

# USE OF CROSS REGRESSION TO MODEL LOCAL SPATIAL AUTOCORRELATION IN PRECISION AGRICULTURE

**J.M. Lowenberg-DeBoer, T.W. Griffin, and R.J.G.M. Florax**

*Site-Specific Management Center  
Dept. of Agricultural Economics  
Purdue University  
West Lafayette, Indiana*

## ABSTRACT

Analysis methods for field-scale precision agriculture datasets are being developed by adapting statistical procedures from a wide range of disciplines. Due to the spatial nature within agriculture fields with respect to yield and factors of production, inference from aspatial analysis is unreliable. When spatial correlation exists within a factor of production or explanatory variables, e.g. elevation, soil characteristics, spatial analysis methods are employed to provide unbiased and efficient coefficients from which to base economic analysis and farm management decisions. We present a simple continuous terrain variable derived from cross-regressive spatial process models that models relative terrain position and allows models to be easily estimated with familiar estimators. In example data sets, cross regressive elevation variables complement other topographic variables, rather than replacing them.

**Keywords:** cross regression, landscape position, terrain, spatial analysis

## INTRODUCTION

The introduction of Global Positioning Systems (GPS) into agriculture production systems have allowed farmers to revisit the old idea of testing input choices before implementing farm management decisions across large areas of the farm. Yield monitor technology has driven the resurgence of on-farm testing because farmers can collect yield measurements without interfering with harvest or other harvest-time field operations. Farmers are making decisions based on yield monitor data. It is the general goal of our work to help them make better, more reliable decisions based on that data.

To analyze yield monitor data in a statistical framework, supporting factors of production are needed for inference. These supporting or explanatory variables include treatment and soils characteristics including elevation and terrain. Landscape position and other terrain attributes are leading factors in explaining site-specific crop productivity, variability, and input application response.

Topographic modeling techniques applied to statistical models include: advanced hydrologic models, indexes of variables, digital elevation models (DEM), elevation as a simple covariate, and derivatives of the elevation surface such as slope, aspect, and curvature. Even though elevation may have no direct interpretation with respect to the dependent variable (e.g., yields, productivity), the elevation covariate typically explains much of the noise component, earning its way as one of the most common topography variable regimes in the literature. An additional advantage of the elevation variable is that since the advent of GPS providing automated guidance farm operators have access to accurate, high resolution and low-cost elevation data collected from field operations.

Problems associated with omitting topography variables or including variables measured incorrectly may be circumvented with the appropriate spatial econometric techniques. We present a method to model relative terrain position by using cross-regressive variables to create a new relative elevation variable not requiring any spatial interpolation methods.

The overall objective of this study is to present a simple yet elegant method to model micro-scale landscape position with a minimal number of variables. Specific objectives were: 1) to offer a relative micro-scale terrain position variable created from a cross-regressive elevation variable and 2) to compare traditional, cross-regressive, and spatial econometric models of empirical on-farm trials. Using spatial econometric techniques that model local spatial autocorrelation, we model relative elevation, slope, and overall micro-scale landscape position with a limited number of continuous covariates. Hypotheses include 1) model specifications including relative elevation variables are no better able to estimate treatment differences or optimal rates than with conventional elevation variables and 2) model specification with the proposed cross regressive variable increases the multicollinearity condition number.

This study demonstrated the implications of using differing topography variables for spatial data analysis of empirical field-scale on-farm comparisons, and assesses the effectiveness of various alternative specifications. The field scale experiments are conducted by farmers in collaboration with the authors. The research questions, i.e. treatments, are pertinent to and chosen by the farmer. The choice of treatments and experimental design was ultimately the farmer's with only guidance from the authors.

## **BACKGROUND**

It is known that relative terrain position influences productivity. Elevation and other topographic information have been used in precision agriculture studies for three broad categories, 1) identification of management zones, 2) empirical crop modeling, and 3) soil mapping (Bishop and McBratney, 2002). Bishop and McBratney's (2002) third point is evident with USDA-NRCS soil mapping units being defined by slope class categories. Points 1 and 2 are of interest to farmers now that elevation measurements are easier to collect. Agricultural technology innovations for data gathering (e.g. combine yield monitors and other site-specific sensors) and navigation (e.g. lightbars and automated guidance) may provide sufficiently accurate elevation measurements for use as covariates in econometric

models. It is only with the highest levels of GPS accuracy that good topographic maps can be developed (Clark and Lee, 1998).

Elevation data is important in estimation of treatment effects in datasets acquired from fields with micro-scale topography differences. In cases such as precision leveled fields in high value crops such as rice (Griffin et al., 2006) and cotton (Griffin et al., 2005a), elevation would not be considered as an important covariate.

Although elevation has been successfully used as a covariate in field-scale precision agriculture datasets, the elevation variable alone usually can not adequately model the relative terrain position of an observation. Even within the same field, an elevation measurement of 200 meters may be 1) on a hilltop, 2) valley bottom, and 3) hill slope. Advanced elevation modeling techniques interpolate elevation into a so-called digital elevation model (DEM) surface from which slope and other elevation derivatives are created, e.g. plan curvature, profile curvature, and aspect. Although spatial interpolation has its place and elevation derivatives have been useful for many soils and crop modeling procedures, they are not as useful in econometric modeling for two reasons. The first reason is that the elevation measurements must be interpolated into a surface thus introducing a systematic random variable into the data (Anselin, 2001). Unlike random errors, systematic error affects the average of the explanatory variable and biases the estimated coefficient. Second, estimation of econometric models suffers from too many continuous covariates especially when several variables are linearly or nonlinearly dependent resulting in multicollinearity. Multicollinearity is the inter-correlation among explanatory variables in an econometric model. A small number of variables that model relative terrain position without interpolation would be useful to spatial analysis of field-scale precision agriculture datasets.

If slope or other variable created from an interpolation process is conceptually important to the statistical model, an omitted variable problem results from its exclusion potentially leading to biased estimated coefficients. Conversely, if an interpolated surface from a sparse data layer, e.g. elevation or soil fertility measurements, is used as an explanatory variable, errors in variables may result. One potential method to avoid the errors in variable and omitted variable problems is the use of cross-regressive variables.

Cross-regression is a spatial process model that includes spatially weighted exogenous variables and can be estimated with ordinary least squares (OLS) (Anselin, 2002; Florax and Folmer, 1992). Cross-regressive models are an extension of the familiar aspatial model,  $y = X\beta + \mu$ , and are given as  $y = X\beta + WZ\gamma + \mu$  where  $y$  is an  $n \times 1$  vector of observations on the dependent variable,  $X$  is an  $n \times k$  matrix of explanatory variables,  $\beta$  is a  $k \times 1$  vector of regression coefficients,  $W$  is an  $n \times n$  spatial weights matrix,  $Z$  is a  $k^* \times 1$  matrix of  $k^*$  explanatory variables that can be the same as in  $X$  except without the intercept term,  $\gamma$  is a  $k^* \times 1$  vector of regression coefficients on the cross-regressive term  $WZ$ , and  $\mu$  a iid error term. Cross-regression explicitly models local spillovers. When the true model specification includes the  $WZ$  term but is estimated without it with OLS, the expected value of estimates remain unbiased and efficient, however an omitted variable problem results.

In addition, the spatially weighted exogenous variable can be included in other spatial process models such as the spatial autoregressive error (SAR) model. The SAR model is another spatial process model that explicitly models spatial autocorrelation in the error term. For precision agriculture datasets, SAR models were expected rather than other spatial models because spatial correlation is usually due to some omitted variable (e.g. subsoil characteristics, microclimate), rather than effect of plants on each other. Spatial diagnostics performed on OLS residuals support this assumption. The SAR model is given as  $y = X\beta + \varepsilon$ ,  $\varepsilon = \lambda W\varepsilon + \mu$  or in reduced form as  $y = X\beta + (I - \lambda W)^{-1}\mu$  where  $y$  is a  $n \times 1$  vector of dependent variables,  $X$  a  $n \times k$  matrix of explanatory variables,  $\beta$  a  $k \times 1$  vector of regression coefficients,  $\varepsilon$  an  $n \times 1$  vector of residuals,  $\lambda$  a spatial autoregressive parameter,  $\mathbf{W}$  is an  $n \times n$  spatial weights matrix, and  $\mu$  a well behaved, non-heteroskedastic uncorrelated error term (Anselin, 1988). The  $(I - \lambda W)^{-1}$  term is the so-called spatial multiplier. When the spatial autoregressive term,  $\lambda$ , is 0, the spatial error model reverts to the familiar aspatial model,  $y = X\beta + \mu$ . The spatial error process can be characterized by the global spillovers due to the spatial multiplier. The spatial error model has no substantive economic interpretation, i.e. estimated parameter coefficients are asymptotically unaffected. When the spatial error model is appropriate, OLS estimates remain unbiased but are inefficient.

## METHODS

This study conducted a series of spatial analyses on empirical field-scale on-farm trials to demonstrate the usefulness of alternative topographic variables. To evaluate localized terrain effects on yield response and treatment effects, we create cross-regressive variables,  $\mathbf{WZ}$ , for each dataset. The first step in this procedure was to choose the spatial interaction structure for use in calculating the spatially weighted elevation term. In general, spatial weights matrices are constructed such that  $w_{ii} = 0$ ,  $w_{ij} > 0$  for observations considered neighbors, and  $w_{ij} = 0$  for non-neighbors where  $w$  is an element of  $W$  and  $ij$  denotes the matrix position. The spatial weights matrix for local terrain effects (hereafter referred to as  $W_1$ ) was selected on the basis that only immediately neighboring observations are of interest so specifications such as first-order queen contiguity or minimum Euclidean distances are used. In either case, these binary matrices were constructed with zeros as non-neighbors and ones as neighbors then row-standardized in GeoDa (Anselin, 2003).

The  $n \times n$  spatial weights matrix is multiplied by the  $n \times 1$  elevation vector,  $\mathbf{E}$ , producing the  $n \times 1$  cross-regressive term  $W_1E$ , where  $E$  is the continuous elevation variable. This cross-regressive term measures the spatially weighted average elevation of the immediate neighbors as defined by the spatial weights matrix,  $W_1$ . Rather than including the spatially weights average elevation,  $W_1E$ , variable in the econometric model essentially resulting in a smooth elevation variable,  $W_1E$  was used to create a relative elevation variable. The cross-

regressive term,  $W_1E$ , was subtracted from the elevation value of the observation in question providing relative elevation,  $RE = E - W_1E$ .

The new relative elevation, RE, variable captures localized terrain position for use in the econometric model. For instance, when  $RE < 0$  the observation is lower in elevation than the average of its immediate neighbors. When  $RE > 0$  the observation is higher than the average of its neighbors, and is at the same elevation as the average of the neighbors when  $RE = 0$ . When  $RE = 0$ , the observation could either be on a flat plain or hillside such that the average of the neighbors equates to the elevation of the observation. This is a limitation of the relative elevation variable compared to slope which distinguishes observations on hillsides. However, slope is unable to model the difference between hilltop and valley which may be more important to distinguish than observations on hillsides. Not only does the RE variable give indication of direction of relative position but also provides the magnitude of difference. Since terrain slopes are generally calculated from interpolated elevation surfaces, the magnitude of  $RE$  partially substitutes for slope allowing the model to be estimated without the systematic error associated with interpolated values.

To analyze the field-scale precision agriculture datasets, aspatial, cross-regressive, and spatial analyses were conducted. All analyses assume yield monitor data has been appropriately filtered to remove erroneous measurements and observations are at correct locations (Griffin et al., 2005b). Aspatial and cross-regression analyses were performed with the OLS estimator, and differed by the cross regression model specification included RE. The SAR model was estimated with the general moments (GM) estimator for all model specifications. The GM estimator was chosen due to the large sample sizes of these field scale experiments and no assurance of normal error distribution (Kelejian and Prucha, 1999, 2006; Bell and Bockstael, 2000). Sample sizes of precision agriculture datasets reported in this paper are all over 1,000 observations and average over 3,000 observations. Spatial process models estimated with the common maximum likelihood (ML) estimator uses an eigenvalue computation from the Jacobian matrix which loses numerical precision beyond 1,000 observations (Anselin, 1988; Kelejian and Prucha, 1999; Bell and Bockstael, 2000). Kelejian and Prucha (1999) go on to say that the eigenvalues for spatial weight matrices under ML estimation were not correctly estimated with more than 400 observations. Due to the computational limitations of ML for large sample-size datasets, GM estimators are considered because estimation can be conducted for very large datasets of several thousand observations. However, the specification of the spatial weights matrix may influence parameter estimation more than the choice of estimator (Bell and Bockstael, 2000).

As opposed to the spatial weights matrix for cross-regressive variable,  $W_1$ , an inverse distance spatial weights matrix  $w_{ij} = \frac{1}{d_{ij}}$  hereafter referred to as  $W_2$ , was chosen for the SAR model. Although the specification of  $W_2$  was chosen for field-scale precision agriculture datasets from *a priori* information, the distance band in which any observation were no longer correlated and thus observations not considered neighbors were empirically determined by Lagrange Multiplier

(LM) tests on OLS residuals. In order to determine the appropriate distance band for  $W_2$ , a series of row-standardized inverse distance weights matrices were created, each with a differing cutoff distance or distance band. The cutoff distances ranged from below a reasonable level to a distance such that no spatial correlation is expected to exist in the data.

OLS estimation was conducted on the full econometric model and LM spatial diagnostics performed on the residuals with each of the spatial weights matrices. The highest chi-squared coefficients of the LM tests indicated the proper distance band for  $W_2$  (Cressie, 1993; Anselin 1988). The most appropriate cutoff distance for each empirical example was used to create the spatial weights matrix for the SAR model.

Model specifications can be compared by the Akaike Information Criterion (AIC) (Anselin, 1988; Greene, 2000). The AIC degrades as model size increases, i.e. a penalty placed on increase numbers of explanatory variables. The GM routines ran in the R software (R Development Core Team, 2006), with sigma squared ( $\hat{\sigma}_k^2$ ) reported. The measure of fit was calculated as  $AIC = N(\ln 2\pi\hat{\sigma}_k^2 + 1) + k$ , where N was the number of observations and k the number of variables.

Empirical examples include both a categorical and rate trial conducted by farmer collaborators. The rate trial was a soybean seeding rate study (hereafter referred to as SOYSEED). The categorical trial included a popcorn insecticide and fungicide treatments (hereafter referred to as SEEDTRT). Results from each dataset are reported after a demonstration of the spatial correlation of RE with other variables.

## RESULTS

It was suspected that terrain attributes were correlated with yield. This was demonstrated by measurements taken from the Western Illinois University demonstration farm in McDonough County, Illinois hereafter referred to as WIU. The 160 hectare field has the highest level of elevation data quality available, a combination of 286 total station and 1,068 RTK-GPS survey measurements. Total stations electronically collect distance measurements and angles between the base and a given location with which relative elevation can be calculated. Eight locations were measured by both the total station and RTK-GPS to align measurements, resulting in 1,346 total elevation observations.

A total of 3,859 electrical conductivity (EC) measurements were georeferenced on 20-meter transects in April 2002. The topography, EC, and YIELD data were combined into a single dataset resulting in 1,075 observations. The final number of observations were less than the most sparse data layer (N=1,346) because not all data layers had observations within a reasonable vicinity (see Griffin et al., 2005b for discussion of disparate spatial data layer assimilation).

Although no on-farm experiment for this field is reported in this paper, this data was useful to demonstrate the spatial correlation among YIELD and elevation variables. Univariate and bivariate Moran's I measures global spatial autocorrelation between variables and was calculated by using immediately neighboring observations as defined by the spatial weights matrix. Although

YIELD had a high level of spatial autocorrelation of 0.70, elevation (ELEV) had a very high value of 0.97 (Table 1). Electrical conductivity (EC) also had a high Moran's I (0.81). The bivariate Moran's I values between YIELD and ELEV (0.21) and EC (-0.35) show moderate spatial association. Spatial autocorrelation between RE and the other variables were small, but significant.

Although bivariate Moran's I for RE with the other three values were significantly different from zero, the magnitude of the spatial autocorrelation was small. Low levels of spatial autocorrelation indicates that RE may be a candidate explanatory variable in an aspatial model. When an explanatory variable is spatially autocorrelated with the dependent variable, itself, or other explanatory variable, residuals from the aspatial regression model will be spatially autocorrelated resulting in inefficient OLS estimation.

### Popcorn Seed Treatments

Seven combinations of seed applied insecticides and fungicides were compared on irrigated popcorn production in a pseudo-replicated strip-trial experimental design for SEEDTRT in Tazewell County, Illinois. The 10-hectare experiment was planted with two passes of an 8-row planter with the control treatment (CHECK) between each of the six treatments and either side of the experiment. Each treatment strip was harvested by two combine harvester passes. Yields are reported in Mg ha<sup>-1</sup>.

Treatment X1 and X2 were two recommended rates of the same insecticide. Treatment X3 was the fungicide. Treatments X4 and X5 were combinations of X3 with X1 and X2, respectively. Treatments X5 and X6 were two recommended rates of another insecticide. Farmer expectations included Treatment X6 dominating the other treatments from *a priori* expectations and experience in the field.

The full model specification (FULL) included binary variables for treatments (Xi), soil binary variables (Si) elevation (E), elevation squared (E2), RE, and interaction terms between elevation and treatments (EXi). The second model specification (EL) omitted RE. The WE model specification is the same as FULL except the RE variable was replaced by the cross-regressive variable WE. The fourth model specification (RE) included RE but omitted all other topography variables.

Table 1. Bivariate Moran's I between YIELD and topographic variables for WIU

Spatially lagged variable	Variable			
	YIELD	ELEV	EC	RE
Yield	0.70	0.21	-0.35	0.05
ELEV	0.21	0.97	-0.54	0.10
EC	-0.35	-0.54	0.81	-0.08
RE	0.03	0.07	-0.05	0.07

N=1,075

All values significantly different from zero at 1% level except for YIELD with spatially lagged RE significant at the 5% level

Table 2. Regression results for SEEDTRT

Variable	OLS	SAR	OLS	SAR	OLS	SAR	OLS	SAR
	FULL	FULL	EL	EL	WE	WE	RE	RE
Intercept	5.898***	0.727***	5.894***	0.744***	5.903***	0.437***	5.686***	1.569***
X1	-0.211	0.082	-0.173	0.029	-0.211	0.049	-0.101**	-0.096
X2	-0.061	0.950***	-0.078	0.996***	-0.061	0.571***	0.071	-0.011
X3	0.067	0.387***	0.057	0.438***	0.067	0.232***	0.251***	-0.013
X4	-0.024	0.989***	-0.068	1.116***	-0.023	0.594***	-0.111**	0.353***
X5	0.146	-0.012	0.114	0.072	0.146	-0.007	0.009	-0.001
X6	0.248***	0.713***	0.198	0.835***	0.247**	0.428***	0.110**	0.318***
S1	-0.236***	0.063	-0.278***	0.197*	-0.237***	0.038	-0.596***	3.868***
S2	0.073	0.290**	0.043	0.420***	0.073	0.174**	-0.241***	4.178***
S3	-1.452***	-0.145	-1.484***	-0.029	-1.454***	-0.087	-1.698***	3.119***
S4	-0.407	-0.303	-0.455	-0.156	-0.407	-0.181	-0.658**	3.805***
EX1	0.01	-0.022	0.006	-0.018	0.010	-0.013		
EX2	0.019	-0.099***	0.017	-0.099***	0.019	-0.059***		
EX3	0.022	-0.041**	0.022	-0.043**	0.023	-0.025**		
EX4	-0.013	-0.129***	-0.008	-0.145***	-0.013	-0.078***		
EX5	-0.018	0.012	-0.015	0	-0.018	0.007		
EX6	-0.019	-0.047***	-0.015	-0.061***	-0.019	-0.028***		
E	-0.114***	0.734***	-0.109***	0.720***	-0.005	0.325***		
E2	0.006***	-0.024***	0.006***	-0.025***	0.006***	-0.014***		
RE	0.110***	-0.193***					0.091***	0.092***
WE					-0.110***	0.116***		
AIC	20,822	20,056	20,833	20,105	20,822	20,056	20,846	21,585

Significance denoted at 1, 5, 10% levels by \*, \*\*, and \*\*\*, respectively

Spatial diagnostics conducted on the OLS residuals indicated that spatial autocorrelation was in the data. A significant Moran's I value of 0.26 indicated global spatial autocorrelation. LM tests indicated the spatial autocorrelation was present in the dependent variable (lag) and residuals (error). Robust LM tests were also significant but the LM and Robust LM tests had higher chi-squared coefficients for spatial error than for spatial lag indicating the error model was most appropriate to analyze the data.

In the FULL, EL, and WE model specifications, OLS results indicated only Treatment X6 was statistically different from the control treatment, while OLS estimation of the RE model indicated Treatments X1, X3, X4, and X6 were statistically different from the control plot (Table 2). This was similar to results from SAR estimation of the models where Treatments X2, X3, X4, and X6 were statistically different from the mean for FULL, EL, and WE models and Treatments X4 and X6 were under the RE model specification.

The rankings within the spatial autoregressive error (SAR) models more closely resembled farmer expectations than the OLS estimator when evaluated at the mean elevation (Table 3). With the SAR model with Treatment X6 dominated for the FULL, EL, and WE models. The RE model resulted in a different agronomic and economic ranking with Treatment X6 ranked second. The FULL and WE model specifications produced the same agronomic rankings and only a few

Table 3. Topography variable and estimator influence on treatment rankings

	OLS FULL		SAR FULL		OLS EL		SAR EL		OLS WE		SAR WE		OLS RE		SAR RE	
	Yld	Ecn	Yld	Ecn	Yld	Ecn	Yld	Ecn	Yld	Ecn	Yld	Ecn	Yld	Ecn	Yld	Ecn
Check	4	3	5	3	4	3	5	4	4	3	5	4	5	3	3	3
X1	7	6	7	7	6	6	7	7	7	6	7	7	6	6	7	5
X2	3	4	2	4	3	4	2	3	3	4	2	3	3	4	5	6
X3	1	1	4	2	1	1	3	2	1	1	4	2	1	1	6	4
X4	6	5	6	5	7	5	6	5	6	5	6	6	7	5	1	1
X5	5	7	3	6	5	7	4	6	5	7	3	5	4	7	4	7
X6	2	2	1	1	2	2	1	1	2	2	1	1	2	2	2	2

switches for the economic rankings for both OLS and SAR estimation. Although the RE model specification clearly did not dominate the other models based on AIC, the inclusion of the RE variable in the FULL model was beneficial to the overall model fit for both OLS and SAR estimation based on AIC. Since elevation by treatment interaction terms was usually significant under SAR estimation, treatment rankings were sensitive to elevation and terrain position.

### Soybean Seeding Rates

Five soybean seeding rates were replicated four times in a 19-hectare strip-trial design with two harvester passes wide per treatment strip in Montgomery County, Indiana. Seeding rates included a very low rate (197,600) to a reasonably high rate (395,200) in increments of 49,400 seeds ha<sup>-1</sup>. Automated guidance with RTK-GPS on the planter tractor collected elevation data. Yields are reported in Mg ha<sup>-1</sup>.

The full model specification (FULL) include seeding rate, rate squared, elevation, elevation squared, RE, and an interaction term between rate and elevation (Table 4). The POP model specification omitted all topography variables and included only rate and its square. The WE model specification is the same as FULL except the RE variable was replaced by the cross-regressive variable WE. The remaining model specification, EL, dropped RE from FULL. In any case, a Moran's I value of 0.38 and significant LM diagnostics on the OLS residuals from the FULL model indicated that the spatial error model was most appropriate. The LM chi-squared coefficients for the FULL model specification were actually higher than for the EL model, thus the SAR model was chosen.

Small changes in agronomic yield resulted between seeding rates estimated from SAR-GM and other model specifications and/or estimators. If the agronomic optimal seeding rate from any of the estimators or model specifications were applied, economic returns were very similar to SAR-GM with the full model specification.

For economic analysis, the choice of estimator impacted the optimal decision (Table 5). In several model specifications, the economic analysis using OLS regression results calculated the optimal seeding rate below the range of rates used in the experiment. In these cases, the range of seeding rates was constrained to be within the 197,600 to 395,200 per hectare range. Reasonable physical agronomic optimal rates were estimated with OLS, but the unconstrained economic analysis did not result in feasible solutions.

Table 4. SOYSEED econometric results

Variable	OLS	SAR	OLS	SAR	OLS	SAR	OLS	SAR	OLS	SAR
	FULL	FULL	EL	EL	WE	WE	RE	RE	POP	POP
Constant	3.686***	0.022	3.676***	0.015	3.603***	0.022	4.136	0.120**	4.134	0.124**
POP	0.001	0.059***	0.000	0.058***	-0.052**	0.059***	0.003	0.072***	0.003	0.072***
POP_SQ	0.000*	0.000***	0.000*	0.000***	0.000	0.000***	0.000	0.000***	0.000	0.000***
NELEV	0.116***	0.132***	0.117***	0.133***	0.184***	0.181***				
E2	-0.005***	-0.005***	-0.005***	-0.005***	-0.005***	-0.005***				
POP_ELV	0.000***	0.000	0.000***	0.000	0.000**	0.000				
RE	0.051***	0.048***					-0.022	0.011***		
WE					-0.061***	-0.048***				
AIC	23,954	21,461	23,969	21,479	23,960	21,461	24,992	21,731	24,992	21,731

Significance denoted at 1, 5, 10% levels by \*, \*\*, and \*\*\*, respectively

The AIC goodness-of-fit rankings within the GM estimator resulted in the FULL and WE model specifications being superior followed by the EL model. With the AIC smaller is better. With OLS, the AIC rankings held FULL superior to WE and WE superior to EL. The AIC value for RE and POP model specifications were identical, indicating that the RE variable on its own was not beneficial to the model in this case. The SAR model dominated the aspatial and cross-regressive models in every model specification. Overall, regression model diagnostics indicated that the RE model was not useful in this dataset.

#### Multicollinearity Condition Number

Compared to full elevation models that include elevation, its square, and interaction terms, model specifications with RE as the only topographic variable had minimal multicollinearity condition numbers. One way of measuring multicollinearity is the condition number (CN) of a matrix  $\mathbf{A}$  which is a ratio of the largest and smallest characteristic roots or eigenvalues of the matrix

$$CN = \left[ \frac{\max \text{root}}{\min \text{root}} \right]^{0.5} \text{ (Greene, 2000). Values larger than 20 are considered large}$$

meaning that the matrix is nearly singular (Greene, 2000) however multicollinearity typically is not a problem as long as coefficient remain robust. The larger the condition number, the more difficult it is to invert the matrix. In econometrics, the level of multicollinearity in the matrix of explanatory variables  $\mathbf{X}$  was of interest, so the condition number of the cross product of  $\mathbf{X}$  ( $\mathbf{X}\mathbf{X}$ ) was calculated.

Table 5. Recommended seeding rates and economic returns for SOYSEED

	OLS FULL	SAR FULL	OLS EL	SAR EL	OLS RE	SAR RE	OLS POP	SAR POP
Yield max ate (000's seeds ha <sup>-1</sup> )	280.3	307.3	282.3	299.1	267.7	305.3	266.2	305.5
Economic max rate (000's seeds ha <sup>-1</sup> )	197.6	265.7	197.6	265.5	197.6	271.2	197.6	271.2
Economic returns by model <sup>a</sup> (\$ha <sup>-1</sup> )	320	337	320	337	320	337	320	337

N=3,897

<sup>a</sup> Economic returns are returns to fixed costs minus seeding if estimated coefficient from SAR FULL model were used as "true" model in economic analysis.

Table 6. Multicollinearity Condition Number for selected studies

Study	SEEDTRT	SOYSEED
FULL	45	133
EL	45	133
WE	82	154
RE	7	92
No topo	7	92

Model specifications including only the relative elevation variable, RE, substantially reduced the multicollinearity condition number compared to the model specifications using elevation, its square, and interaction terms (EL) (Table 6). The classic cross regressive term WE increased the multicollinearity condition number for both data sets. The full model specification (FULL) including all topography variables (RE, elevation, elevation squared) and the EL model had the same multicollinearity number. In the SOYSEED seeding rate trial, dropping the RE variable leaving only the seeding rate and its square as the only explanatory variables in the model, no difference in condition number occurred.

## CONCLUSIONS

Cross-regressive variables were somewhat useful in modeling field-scale precision agriculture datasets, however rather than substituting for conventional terrain variables such as elevation, its square and interaction terms, the WE and RE variables complemented these variables in the examples studied. The proposed relative elevation (RE) variable did not strictly dominate nor was it strictly dominated by the other model specifications including conventional topography variables. This was demonstrated by the AIC regression diagnostics and the low bivariate Moran's I values for RE relative to other continuous variables.

When the AIC goodness-of-fit criteria were compared, the FULL and WE model specifications dominated the EL model specification which dominated the RE model specification. AIC diagnostics suggests that the WE and RE variables contribute more to model estimation than was potentially penalized by over specification of the model, with the WE variable being better than RE. However, model specifications without the conventional topography variables were

dominated by the FULL and EL model specifications. While the AIC measures suggest model specifications benefit from inclusion of RE, RE was not capable of replacing the conventional topography variables but rather complementing their use. Thus, hypothesis 1 was not supported.

In the empirical examples, models with the cross regressive relative elevation variable have lower multicollinearity condition numbers than models which use elevation and elevation interaction terms directly. Condition numbers indicated that RE variables did not cause additional multicollinearity to the model as did WE. If multicollinearity is suspected to be a problem in a dataset, the use of the RE variable rather than the WE variable may be considered. Thus, hypothesis 2 is not supported.

Models that already suffer from near singular explanatory matrices may benefit from including relative elevation rather than the conventional topography variables. The relative elevation variable may be able to avoid other theoretical econometric problems of using interpolated values, errors in variables, or not including a conceptually important variable, omitted variable problem. Further cost:benefit analysis on use of interpolated values versus specification of the spatial weights matrix is needed.

In both of the example data sets, spatial diagnostic statistics pointed to use of a spatial error model even when the elevation cross regressive terms were used. It is possible that for some datasets cross regressive independent variables would model spatial structure well enough to avoid use of spatial error models and the estimation problems they entail. Cross-regression can be carried out in most any statistical software package and is much simpler than other spatial process models. In the examples presented here only elevation was treated as a cross regressive variable. It is possible that having additional independent cross regressive variables would improve error structure modeling.

For relative elevation or any topographic variable to be useful as a covariate, high quality GPS signals are needed. Even with RTK-GPS automated guidance equipment, not all elevation measurements are reliable enough to include in the model and some data cleaning procedure is recommended. Elevation measurements are still recorded when base station signals become weak and diminishing data quality may not be readily apparent in the database or maps. It was noted that suspect features of the relative elevation maps were identified even when the elevation map was particularly smooth. The slight differences in elevation measured while equipment moved in opposite directions may be enough to distort the relative elevation variable.

## ACKNOWLEDGEMENTS

Routines for the GM estimator within the R (R Development Core Team, 2006) statistical language and environment were provided by Luc Anselin and Julia Koschinsky, University of Illinois. Total station elevation measurements were provided by the cooperation of Dr. Buck Tillotson, Professor of Agriculture at Western Illinois University, Jay Peters, former student at Western Illinois University, and Sean Evans, former Crops Systems Educator with University of Illinois Extension. RTK-GPS elevation data were provided by Steve Hobson,

Area Engineer with USDA-NRCS in Macomb, Illinois. Electrical conductivity measurements were provided by David Brummer of KSI in Shelbyville, Illinois.

Special appreciate is given to USDA-SARE for providing funding to facilitate the research of Griffin's PhD dissertation.

## REFERENCES

- Anselin, L. 2003. GeoDa 0.9 User's Guide. Spatial Analysis Laboratory, University of Illinois, Urbana-Champaign, IL.  
[http://sal.agecon.uiuc.edu/geoda\\_main.php](http://sal.agecon.uiuc.edu/geoda_main.php)
- Anselin, L. 2001. Spatial Effects in Econometric Practice in Environmental and Resource Economics. *American Journal of Agricultural Economics* 83(3) August 2001: 705-710.
- Anselin, L. 2002. Under the Hood. Issues in the Specification and Interpretation of Spatial Regression Models. Available on-line at:  
<http://sal.uiuc.edu/users/anselin/papers/hood.pdf> Accessed March 1, 2006.
- Anselin, L. 1988. *Spatial Econometrics: Methods and Models*, Kluwer Academic Publishers, Dordrecht, Netherlands.
- Bell, K.P. and Bockstael, N.E. 2000. Applying the Generalized-Moment Estimation Approach to Spatial Problems Involving Micro level Data. *The Review of Economics and Statistics*, February 2000, 82 (1): 72-82.
- Bishop, T. F. A. and A. B. Mcbratney. 2002. Creating Field Extent Digital Elevation Models for Precision Agriculture. *Precision Agriculture*, 3, 37-46, 2002
- Clark, R. L. and R. Lee, 1998. Development of Topographic Maps for Precision Farming with Kinematic GPS. *Transactions of the ASAE VOL. 41(4)*: 909-916.
- Cressie, Noel A.C. 1993. *Statistics for Spatial Data*. John Wiley & Sons: New York.
- Florax, R. and Folmer, H. 1992. Specification and Estimation of Spatial Linear Regression Models: Monte Carlo Evaluation of Pre-Test Estimators. *Regional Science and Urban Economics*, 22:405-432.
- Greene, W. H. 1993. *Econometric Analysis*, Macmillan, New York.
- Griffin, T.W., Fitzgerald, G, Lambert, D.M, Lowenberg-DeBoer, J., Barnes, E.M., and Roth, R. 2005a. "Testing Appropriate On-Farm Trial Designs and Statistical Methods for Cotton Precision Farming," *Proceedings of the Beltwide Cotton Conference*, January 4 - 7, 2005, New Orleans, LA. Available at:  
<http://www.cotton.org/beltwide>
- Griffin, T.W., Brown, J.P., and Lowenberg-DeBoer, J. 2005b. Yield Monitor Data Analysis: Data Acquisition, Management, and Analysis Protocol. Available on-line at: <http://www.purdue.edu/ssmc>

- Griffin, T.W., Florax, R.J.G.M., Lowenberg-DeBoer, J. 2006. Field-Scale Experimental Designs and Spatial Econometric Methods for Precision Farming: Strip-Trial Designs for Rice Production Decision Making, Southern Agricultural Economics Association Annual Meetings Selected Paper Orlando, Florida, February 5-8, 2006. Available on-line at: <http://agecon.lib.umn.edu/>
- Kelejian, H.H. and Prucha, I. 2006 "Specification and Estimation of Spatial Autoregressive Models with Autoregressive and Heteroskedastic Disturbances," Department of Economics, University of Maryland. Working Paper. Available on-line at: [http://www.econ.umd.edu/~prucha/Papers/WP\\_Prucha\\_3\\_2006.pdf](http://www.econ.umd.edu/~prucha/Papers/WP_Prucha_3_2006.pdf)
- Kelejian, H.H. and Prucha, I.R. 1999. A Generalized Moments Estimator for the Autoregressive Parameter in a Spatial Model. *International Economic Review*, 40, 509-533. Available on-line at: [http://www.econ.umd.edu/~prucha/Papers/IER40\(1999\).pdf](http://www.econ.umd.edu/~prucha/Papers/IER40(1999).pdf)
- R Development Core Team. 2006. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.