random fields

The steps for this part of the analysis were:

1) Generate n values for three mutually independent, spatially structured random variables in a field of dimensions $\sqrt{n} \times \sqrt{n}$. Call these vectors $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ and $\boldsymbol{\varepsilon}$.
2) Generate a dependent variable $\boldsymbol{y}$ by combining the vectors from the previous step in a model of the form: $\boldsymbol{y}=\beta_o+\beta_1\boldsymbol{X}_1+\beta_2\boldsymbol{X}_2+\boldsymbol{\varepsilon}$.

3) Fit a linear model $\boldsymbol{y}=f(X_1,X_2)$ and estimate β_0,β_1 and β_2 by OLS, direct covariance representation and SAR-error, using different structures for $\boldsymbol{V}$ and $\boldsymbol{W}$ to evaluate the consequences of a misspecification in the correlation structure on the estimation results.
4) Save the output from 3 and return to 1, repeat the sequence *s* times.

The spatial random fields from step 1 were generated in PROC SIM2D in SAS®, using LU (Cholesky) decomposition of the covariance matrix. In the past this method was usually recommended only for small grids of less than 100 points (Goovaerts, 1997), but our simulation of 441 points took less than 0.2 seconds per field in a PC with a Pentium 4 processor. This was done for a square field where the minimum distance between 2 points was assigned an arbitrary value of 20 m. The spatial structure was based on spherical covariance models for the three variables, with mean and covariance parameters as specified in Table 2. The values chosen for the parameters in the model were : β_0=10, β_1=0.6 and β_2=1.2, which together with the covariance parameters assigned to the variables would result in an adjusted-R^2 of about 0.37, well within the order of those frequently found in the literature for regressions like the one performed in the previous section.

This sequence was repeated 1000 times, thus resulting in an equal number of realizations of this hypothetical field. The corresponding 1000 regressions were fit in SAS® for the direct covariance representation with 4 different error covariance specifications: spherical-no nugget (true structure), spherical-nugget, exponential-no nugget and exponential-nugget. Similarly, the same data were fit under the SAR-error model using the same two neighborhood structures as in the previous section.

5.2. Simulation results and discussion

In agreement with theory, the empirical distribution of $\hat{\beta}$ under OLS showed by far the largest standard errors among the three approaches (Figure 6). Furthermore, the incorrect OLS-based estimated standard errors for each run grossly underestimated the one derived from the empirical distribution (Figure 8). Interestingly, among the spatial approaches the d50 structure resulted in standard errors considerably larger than the rest (Figures 6 and 8). The empirical standard error of β_1 for d50 was equal to 0.08, whereas those for all other spatial models were between 0.054 (spherical no nugget) and 0.06 (queen). Similarly, for $\hat{\beta}_2$ the standard error of d50 was 0.09 while for others ranged between 0.064 (spherical no nugget) and 0.07 (Queen). This is likely due to the poor representation of the true covariance values at short ranges below 40m that results from the d50 specification as shown in Figure 2 for a specific point in the grid. Although the true covariance structure was a spherical with no nugget, all other geostatistical models evaluated and the SAR-error based on a Queen structure yielded very consistent estimates, with R^2 of about 0.99 for the regression of the 1000 parameter estimates from any of them against those from the spherical-no nugget model. When a nugget was specified, the values for this structure estimated by REML were never higher than 0.6 and averaged 0.1 and 0.005 for the spherical and exponential models, respectively, even though the initial value for the iterations was set at 1 in the `parms` statement in SAS®.

Finally, an intuitive idea of the practical significance of the differences mentioned above can be given by expressing the fraction of estimates in each empirical distribution that were beyond an interval of ± 15% of the true value for that parameter. That is, within the range 0.51-0.69 for β_1 and 1.02-1.38 for β_2. For the first parameter, only 55% of all OLS estimates were within this range, followed by SAR-error d50 where this fraction was 72%. For the other spatial methods, the values were considerably larger, ranging from 86% (Queen) to 90% (spherical-no nugget). In the case of β_2 these differences were not as important due to the ratio standard error/parameter value being 1.8 smaller to that for β_1. For OLS and d50 the corresponding fractions of the estimates within the mentioned range were 84% and 95%, respectively, whereas for the other models these values were above 99%.

6. Conclusions and summary

Our example of a regression analysis similar to those performed in precision agriculture research made evident that large inconsistencies are possible between OLS-based estimation and any other method that accounts for correlation in the residuals. In particular, all methods were fairly consistent when the residuals presented a correlation range of only about 70 m and a nugget/sill ratio close to 0.4, but significant differences between both approaches resulted when the range was in the order of 120 m and the nugget/sill ratio was about 0.15. Under this last situation the parameter estimates from the spatial approaches did not show significant differences among them from either statistical or practical standpoints. This was in accordance to the suggestion given by Zimmerman and Harville (1991) about the use of geostatistical models, and the results obtained by Lambert et al. (2002) in a comparison of geostatistical approaches vs. SAR-error. Moreover, the Monte Carlo simulation exercise showed that differences of practical relevance were certainly likely between OLS and the other spatial approaches, but within this last group only SAR-error d50 was an evidently inferior specification compared to the rest, for the particular simulation model we chose. It seems important then to correctly specify the neighbor structure for the SAR-error model in order to obtain appropriate estimates and standard errors.

Further explorations into the comparative performance of these methods will involve the simulation of spatial random fields with a variety of spatial structures in the error, ranging from situations with short correlation ranges and high nugget/sill ratios to more spatially continuous patterns. This would provide a better illustration of the situation observed in our example, where in one of the years the effect of the spatial correction appear negligible while in the other this had a strong influence.

References

Anselin, L. 1988. *Spatial Econometrics, Methods and Models*. Boston: Kluwer Academic.

Anselin, L and A. Bera. 1998. Spatial Dependence in Linear Regression Models with an Introduction to Spatial Econometrics. In A. Ullah and D. Giles (Eds.*),Handbook of Applied Economic Statistics*, pp. 237–289. New York: Marcel Dekker.

Anselin, L. 2003. Spatial Externalities, Spatial Multipliers and Spatial Econometrics. *International Regional Science Review* 26, 153-166.

Anselin, L. 2002. Under the Hood. Issues in the Specification and Interpretation of Spatial Regression Models. *Agricultural Economics* 27, 247-267.

Anselin, L., R. Bongiovanni and J. Lowenberg-DeBoer. 2002. A spatial econometric approach to the economics of site-specific nitrogen management in corn production. REAL paper 02-T-2. Available online at: http://www2.uiuc.edu/unit/real/d-paper/real02-t-2.pdf

Anselin, L. 2004. GeoDa 0.95i Release Notes. Spatial Analysis Laboratory (SAL). Department of Agricultural and Consumer Economics, University of Illinois, Urbana-Champaign, IL. Available online at: http://sal.agecon.uiuc.edu/geoda_download.php

Bermudez M. and A.P. Mallarino. 2002. Yield and early growth responses to starter fertilizer in no-till **corn** assessed with precision agriculture technologies Agron J. 94 (5): 1024-1033.

Bhatti A.U., D.J. Mulla, F.E. Koehler and A.H. Gurmani. 1991. Identifying and removing spatial correlation from yield experiments. Soil Sci. Soc. Am. J. 55 (6): 1523-1528.

Bivand, R. 1980. A Monte Carlo study of correlation coefficient estimation with spatially autocorrelated obdervatioins. Quaestiones Geographicae. 6:5-10.

Cambardella, C.A. and D.L. Karlen. 1999. Spatial analysis of soil fertility parameters. Precision Agriculture 1:5-14.

Cambardella C.A., T.B. Moorman, J.M. Novak, T.B. Parkin, D.L. Karlen, R.F. Turco and A.E. Konopka. 1994. Field scale variability of soil properties in Central Iowa soils. Soil Sci. Soc. Am. J. 58:1501-1511.

Chamran, F., P.E. Gessler, and O.A. Chadwick. 2002. Spatially explicit treatment of soil-water dynamics along a semiarid catena. Soil Sci. Soc. Am. J. 66:1571-1583.

Chang J.Y., D.E. Clay, K. Dalsted, S. Clay and M. O'Neill. 2003. Corn (Zea mays L.) yield prediction using multispectral and multidate reflectance. Agronomy Journal 95 (6): 1447-1453.

Christman, M.C. and R.W. Jernigan. 1997. Spatial correlation models as applied to evolutionary biology. pp. 221-232 in: Gregoire, T.C., D.R. Brillinger, P.J. Diggle, E. Russek-Cohen, W.G. Warren and R.D. Wolfinger (Eds.) Modelling longitudinal and spatially correlated data: Methods, applications and future directions. Springer, New York.

Clifford P,S. Richardson and D. Hemon. 1989. Assessing the significance of the correlation between 2 spatial processes. Biometrics 45 (1): 123-134.

Cook D.G. and S.J. Pocock. 1983. Multiple-regression in geographical mortality studies, with allowance for spatially correlated errors Biometrics 39 (2): 361-371

Dubin, R. 1988. Estimation of regression coefficients in the presence of spatially autocorrelated errors. *The Review of Economics and Statistics* 70, 466-474.

Dutilleul P. 1993. Modifying the t-test for assessing the correlation between 2 spatial processes. Biometrics 49 (1): 305-314.

Florax R.J.G.M., R.L. Voortman, J. Brouwer. 2002. Spatial dimensions of precision agriculture: a spatial econometric analysis of millet yield on Sahelian coversands. Agr. Econ. 27 (3): 425-443.

Goovaerts, P. 1997. Geostatistics for natural resources evaluation. Oxford University Press. New York.

Goovaerts, P. 1998. Geostatistical tools for characterizing the spatial variability of microbiological and physico-chemical soil properties. Biology and Fertility of Soils, 27:315- 334.

Haining, R.P. 1990. Spatial data analysis in the social and environmental sciences. New York : Cambridge University Press, 1990.

Halvorson G.A. and E.C. Doll. 1991. Topographic Effects On Spring Wheat Yields And Water-Use. Soil Sci Soc Am J 55 (6): 1680-1685.

Kitchen NR, S.T. Drummond, E.D. Lund, K.A. Sudduth and G.W. Buchleiter. 2003. Soil electrical conductivity and topography related to yield for three contrasting soil-crop systems. Agronomy Journal 95 (3): 483-495.

Kravchenko A.N. and D.G. Bullock. 2000. Correlation of corn and soybean grain yield with topography and soil properties. Agronomy Journal 92 (1): 75-83.

Lambert, D.M., J. Lowenberg-DeBoer and R. Bongiovanni. 2002. Spatial Regression, an Alternative Statistical Analysis for Landscape Scale on-farm Trials: Case Study of Variable Rate Nitrogen Application in Argentina. In: Robert et al., (eds.) Proceedings from the 6th International Conference on Precision Agriculture. Minneapolis, MN, USA, July 14-17, 2002.

Lark, R.M. 1997. An empirical method for describing the joint effects of environmental and other variables on crop yield. Ann Appl Biol 131 (1): 141-159.

Lark R.M. 2000. Regression analysis with spatially autocorrelated error: simulation studies and application to mapping of soil organic matter. International Journal Of Geographical Information Science. 14 (3): 247-264.

Lark R.M. and H.C. Wheeler. 2003. A method to investigate within-field variation of the response of combinable crops to an input. Agron. J 95 (5): 1093-1104.

Littell, R.C., G.A. Milliken, W.W. Stroup, and R.D. Wolfinger. 1996. SAS System for Mixed Models, Cary, NC: SAS Institute Inc.

Long, D.S. 1998. Spatial autoregression modeling of site-specific wheat yield. Geoderma 85 (2-3): 181-197

Machado, S., E.D.Jr. Bynum, T.L. Archer, R.J. Lascano, L.T. Wilson, J. Bordovsky, E. Segarra, K. Bronson, , D.M. Nesmith and W. Xu. 2000. Spatial and temporal variability of corn grain yield: site-specific relationships of biotic and abiotic factors. Precision Agriculture 2 (4) p. 359-376.

Mallarino, A.P., E.S. Oyarzabal and P.N. Hinz. 1998. Interpreting within-field relationships between crop yields and soil and plant variables using factor analysis. Precis. Agric. 1 (1) p. 15-25.

Moore, I.D., R.B. Grayson, and A.R. Ladson. 1991. Digital terrain modeling. A review of hydrological., geomorphological., and biological applications. Hydrological Processes 5:3-30.

Perez-Quezada J.F., G.S. Pettygrove and R.E. Plant. 2003. Spatial-temporal analysis of yield and soil factors in two four-crop-rotation fields in the Sacramento Valley, California. Agronomy Journal 95 (3): 676-687.

Pierce, F.J. and P. Nowak. 1999. Aspects of precision agriculture. Advances in Agronomy, Vol. 67:1-85.

Plant, R.E., A. Mermer, G.S. Pettygrove, M.P. Vayssieres, J.A. Young, R.O. Miller, L.F. Jackson, R.F. Denison and K. Phelps. 1999. Factors underlying grain yield spatial variability in three irrigated wheat fields. Transactions of the ASAE 42 (5): 1187-1202.

Robertson, G.P., M.A. Huston, F.C. Evans and J.M Tiedje. 1988. Spatial variability in a successional plant community: patterns of nitrogen availability. Ecology 69(5):1517-1524.

Schabenberger, O. and F.J. Pierce. 2002. Contemporary statistical models for the plant and soil sciences. Boca Raton, Fla.: CRC Press.

Schepers A.R., J.F. Shanahan, M.A. Liebig, J.S. Schepers, S.H. Johnson and A. Luchiari. 2004. Appropriateness of management zones for characterizing spatial variability of soil properties and irrigated corn yields across years. Agron J 96 (1): 195-203.

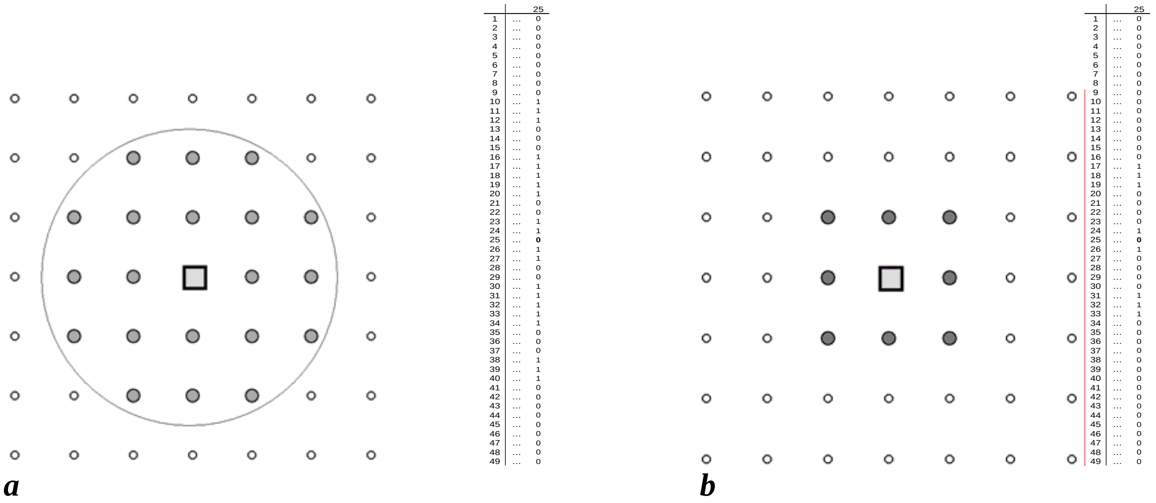
Upton, G. and B. Fingleton. 1985. Spatial data analysis by example. Volume I: Point pattern and quantitative data. Wiley, New York.

Western, A.W., R.B. Grayson, G. Blöschl, G.R. Willgoose, and T.A. McMahon. 1999. Observed spatial organization of soil moisture and its relation to terrain indices. Water Resour. Res. 35:797-810.

Wilkinson G.N., S.R. Eckert, T.W. Hancock and O. Mayo. 1983. Nearest neighbor (NN) analysis of field experiments Journal Of The Royal Statistical Society Series B-Methodological 45 (2): 151-211.

Wilson, J. P. and J.C. Gallant. Digital Terrain Analysis. 2000. In Terrain Analysis: Principles and Applications Wilson, J. P. And J.C. Gallant (Ed.). John Wiley & Sons, Inc.

Zimmerman D.L. and D.A. Harville. 1991. A random field approach to the analysis of field-plot experiments and other spatial experiments. Biometrics 47 (1): 223-239.



		25
1	...	0
2	...	0
3	...	0
4	...	0
5	...	0
6	...	0
7	...	0
8	...	0
9	...	0
10	...	1
11	...	1
12	...	1
13	...	0
14	...	0
15	...	0
16	...	1
17	...	1
18	...	1
19	...	1
20	...	1
21	...	0
22	...	0
23	...	1
24	...	1
25	...	**0**
26	...	1
27	...	1
28	...	0
29	...	0
30	...	1
31	...	1
32	...	1
33	...	1
34	...	1
35	...	0
36	...	0
37	...	0
38	...	1
39	...	1
40	...	1
41	...	0
42	...	0
43	...	0
44	...	0
45	...	0
46	...	0
47	...	0
48	...	0
49	...	0

		25
1	...	0
2	...	0
3	...	0
4	...	0
5	...	0
6	...	0
7	...	0
8	...	0
9	...	0
10	...	0
11	...	0
12	...	0
13	...	0
14	...	0
15	...	0
16	...	0
17	...	1
18	...	1
19	...	1
20	...	0
21	...	0
22	...	0
23	...	0
24	...	1
25	...	**0**
26	...	1
27	...	0
28	...	0
29	...	0
30	...	0
31	...	1
32	...	1
33	...	1
34	...	0
35	...	0
36	...	0
37	...	0
38	...	0
39	...	0
40	...	0
41	...	0
42	...	0
43	...	0
44	...	0
45	...	0
46	...	0
47	...	0
48	...	0
49	...	0

Figure 1. Representation of two possible neighborhood structures for ***W*** in the SAR-error model. a. distance-based criterion, using a 50-m radius around the central point. b. queen criterion. Neighbors (grey-filled circles) and non-neighbors (empty circles) are defined relative to the central point (square). In this case, ***W*** would be a 49 x 49 matrix. If we number the points 1to 49 from left to right and up to bottom the central point is assigned number 25, corresponding to the 25th column. The values for this column in a non row-standardized version of ***W*** are shown to the right of each graph. Notice that point 25 itself is assigned a value of 0.

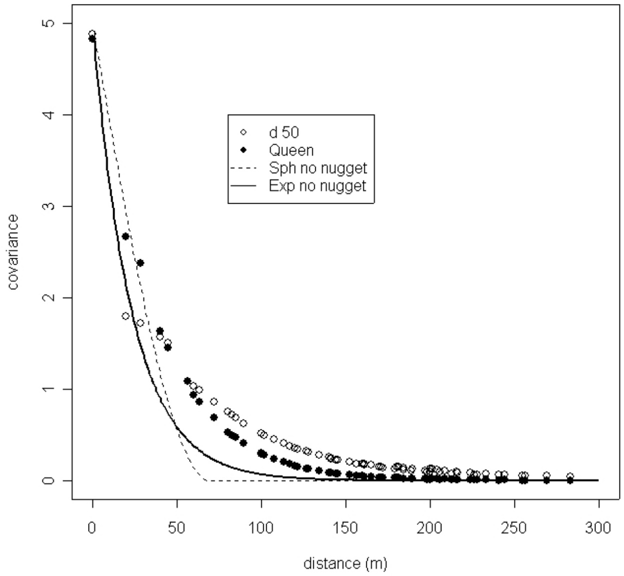
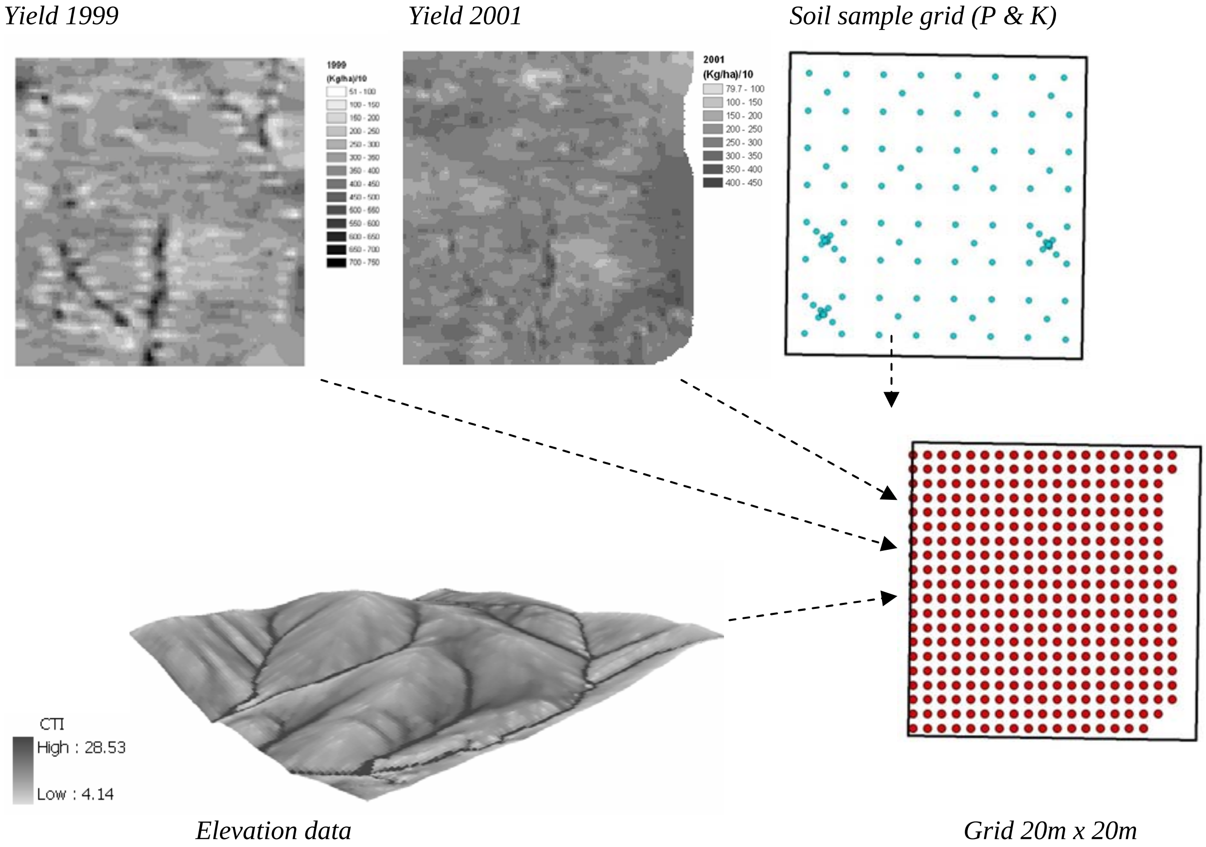


Figure 2. Covariance values between pairs of points at increasing separation distances. The values for the two geostatistical models apply to any point in the field, but those for the SAR-error models are specifically computed relative to the central point in a 21 x 21 regular grid in a field, or equivalently correspond to the 221th column in the matrix ***V***. Due to the induced heteroscedasticity in the SAR-error model, covariance values for other points would differ slightly from the ones shown here.



Yield 1999

Yield 2001

Soil sample grid (P & K)

Elevation data

Grid 20m x 20m

Figure 3. Data used in the example on regression application to precision agriculture: Grain yield for both years, soil sample grid and CTI map overlain onto the topographical relief surface. The arrows represent the operation of ordinary kriging to a common point grid. The support used in the spatial estimation was approximately the same for all datasets.

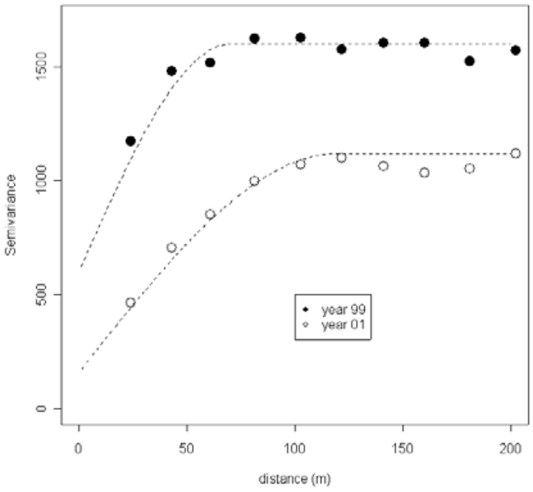


Figure 4. Experimental semivariograms of OLS residuals for both years, with corresponding spherical models (dotted line) fitted to them.

	OLS			SAR-error Queen			SAR-error d 50			DCR (noiter)			DCR (REML)		
Effect	**beta**	**s.e.**	**p-val.**	**beta**	**s.e.**	**p-val.**	**beta**	**s.e.**	**p-val.**	**beta**	**s.e.**	**p-val.**	**beta**	**s.e.**	**p-val.**
Intercept 99	528.6	39.06	<.001	569.7	50.93	<.001	559.4	48.48	<.001	549.0	52.10	<.001	531.6	50.20	<.001
P 99	0.60	0.25	0.02	0.64	0.40	0.11	0.53	0.38	0.16	0.62	0.41	0.13	0.59	0.37	0.11
K 99	-0.28	0.07	<.001	-0.32	0.11	0.00	-0.32	0.11	0.00	-0.30	0.12	0.02	-0.29	0.11	0.01
Prof 99	-283.7	79.20	0.00	-265.8	79.64	0.00	-261.8	78.98	0.00	-251.4	76.56	0.00	-276.6	80.43	0.00
Plan 99	-337.0	47.57	<.001	-363.0	48.04	0.00	-342.6	47.71	0.00	-341.1	45.88	<.001	-344.2	48.36	<.001
Spi 99	1.10	0.54	0.04	1.93	0.59	0.00	1.76	0.58	0.00	2.03	0.57	0.00	1.67	0.56	0.00
Cti 99	-20.46	3.87	<.001	-25.02	4.36	0.00	-23.12	4.17	0.00	-22.65	4.38	<.001	-20.79	4.52	<.001
Ang 99	5.81	1.50	0.00	6.77	1.78	0.00	6.58	1.67	0.00	5.84	1.78	0.00	5.97	1.88	0.00
Intercept 01	**179.7**	31.72	<.001	**330.1**	46.32	0.00	**358.8**	45.66	0.00	**298.1**	42.92	<.001	**339.9**	55.85	<.001
P 01	**1.08**	0.20	<.001	**0.14**	0.39	0.72	**0.08**	0.34	0.80	0.59	0.39	0.14	0.31	0.51	0.55
K 01	**-0.12**	0.06	0.04	**-0.24**	0.13	0.06	**-0.35**	0.11	0.00	-0.24	0.12	0.06	-0.29	0.17	0.08
Prof 01	**36.4**	64.32	0.57	-86.2	47.75	0.07	**-130.0**	49.58	0.01	**-97.9**	39.66	0.01	**-96.0**	47.02	0.04
Plan 01	**3.95**	38.63	0.92	**-93.4**	28.86	0.00	**-103.4**	30.07	0.00	**-83.0**	24.22	0.00	**-87.8**	28.12	0.00
Spi 01	**0.98**	0.44	0.03	**-0.25**	0.37	0.50	**-0.48**	0.37	0.20	**-0.07**	0.30	0.82	**-0.01**	0.36	0.97
Cti 01	**10.35**	3.14	0.00	**-0.02**	2.80	0.99	**2.12**	2.74	0.44	**1.51**	2.47	0.54	**-0.38**	2.91	0.90
Ang 01	1.17	1.22	0.34	1.86	1.16	0.11	1.42	1.10	0.20	2.24	0.98	0.02	1.74	1.19	0.14

Table 1. Regression parameter estimates, standard errors and p-values for OLS, SAR-error and direct covariance representation (DCR) corresponding to model (7) for both years. Shaded parameter estimates are those showing the largest absolute differences with OLS parameter estimates. .

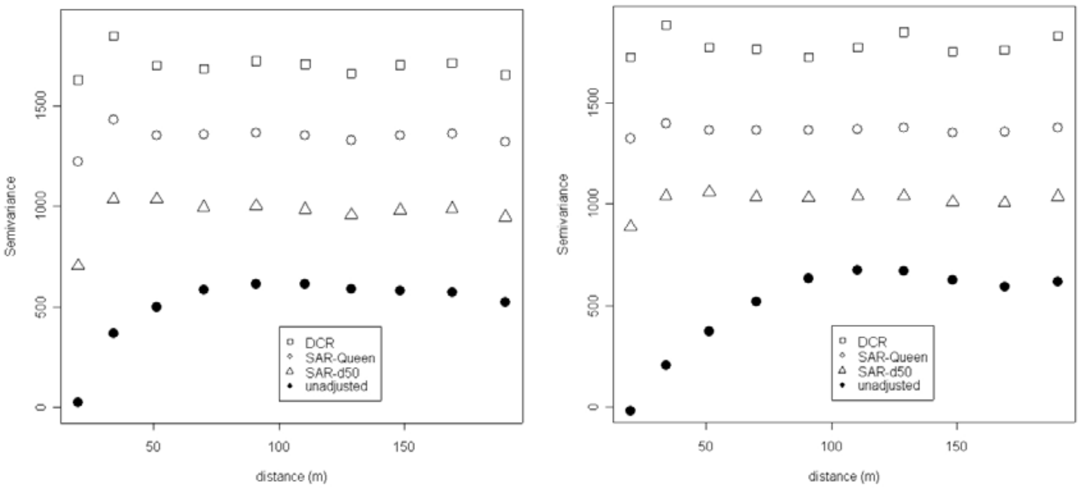


Figure 5. Experimental semivariograms of residuals for 1999 (left) and 2001 (right). DCR: direct covariance representation. 'Unadjusted' corresponds to OLS residuals. The sills for the 4 variograms have been shifted vertically to allow for a better comparison of the spatial structure among the 4 cases.

	μ	σ_0^2	σ_s^2	a
X_1	5	0.5	2	80
X_2	5	0	1.4	50
ε	0	0	0	70

Table 2. Mean and covariance parameters used to generate the spatial random fields for the Monte Carlo simulation.

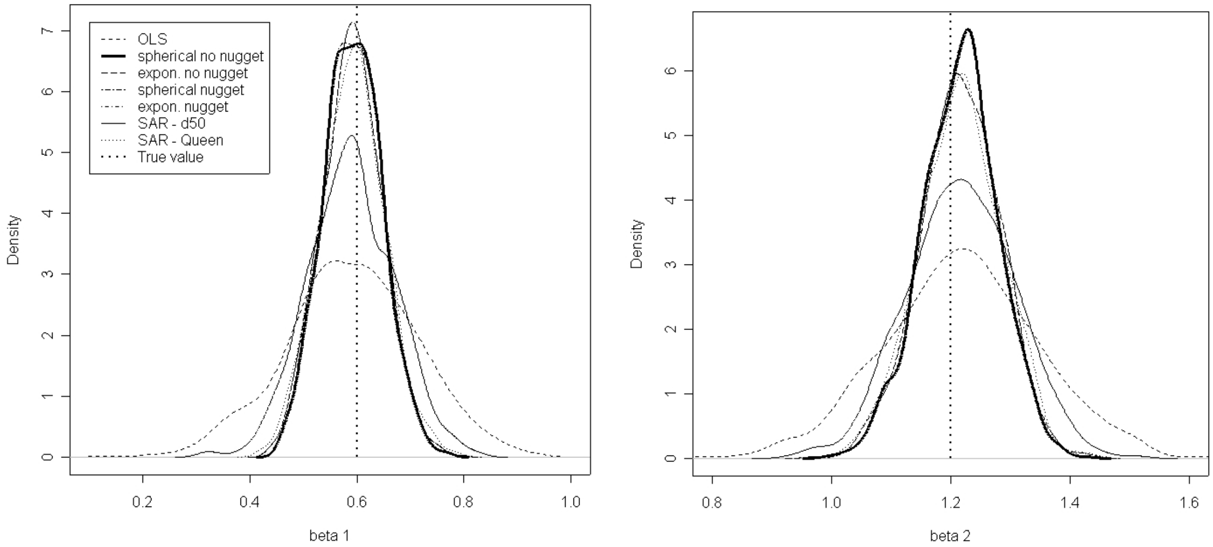


Figure 6. Empirical distributions of the estimates for β_1 (left) and β_2 (right) for the 1000 simulated spatial random fields. The dashed vertical line represents the true value for each parameter.

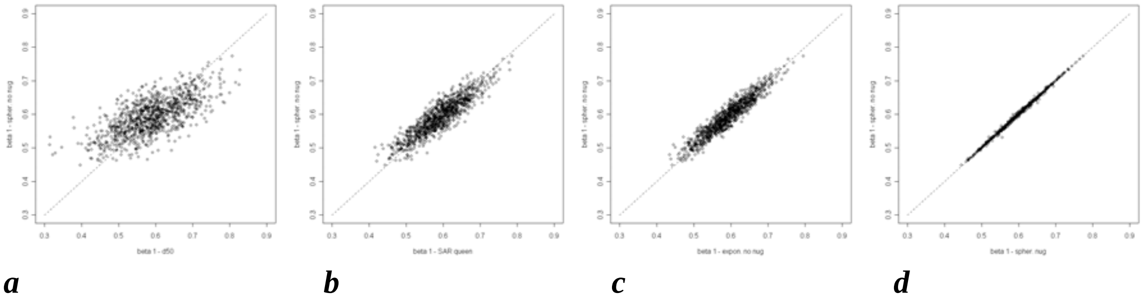


Figure 7. Scatterplots for estimates of β_1 at each of the 1000 simulations obtained with the true ***V*** structure (spherical., no nugget) against those estimates obtained with alternative specifications. *a*: SAR-error d50 *b*. SAR-error queen. *c*. Exponential-no nugget. *d*. Spherical-nugget.

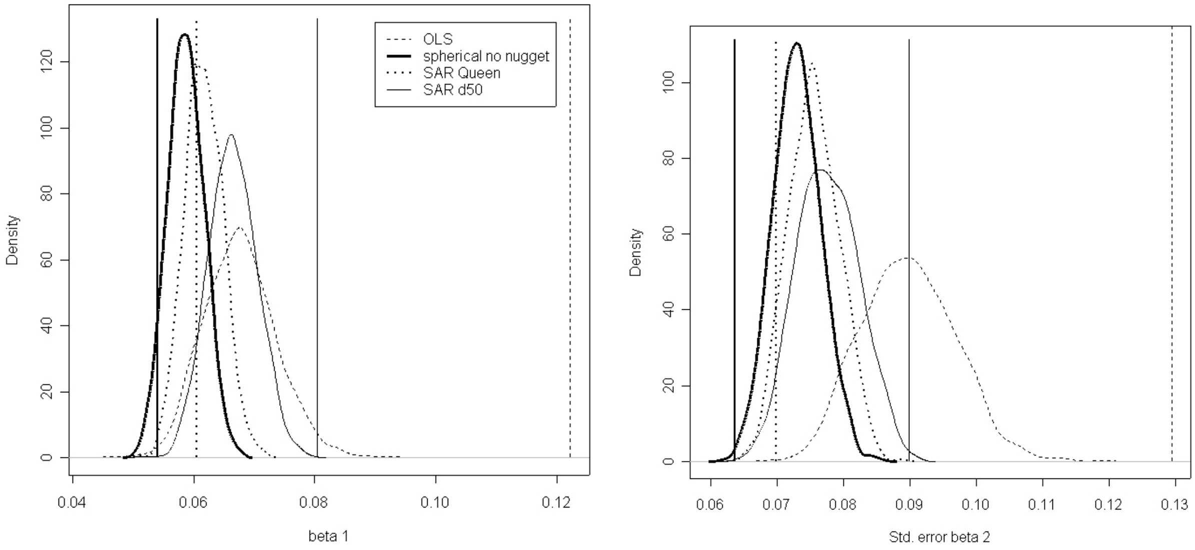


Figure 8. Distributions of the reported standard errors for β$_1$ (left) and β$_2$ (right) for the 1000 simulated spatial random fields. The vertical lines represent the standard errors computed from the empirical distributions of each parameter shown in Figure 6.