1-1-2015

Spatial Pattern Analysis of Agricultural Soil Properties using GIS

Corrin Jared McCarn

Spatial pattern analysis of agricultural soil properties using GIS

By

Corrin J. McCarn

A Thesis
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Master of Science
in Geosciences
in the Department of Geosciences

Mississippi State, Mississippi

December 2015

Spatial pattern analysis of agricultural soil properties using GIS

By

Corrin J. McCarn

Approved:

Qingmin Meng
(Major Professor)

John C. Rodgers III
(Committee Member)

William H. Cooke III
(Committee Member)

Michael E. Brown
(Graduate Coordinator)

R. Gregory Dunaway
Dean
College of Arts & Sciences

Name: Corrin J. McCarn

Date of Degree: December 11, 2015

Institution: Mississippi State University

Major Field: Geosciences

Major Professor: Qingmin Meng

Title of Study: Spatial pattern analysis of agricultural soil properties using GIS

Pages in Study: 48

Candidate for Degree of Master of Science

Agricultural soil properties exhibit variation over field plot scales that can ultimately effect the yield. This study performs multiple spatial pattern analyses in order to design spatially dependent regression models to better understand the interaction between these soil properties. The Cation Exchange Capacity (CEC) and Calcium-Magnesium Ratio (CaMgR) are analyzed with respect to Calcium, Magnesium, and soil moisture values. The CEC and CaMgR are then used to determine impact on the yield values present for the field. Results of this study show a significant measure of model parsimony (0.979) for the Geographically Weighted Regression (GWR) model of the CEC with free Ca, Mg, and soil moisture as explanatory variables. The model for CaMgR using the same explanatory variables has a much lower measure of model fit. The yield model using the CEC and CaMgR as explanatory variables is also low, which is representative of the underlying processes also impacting yield.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

The global population is slated to increase drastically by the year 2050 requiring agricultural yield and crop production to meet the increase in food demand resulting from future growth (Dawson 2014). While crop production is still increasing, the percentage of crop yield increase is decaying at a rate that will diminish over time into static or reduced production. In order to address potential crop optimization strategies, current research has developed crop modeling parameters (Mendelsohn 2007, Drewniak 2012), such as the Community Land Model and multiple crop failure scenarios. Many of these models address different aspects of crop production, but complex models are needed to explain complex processes. As a quickly evolving scientific field, Geospatial Information Systems (GIS) are well-suited for the spatial and statistical analyses that are necessary to understand the effects of multiple variables on crop yield. More simplistic crop production models address singular variables such as temperature, or atmospheric impact. These variables modeled individually produce skewed results that do not account for the presence of multiple causative factors. This study will attempt to use GIS to model regression equations including multiple variables that act upon crop production.

1.1 Importance of Study

Increasing global population and the future acceleration of that increase necessitate advancements in food production. Currently research has been focused on developing chemical applications and increasingly complex crop simulation models. Plant growth is a complex process and simulating entire field plots worth of growth requires consideration of a large number of variables. Complex processes occur within agricultural soils to a degree that including all of the necessary factors in a single model is extremely difficult. The large number of variables necessary for crop modeling can result in reduced ability to account for minute changes within those variables and lowers overall model precision. Focusing on specific target variables for these models can maintain precision while still allowing for the inclusion of multiple variables in the analysis. The target variables for this study will be the Cation Exchange Capacity (CEC), and the Calcium - Magnesium Ratio (CaMgR), soil moisture content, and elevation. Other variables included for this study will be the elemental base percentages comprising the CEC (Hydrogen, Potassium, Calcium, and Magnesium), and individual measures of Calcium and Magnesium. Focusing primarily on the soil characteristics will develop values for these characteristics that can in turn be used within more complex models in place of less precise existing soil values. Alternatively, if this modeling is implemented at a field level in a precision agricultural design, the optimization of soil characteristics could be used to inform management decisions.

1.2 Study Purpose

The primary purpose of this study is to model spatial relationships of CEC and CaMgR in order to better understand potential agricultural soil optimization. Projected global population increases necessitate agricultural advances in order to sustain food sources. The CEC will be analyzed using regression and interpolation methods in order to determine elemental composition of the CEC values, as well as the impacts of soil moisture and elevation. Calcium, Magnesium, soil moisture, and elevation will be used to develop understanding of the CaMgR and model these values for the target field. The CEC and CaMgR values will then be analyzed using multivariate statistics in order to determine the relationship between the two. That relationship will then be measured in terms of impact on crop yield within the target field. This study has been designed to address the following questions:

1. Can the relationship between CEC and specific independent variables be defined?
2. Could spatial regression techniques be used to define how Ca, Mg, and additional explanatory variables act upon the CaMgR?
3. How do the values for CEC and CaMgR interact on an individual field basis and how does this interaction impact yield?

Answering these initial questions will allow for modification of existing field methods in order to maximize potential yield. Gaining a better understanding of soil processes and the relationships between soil characteristics would necessitate changes in agricultural management decisions. Soil chemical applications and soil treatment methods could be more effectively used to develop optimized ratios between the CEC and free Ca and Mg. There is potential for the soil moisture and elevation to alter the effects of the CEC and CaMgR on the yield which must also be addressed in this study. The largest potential impact of the soil moisture and elevation at the field level would be potential for water to leach nutrients from the soil. If these relationships can be defined, they could benefit knowledge of interacting soil processes, which could then in turn be used to provide better management of agricultural field plots.

As a global purpose for this study, future research must address the growing agricultural needs of an ever expanding population. The goal of providing adequate agricultural crop yields to meet increasing demand necessitates improving current crop growth techniques. One way of accomplishing this is by optimizing known soil properties that affect the nutrients available to crop plants. The maintenance of nutrients in the soil and defining an optimal ratio for these nutrients could potentially boost yield, or provide further knowledge of how they might be detrimental to yield if not optimal. A localized goal for this study is to increase understanding of soil properties in light of precision agricultural practices. Chemically analyzed soil samples provide Cation Exchange Capacity and Calcium-Magnesium Ratio values for each sampled feature, but these values are difficult to interpret. Optimal ranges for such values exist, but the true impact on yield of different values can vary across the spatial range of a single field plot. Identifying how these processes interact with each other as well as yield could inform management and administrative decisions for precision agriculture at a field level.

CHAPTER II

LITERATURE REVIEW

2.1 Soil Moisture

Soil moisture, CEC, CaMgR, and topographical characteristics all vary within individual field plots (Anderson-Cook 2002). Soil moisture plays an important role in plant growth (Volkmar 1997) with many plant processes depending on water for basic metabolic functions and development. Preexisting water stress can significantly reduce the ability of the plant to respond to additional stress factors which can ultimately result in greater vulnerability. Green (2004) used a statistical crop model with weighting to identify the topographical moisture index as accounting for 38-48% of variance for winter wheat yield. This amount of variance explaining nearly half of the crop production data dictates that one of the main variables in geospatial regression analysis should be the soil moisture content. Application of water to the field plot can also effect the soil moisture content (Marques Da Silva 2008) which also functions as an explanatory factor in spatial variance on crop yield. Climate change could potentially be detrimental to crop growth for a range of reasons (Hu 2003), and the addition of more soil moisture can result in transportation or leeching of soil minerals.

2.2 Cation Exchange Capacity

Ionic charge of the soil and the quantified variable of CEC within the soil samples is another variable that influences crop yield. In precision agriculture, the CEC is often used to measure overall soil salinity as well as other physic-chemical properties (Corwin 2005). Soil type can be defined through the soil electrical charge (Kühn 2008) and in turn this can be useful when establishing the amount of spatial autocorrelation present within soil sample data. As with the soil mineral content, the CEC is highly variable and differentiation is present at a field plot basis. This differentiation is observable to a degree in recorded soil samples which can be useful for discerning miniscule changes in soil types within the same field plot. Officer (2004) used fields in Illinois and Missouri within a principle component analysis to correct soil maps. The principle component analysis resulted in establishing concave elevation characteristics associated with soil electrical charge that had a large impact on soil fertility. Roughly the lower areas in the fields were collecting greater ionic charge due to the leeching of nutrients in the soil and the transportation of soil nutrients by water. This leeching of soil nutrients can be attributed to the CEC as a measure of the soil's ability to maintain nutrients. CEC is currently viewed with a generalized optimal range; low CEC results in the soil being unable to maintain nutrients long enough for adequate plant absorption, and too high of a CEC attributes to nutrients being maintained too efficiently for the plants to be able to uptake.

2.3 Calcium Magnesium Ratio

Soil mineral content such as the CaMg ratio has been shown to directly affect plant growth and development. The mineral content of the soil is highly variable and is known to vary within the same field plot as well as within different soil types. Soil maps are available for many areas such as the location of this study, but there are often large amounts of error within soil maps. This can be alleviated by using soil prediction algorithms (Moore 1993) which primarily identify the soil but also to establish the validity of the soil samples taken on site representing ground reference data. By modeling crop yield and soil mineral content (Vrindts 2003, Villamil 2012) the correlation coefficient of the variable can be used to establish management plans for field plots. There have been multiple long-term field studies (Olness 2001, 2002) which have addressed maize and soybean mineral sensitivities as measured by genetic variations within species. Nitrogen and potassium have both been established as influential growth minerals for plants, but the relationship between crop yield and CaMg ratio are more difficult to define.

2.4 Elevation

Errors identified in DEMs (Aziz 2008, Holmes 2000) necessitate correctional algorithms in order to use those models for agricultural research. Even when using 10-m DEMs (Green 2007) the amount of associated error is too large for agricultural terrain modeling. Field plots represent a difficult surface to model due to the precision and accuracy necessary to create functioning topographical models. Hydrological function models are directly dependent upon correct fit and representation of the elevation model when formulated for field plots. Kravchenko (2000) used a weighted model to identify physical factors such as slope and elevation, and the degree that they effect crop production. This study established topography as accounting for 20% of the statistical variance for crop production in the target fields used which indicated that a large amount of data variation could be attributable to the errors associated with DEMs. Roughly half of the variability of soil properties can be attributed to elevation, slope, and moisture changes within the soil over the length of a field plot (Moore 1993). In precision agriculture, often Real-Time Kinetic (RTK) technology and sensors are fitted to the harvesting equipment which allows collection of ground reference data. Soil topography measurements vary depending on data collection methods (Schmidt 2003) with RTK being more suitable for agricultural field use. By using the more precise RTK elevation data collected during crop harvesting periods, the error associated with DEMs and agricultural implementation can be circumvented. Each harvested cell of the field, which encompasses a space roughly the size of the harvester, is assigned an elevation value as measured by the sensor attached to the harvester. The sensor attached to the harvester also collects raw harvest data and base soil moisture readings for each of the cells.

2.5 Spatial Analysis Methods

Spatial analysis of soil properties has been accomplished using a multitude of different methods. Semivariogram analysis is one of the most common and has revealed normal trends and variabilities within the soil properties of a regionalized area such as a field plot (Trangmar 1985). These trends within the data can be measured using the results of semivariograms and the extent of spatial dependence can be established. These spatial dependencies are commonly also used to show dissimilarity in terms of distance for sample points (Goovaerts 1998). Interpolation techniques such as kriging are typically used in conjunction with variogram analyses in order to estimate values for areas not sampled and address potential directionality. No interpolation methods will be used for this study meaning spatial pattern analysis using various hot spot and clustering analyses will be used.

Ordinary-Least Squares (OLS) regression will be used for the initial exploratory regression and variable fitting of the model. The main difference between the OLS regression and the GWR methods are in terms of spatial distribution and dependence.

Spatial heterogeneity of the variables must be corrected using algorithms and weighted matrices (Charlton 2009). The required corrections for OLS regressions using spatial data necessitate using a different regression technique. The GWR method is designed to incorporate the spatial distances between the points. Spatial changes from nearest neighbor points are used when estimating the dependent variable values. These distances influence the estimations using varying amounts of spatial autocorrelation.

CHAPTER III

METHODS

3.1 Study Area

The location of the study area is within the boundary of Lake County, Tennessee. The primary field plot used for this study will be referred to as ShopSuper, which encompasses 293.6 total acres of arable land. Corn, soybeans, and wheat are the primary rotational field crops planted on this field plot with annual soil samples taken for chemical analysis. Out of the potential field plots that could be used as the basis for this study, this plot was the largest in terms of acreage and had recent crop yield data.

3.1.1 Data Collection

The crop yield data used for this study were collected using RTK precision agricultural practices. These shapefiles were continuously uploaded to the Greenstar 2 controller module used for this process and were then formatted after for the purpose of this study. Each of the shapefiles were projected using the Universal Transverse Mercator (UTM) Zone 16 North projection. These shapefiles encompass the entire field and are comprised of yield cells which are roughly the width of the harvester being used. The RTK process of crop harvesting resulted in yield cell shape files that have an attributed soil moisture content for each cell. These soil moisture values are collected using in situ measurements recorded during the harvesting process by the sensor mounted on the harvester. The elevation of each of the yield cells is averaged concurrently during this process which results in an elevation value assigned to each of the cells. The RTK elevation values produced during harvesting were used for this study as opposed to DEM layers due to the accuracy associated with the RTK values. Due to the high number of the sample points and the constant GPS receiver correction, the elevation error for the entire field would be negligible (Schmidt 2003). Using the RTK data to produce continuous elevation values for the target field allowed for inclusion as a variable being used for this study. The large number of cells present within the yield shapefiles warranted clipping by the soil sample points file. This clipping process resulted in a joined file with both yield values and soil sample information for each of the samples. While this is not a full representation of the fluid yield values over the entire field plot, it was an effective way to make sure soil properties were not assumed for larger polygon sections.

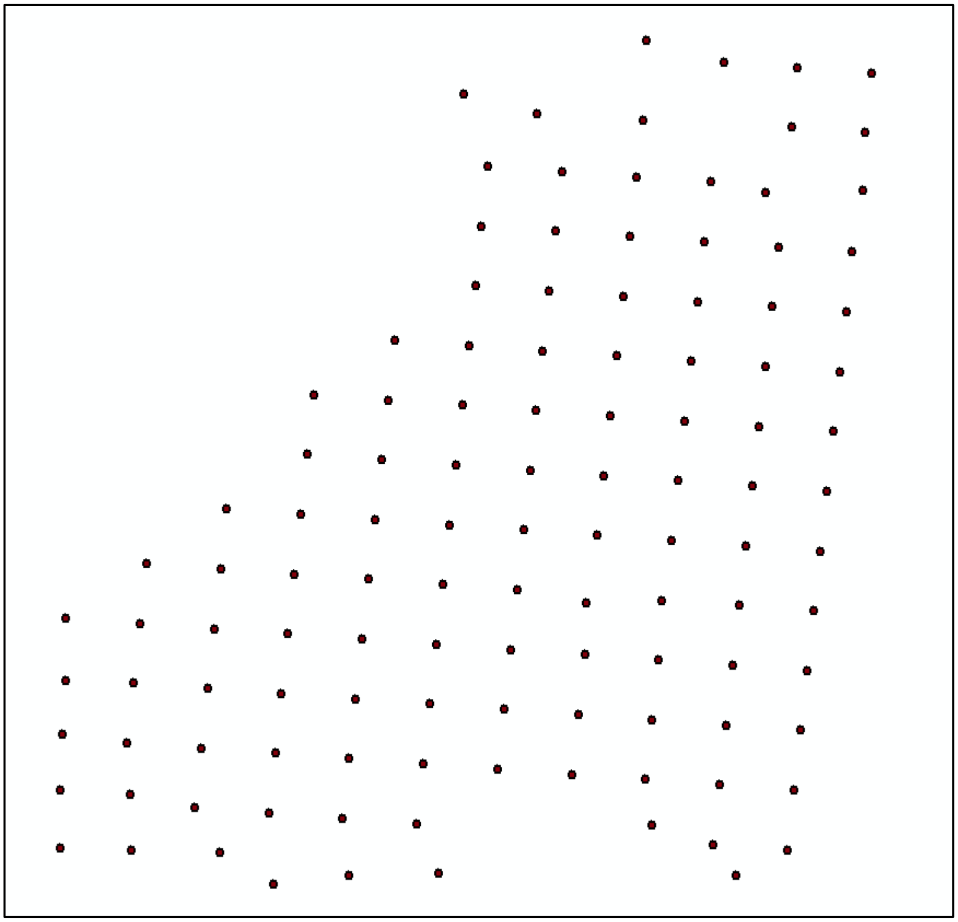

Figure 3.1 Soil Sample Point Layer of Target Field

3.1.2 Soil Composition

The soil samples for this study were collected in October of 2014. There are 113 soil sample points for the soil sample shapefile which was chemically analyzed by a private agricultural laboratory. The attributes of the soil samples include the CEC, CEC base percentages (Hydrogen, Potassium, Calcium, and Magnesium), CaMgR, Calcium level, Magnesium level, and several other mineral contents. Latitude and longitude of each of the samples were recorded during the collection process and were later projected to Universal Transverse Mercator (UTM) Zone 16 North. The soil map for this field plot

(NRCS 2014) establishes numerous soil types present within the study area. The majority of the soils are silt-base, loam, or clay soils in composition. These types of soils have been shown to correlate with CEC and in some cases improved crop production (Sudduth 2005).

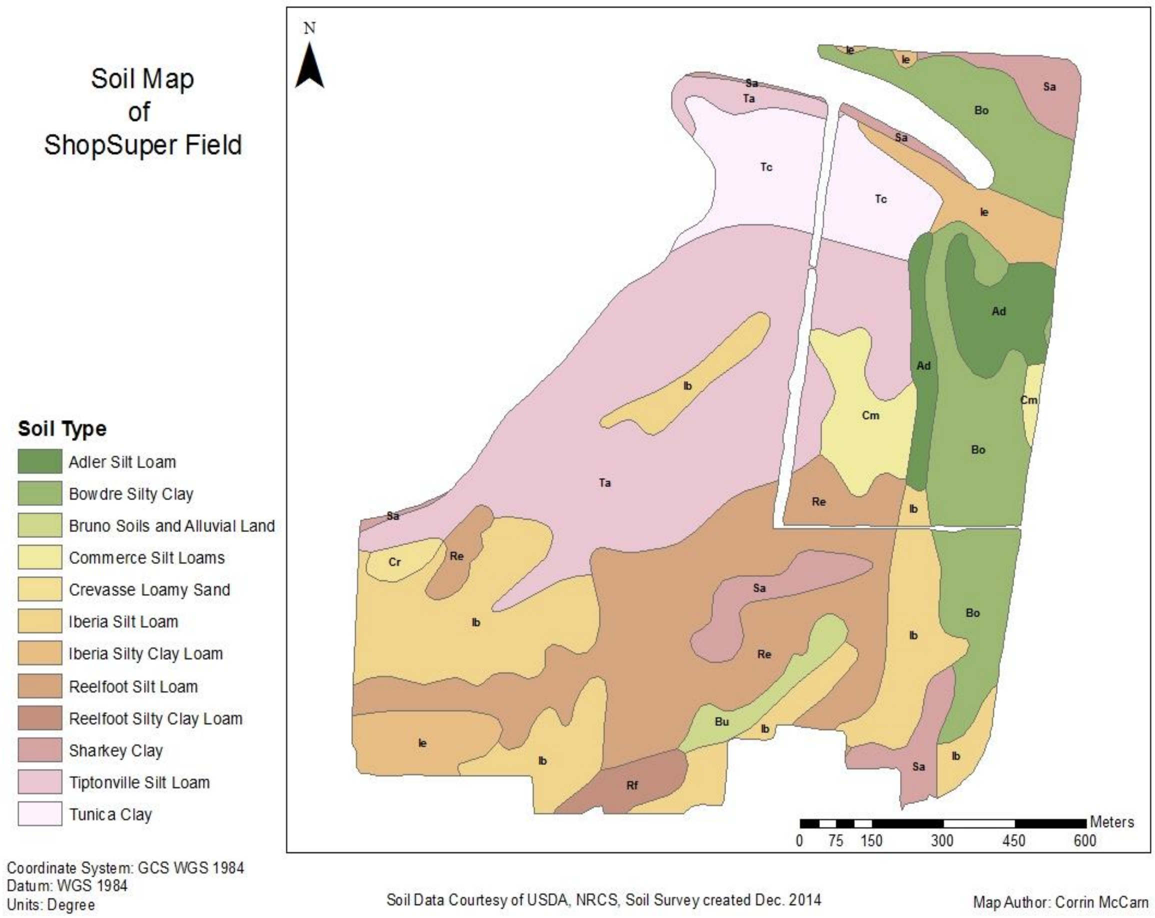

Figure 3.2 Soil Identification Map of Target Field

3.2 Point Pattern and Clustering

3.2.1 Multi-Distance Spatial Cluster Analysis

The initial analysis will be to determine the orientation of the variables in relationship to one another over the area of the entire field. By using a Multi-Distance Spatial Cluster Analysis (Ripley's K) function the clustering or dispersion of the variables can be established. The formula is able to be weighted proportional to the distance between point features:

$$L(d) = \sqrt{\frac{A\sum_{i=1}^{N}\sum_{j=1, j\neq i}^{N} k(i,j)}{\pi N(N-1)}} \tag{3.1}$$

The formula is designed to compare the spatial point values with a complete spatial randomness model using an index of dispersion (Pfeiffer 1996). This is to define whether the spatial orientation of the points impacts the values and to what degree the points display spatial autocorrelation. Expected point values are created using this formula and then compared to the actual values to show the level of clustering or dispersion present within the data. Since these soil sample points are fixed with predefined distances, this is measuring the extent of the dispersion for the variable values present for the points.

3.2.2 Spatial Autocorrelation

The variables chosen for this study will be analyzed individually using both the spatial location and the existing values using a spatial autocorrelation method (Global Moran's I). The principle design of this inferential statistic method measures the result of the analysis in terms of the null hypothesis. The null hypothesis states that the variable values are present randomly within the study area, with the probability score and Z-score being used to determine whether to accept or decline the null hypothesis. If the analysis shows that the probability score is not significant, the null hypothesis must be accepted. A significant probability score can infer that the spatial orientation of the data is clustered if the Z-score is positive, and if the Z-score is negative the data is spatially dispersed. This form of analysis has multiple distance options to ensure that the distances between neighboring points is uniform and does not influence the results. For the data used in this study, the point distances are uniform so there is no need to determine a distance modifier. Unlike the Ripley's K analysis, this analysis will have to be performed individually for each of the variables being studied. The spatial autocorrelation of the variables will aid in designing regression models by identifying patterns of dispersion within the data that could be too strongly related and warrant multicollinearity.

3.2.3 Cluster and Outlier Analysis

The cluster and outlier analysis (Anselin Local Moran's I) will result in the creation of a new layer to be used within the GIS. Continuing the idea of spatial clustering introduced during the Global Moran's I analysis, this layer will show where clusters are spatially located within the data and whether the clustered values are low or high. This function uses a standard statistical confidence level of 0.05 for determining whether values are clustered high, clustered low, or surrounded by values that indicate an outlier. This high and low clustered value identification will be especially useful when determining the spatial orientation and relationships of the soil variables.

3.2.4 Hot Spot Analysis

Each of the primary target variables was processed using the Getis-Ord Gi* technique of hot spot analysis. Hot spot analysis will be used to measure each of the variables to identify where there are large quantities of high values and large quantities of low values. This will allow each of the variables to be compared to find correlations in terms of the hot spot groupings and the low value groupings. The point layers used for input will need to use a fixed distance band with Euclidean distance between the points to identify these hot spots. This hot spot identification will also be used to verify the clustering results of the Anselin Local Moran's I toolset. Determining the clusters of the variable values within the spatial boundary of the field will be useful when paired with the results of the exploratory regression for deciding which variables to use for the finalized regression models.

3.2.5 Grouping Analysis

The hot spots determined using the Getis-Ord Gi* method above will then be grouped using the grouping analysis toolset. The first group of correlated hot spots will be physical variables (moisture, elevation) compared to the yield. The second grouping will be of the chemical properties of the soil (CEC, CaMgR) compared to the yield. A final grouping will then be created comparing all of the variables to the yield and all three will be mapped in order to identify trends within the groups. Each of these groups will need to use Delaunay triangulation in order to be spatially grouped, and can then be analyzed to determine the R-squared value of group parsimony. The measure of fit for the gorupings can then be used to decide which variables exhibit like values and are potentially correlated for use in the regression model design process.

3.3 Regression Techniques

3.3.1 Exploratory Regression

This study has numerous variables with complex relationships which could make the regression process more difficult to design. The CEC values of the field overall are comprised of individual base percentages of Hydrogen, Potassium, Calcium, and Magnesium. As these base percentages are being weighed as variables and spatially modeled using the techniques previously listed, they are still necessary for this analysis, but exploratory regression is needed to determine proper usage. If the CEC is used as the dependent variable and those base percentages are set as explanatory variables the regression design will be fatally flawed due to multicollinearity.

The exploratory regression process within the GIS is scripted to evaluate every possible combination of variables and designate which combinations meet the criteria assigned by the design of the study. Each of the candidate explanatory variables is assessed using specified thresholds for model fit and each of the combinations return a score in terms of those criteria. While this is a useful tool in designing the final regression model, there is potential for this approach to influence the variable selection process to include only variables that contribute to a successful model. Bearing this problem in mind and understanding the nutrient processes of the soil characteristics will minimize the potential framing effects of this tool. The regression design of this study will use this tool solely to reduce unnecessary multicollinearity within the final regression model.

3.3.2 Geographically Weighted Regression

The Geographically Weighted Regression (GWR) method will be used as the primary regression modeling process for this study. Using the results of the pattern and cluster analyses listed above (Ripley's K, Global Moran's I, and Anselin Local Moran's I) the spatial distribution of the target variables will be used to determine the input for this regression. The exploratory regression technique using primarily OLS will be used to reduce the overall multicollinearity present within the model.

Typical linear regression utilizes paired values for independent variables in order to estimate the corresponding value of the dependent variable. With the addition of spatial orientation for botht eh dependent and independent variables, the spatial locations are added to this regression function. The formula for GWR is:

$$y_i(\mathbf{u}) = \beta_{0i}(\mathbf{u}) + \beta_{1i}(\mathbf{u})x_{1i} + \beta_{2i}(\mathbf{u})x_{2i} + \ldots + \beta_{mi}(\mathbf{u})x_{mi} \tag{3.2}$$

In this formula, 'y' would be the dependent variable with the subsequent 'x' variables being the independent variables and 'u' serving as spatial locations (Charlton 2009). This form of regression is based on dependent variable values being estimated at each of the spatial locations in terms of the relationship with the independent variables and the respective values at those points. There are multiple weighting methods that can be used for this function, but because the soil sample layers are so uniformly spaced, this study will use a fixed kernel with the bandwidth established using the Akaike Information Criterion method.

CHAPTER IV

RESULTS

4.1 Spatial Distribution

The Ripley's K function used to measure the multi-distance spatial orientation of the study variables shows that the data is more dispersed than clustered. All of the variables were included for this K function in order to determine the spatial orientation for the data as a whole. Each of the feature values are assigned an expected K value representing a random distribution, which are then compared to the observed K values. Due to largely negative differential between expected and observed K values, the trend of the data is dispersed instead of clustered.

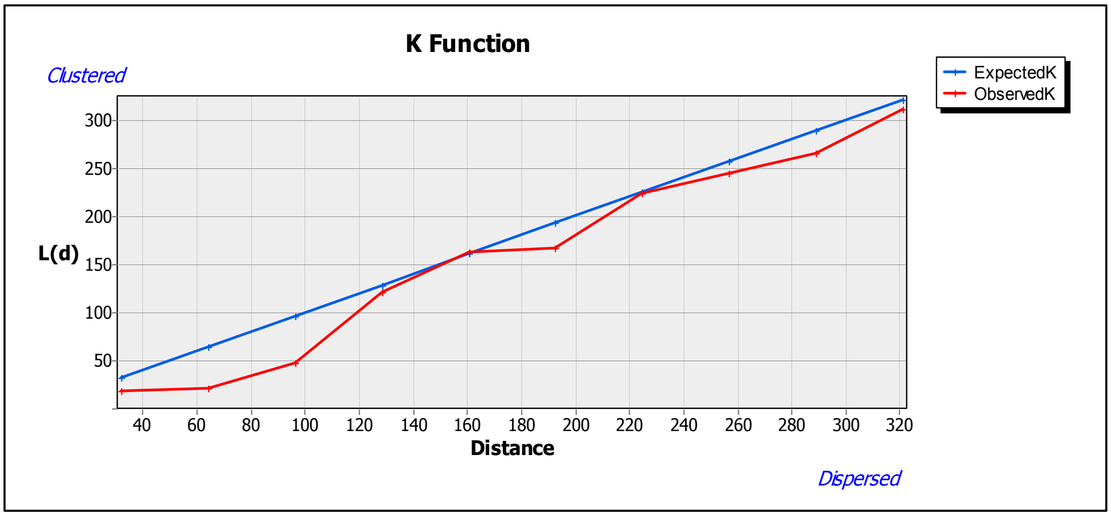

Figure 4.1 Ripley's K Function Results for All Variables

The trend of dispersion is observable for both CEC and yield using spatial autocorrelation methods. The Global Moran's I function provides Z-Scores for both CEC and yield that are not significantly different than randomly dispersed. The soil moisture variable orientation was significant within a 0.05 confidence interval due to the clustering of the variables. The elevation and CaMgR were significant to the extent that the clustered pattern of both variables had less than a 1% chance of being randomly clustered. The variable clustering of the CaMgR would mean that within the study area, the CaMgR values display a degree of spatial autocorrelation. The elevation variable is more so expected because of the inherent flatness of field plots, which normally have a gradual slope over the course of the entire plot.

Table 4.1 Global Moran's I Values

	CEC	CaMgR	Moisture	Elevation	Yield
Moran's Index:	0.264587	0.640665	0.447368	0.626295	0.067119
Expected Index:	-0.009091	-0.009091	-0.009091	-0.009091	-0.009091
Variance:	0.049440	0.049248	0.049273	0.048527	0.048729
z-score:	1.230835	2.927896	2.056356	2.884334	0.345234
p-value:	0.218385	0.003413	0.039748	0.003922	0.729919

Local cluster and outlier analysis was performed for the soil properties of CEC and CaMgR. The Global Moran's I results show that the CaMgR is clustered as opposed to dispersed and that CEC is randomly distributed for the study plot. Using a more localized for of cluster analysis the CEC has observable clustering of both low values and high values in the plot. This clustering of values is likely due in large part to the different soil types present within the field plot, soil moisture properties, and the gradual sloping elevation of the surface.

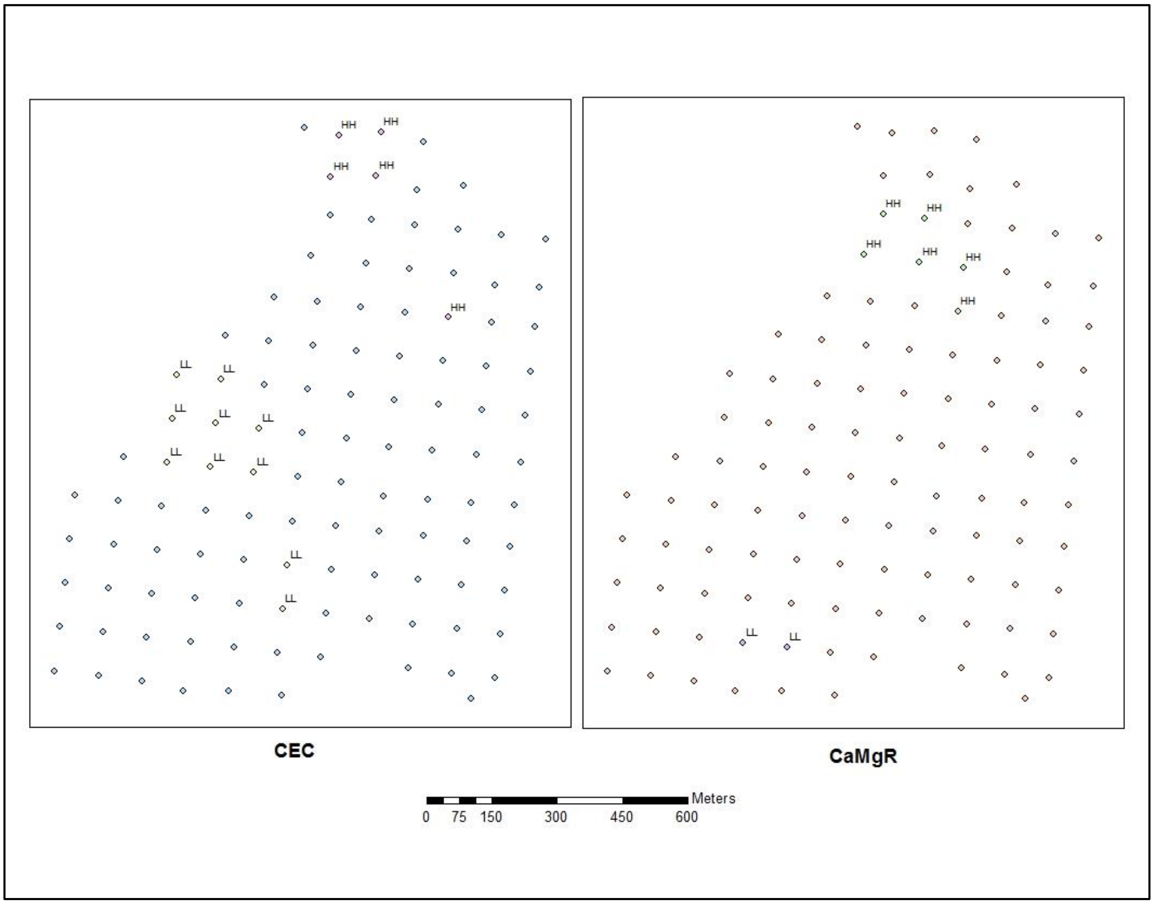

Figure 4.2 Anselin Local Moran's I Results for Cation Exchange Capacity and Calcium-Magnesium Ratio

The clustering present within the local Moran's I results addressed in greater detail using the Getis-Ord Gi* function. The results of this function creates a simple Thiessen polygon created for each soil sample point with an assigned Z-score. Each of the Z-scores correlate to clustered values, with negative scores associated to clustered low values and high scores assigned to clustered high values. The class breaks of each variable clustering layer have been modified so that the highest quantity of positive groupings are representative of Z-scores that range above 1.97. Likewise, the negative Z-scores expressed through class breaks are representative of scores lower than -1.97. The values of both 1.97 and -1.97 are used to denote a confidence level greater than 0.95 for the clustering of these values.

Each of the variable clustering results are mapped and can be compared to determine which hot spots could be the result of correlation between variables. Mapping the clusters can also aid in initial visual determination of spatial dependency which aids in identifying areas of interest for use in the regression modeling. The hot spot clustering of the yield variable produces no positive groupings of high values above a 0.95 confidence level and only one polygon located in the southeast corner that exhibits a clustering of low values above a 0.96 confidence level.

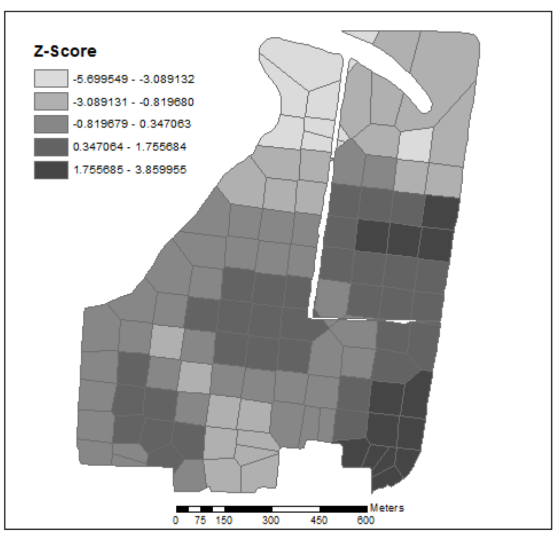

Figure 4.3 Getis-Ord Gi* Clustering Results for Elevation Variable

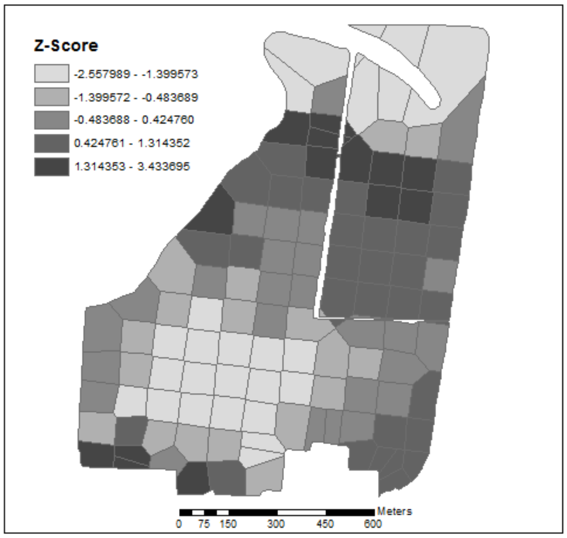

Figure 4.4 Getis-Ord Gi* Clustering Results for Calcium-Magnesium Ratio Variable

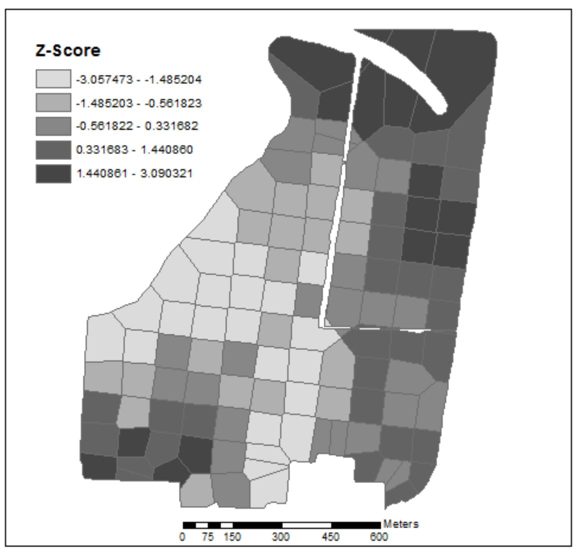

Figure 4.5 Getis Ord Gi* Clustering Result for Cation Exchange Capacity

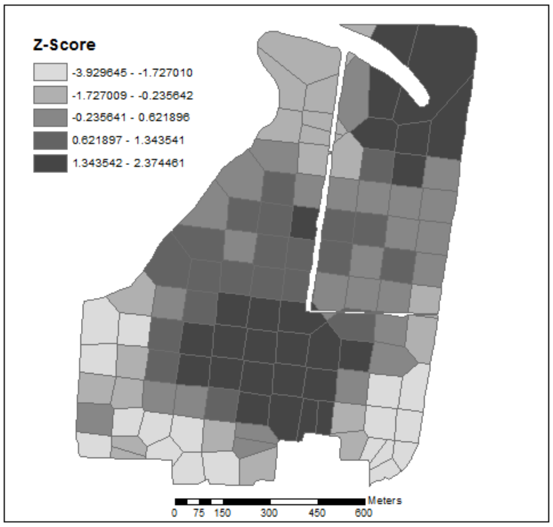

Figure 4.6 Getis-Ord Gi* Clustering Result for Soil Moisture Variable

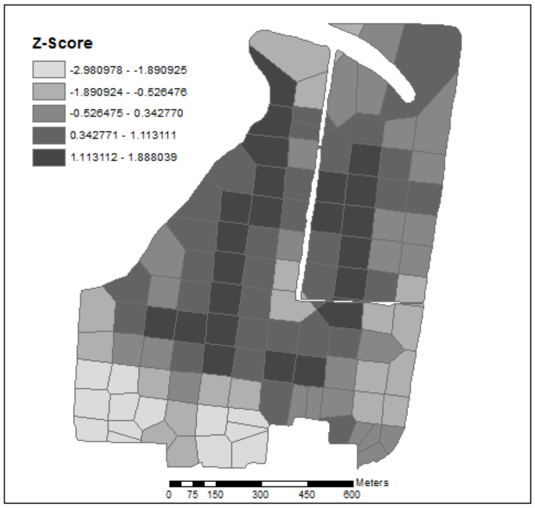

Figure 4.7 Getis-Ord Gi* Clustering Result for Yield Variable

The results of the hotspot analysis were then compared to the results of the grouping analysis performed. The first grouping of variables includes soil moisture, elevation, and the yield which make up the physical properties of the soil. Soil moisture in conjunction with elevation can be used to identify potential leaching of soil nutrients due to transportation by moisture present within the soil. The second grouping of variables includes the CaMgR, CEC, and yield which is representative of the chemical properties of the soil samples. The final grouping is an overall grouping containing the variables from both the physical and chemical groupings.

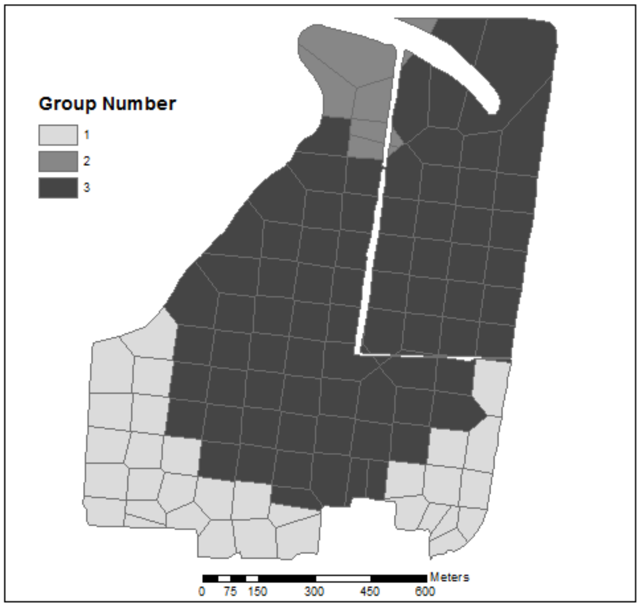

Figure 4.8 Grouping Analysis Result for Physical Properties

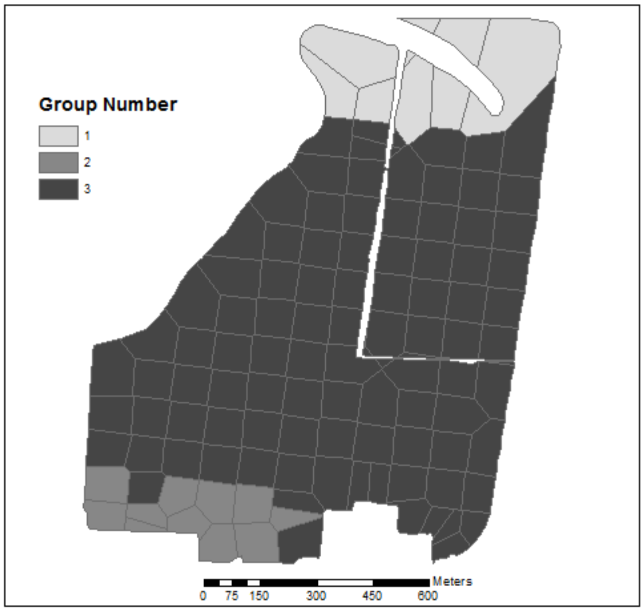

Figure 4.9 Grouping Analysis Result for Chemical Properties

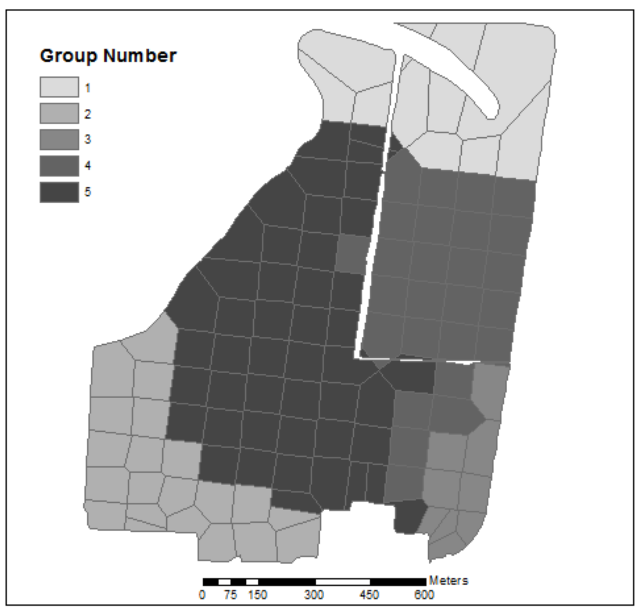

Figure 4.10 Grouping Analysis Result for Overall Soil Properties

The physical and chemical groupings both used three classes to identify groups where all of the variable values were the most similar while maintaining the largest amount of difference from the additional groups. The overall soil properties grouping was created with the same method but five classes were used to create a more meaningful set of groups. The physical and chemical groupings both display a primary group which encompasses the majority of the target area, and additional smaller groups at the top and bottom of the target area. This same trend is observable for the overall soil properties grouping, with the exception of an additional group present in the central area. These groupings within the data link hot spots and clusters for individual variables with additional similar variable values to show which spatial areas within the target location are related. This does not necessarily identify relationships between specific variables, but instead determines the spatial location of similar variable values.

4.2 Regression Modeling

The GIS exploratory regression toolset compares multiple models using assigned dependent and independent variables. Each model was scored for the purposes of reducing multicollinearity and improving performance of the chosen regression model techniques. By allowing the models to be scored, potential models with overlapping variables could be paired out of the design for the finalized model. As mentioned earlier in this study, there have been scientific objections to using exploratory regression in lieu of initial hypotheses and the idea of potentially framing the study. Many of these variables are dependent upon multiple factors and the risk of fatal model design due to multicollinearity is accordingly high. Another potential problem with OLS regression compared to the GWR method is the idea of spatial dependency and heterogeneity within the data. The OLS regression would have to be modified in order to deal with the differential values of the variables over spatial distances.

Table 4.2 Cation Exchange Capacity Independent Variable Summary of Variable Significance

Summary of Variable Significance			
Variable	Significant	Negative	Positive
Ca	100.00	0.00	100.00
Mg	100.00	0.00	100.00
Moist	62.50	100.00	0.00
CaMgR	50.00	25.00	75.00
Elev	37.50	43.75	56.25

Table 4.3 Cation Exchange Capacity Independent Variable Summary of Multicollinearity

Summary of Multicollinearity			
Variable	VIF	Violations	Covariates
Elev	1.13	0	-
Moist	1.13	0	-
Ca	5.87	0	-
CaMgR	1.35	0	-
Mg	5.99	0	-

Table 4.4 Calcium-Magnesium Ratio Independent Variable Summary of Variable Significance

Summary of Variable Significance			
Variable	Significant	Negative	Positive
Moist	81.25	100.00	0.00
Mg	75.00	100.00	0.00
Ca	56.25	0.00	100.00
CEC	50.00	25.00	75.00
Elev	0.00	0.00	100.00

Table 4.5 Calcium-Magnesium Ratio Independent Variable Summary of Multicollinearity

Summary of Multicollinearity			
Variable	VIF	Violations	Covariates
Elev	1.16	0	-
Moist	1.21	0	-
Ca	26.70	8	CEC (88.89)
CEC	45.02	10	Ca (88.89)
Mg	9.46	6	CEC (55.56)

Table 4.6 Yield Independent Variable Summary of Variable Significance

Summary of Variable Significance			
Variable	Significant	Negative	Positive
Moist	100.00	0.00	100.00
CEC	87.10	100.00	0.00
Ca	48.39	38.71	61.29
Mg	12.90	51.61	48.39
Elev	0.00	35.48	64.52
CaMgR	0.00	51.61	48.39

Table 4.7 Yield Independent Variable Summary of Multicollinearity

Summary of Multicollinearity			
Variable	VIF	Violations	Covariates
Elev	1.16	0	-
Moist	1.24	0	-
Ca	27.38	15	CEC (93.75)
CaMgR	1.35	0	-
CEC	45.02	21	Ca (93.75)
Mg	10.02	13	CEC (75.00)

The tables produced through using the exploratory regression present variables that have a large amount of multicollinearity. For the CEC GWR model, the variables of free Ca, free Mg, and soil moisture have the highest significance resulting from the multiple OLS regression attempts that are performed by this function. There are also no multicollinear variables which would present redundant information. The CaMgR GWR model exhibits a lower significance but still identifies the free Ca, free Mg, and soil moisture variables that were chosen for the CEC GWR model. The multicollinearity testing flagged the variable CEC from being too similar to both free Ca and Mg. The final exploratory regression is for the yield GWR model, and OLS determined significance estimation is not necessary as both soil properties are going to be used in this model, and only the multicollinearity between those two variables necessitated the function.

After reducing the redundant variables from the final model design, the GWR variables were finalized. The GWR model featuring CEC as the dependent variable with free Ca, free Mg, and soil moisture as the explanatory variables. The GWR model for the CaMgR uses the CaMgR as the dependent variable with the same design of free Ca, free

Mg, and soil moisture as the independent variables. For both of the GWR models, the kernel is set as fixed due to the Gaussian fixed distance kernel being the typical choice for normally distributed data. The bandwidth for both of the models is set to be determined using the Akaike information criterion. The results of the GWR model are displayed using Voronoi polygons because the initial yield cell polygons clipped for each soil sample were not able to be differentiated using a graduated color scheme due to size. The yield cell layer was clipped by the soil sample points in order to reduce the amount of data that was not truly representative of the soil properties and values that it was linked to. The result of that clipping process resulted in small polgyons with large distances between. The process of mapping simple Voronoi polygons uses the value of each individual soil sample to create a single polygon for each of the soil samples that encompasses the space between the nearest neighboring point and respective polygon.

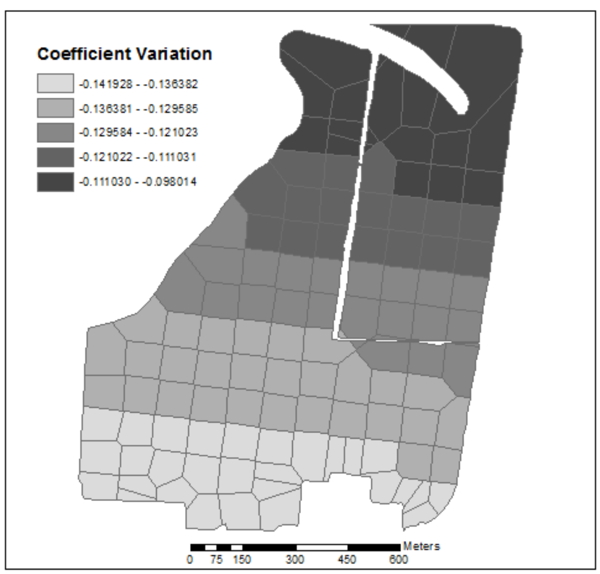

Figure 4.11 Soil Moisture Coefficient Variation from Cation Exchange Capacity Geographically Weighted Regression

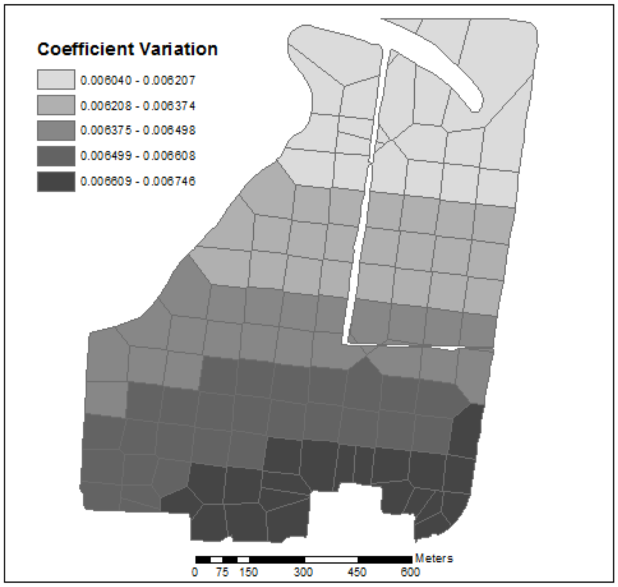

Figure 4.12 Magnesium Coefficient Variation from Cation Exchange Capacity Geographically Weighted Regression

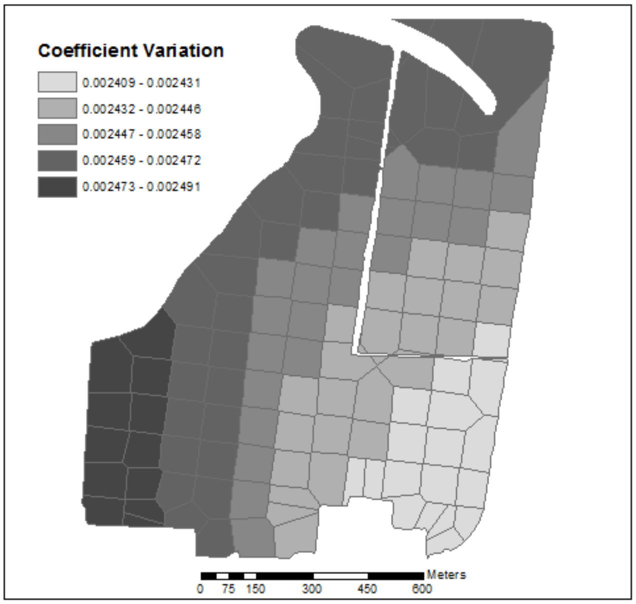

Figure 4.13 Calcium Coefficient Variation from Cation Exchange Capacity Geographically Weighted Regression

The model for CEC resulted in a local R-squared value of 0.98 and an adjusted R-squared value of 0.979. The Ca, Mg, and soil moisture variable parameters are mapped using the difference between the GWR estimated values and the observed values for those variables. Both Ca and Mg are positive with relatively low difference between the observed values and predicted values. The spatial orientation of both soil moisture and Mg are relatively linear across the plot from the northern edge to the southern edge. The soil moisture variable is negative but the differential is still very miniscule considering the overall fit of the model. The positive Ca and Mg variable residuals have a small range of variation which are both oriented in different linear directions. Using explanations of the soil properties and characteristics, the soil type and elevation changes across the field plot could account for these changes in free Ca and Mg. CEC is comprised of soil mineral base percentages meaning that the free Ca and Mg would not be included in the base percentages of Ca and Mg that make up the CEC. The fluctuations of the freely available Ca and Mg in the field would be separate from those that determine the CEC.

The CEC is a measure of the ability of the soil to maintain nutrients and varies by soil type, composition, and moisture. Variable soil moisture could be causing nutrient leaching within the soil and transportation of salt causing lower soil electrical conductivity values. The high correlation between the free Ca and Mg in the soil and the CEC values could be a measure of the impact of this leaching observable in the values of those free nutrients. It is also of note that the amounts of free Ca and Mg in the soil have been previously found to be not spatially dependent which implicates soil processes (Cambardella 1994). Leaching of the free Ca and Mg could be directly reflected by the

CEC variable values which would link those values not to the composition of the soil CEC but how soil properties are acting upon the CEC as a soil process.

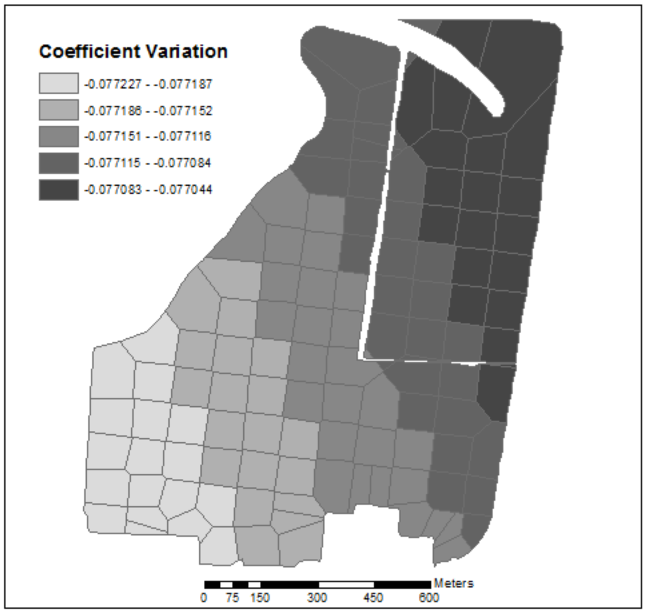

Figure 4.14 Soil Moisture Coefficient Variation from Calcium-Magnesium Ratio Geographically Weighted Regression

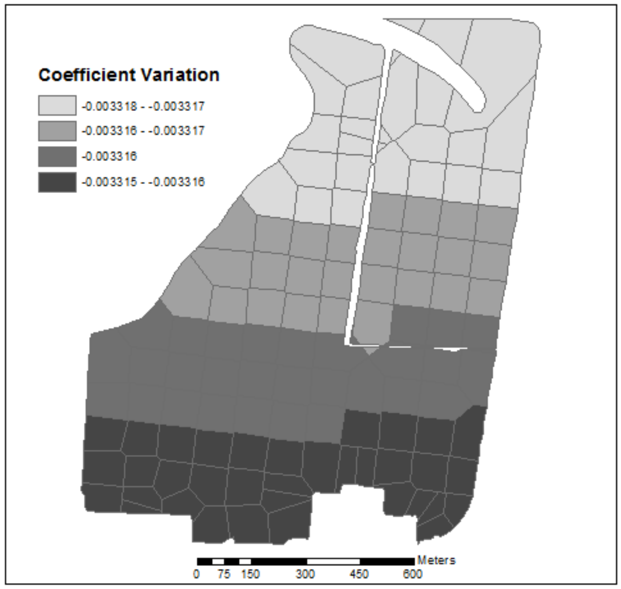

Figure 4.15 Magnesium Coefficient Variation from Calcium-Magnesium Ratio Geographically Weighted Regression

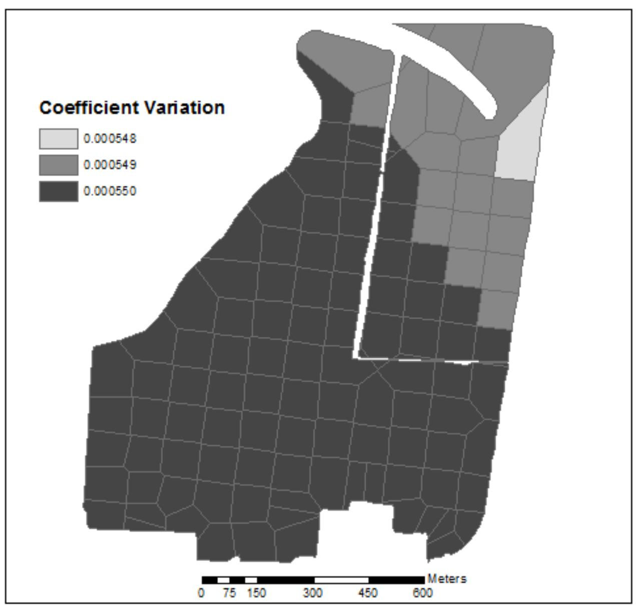

Figure 4.16 Calcium Coefficient Variation from Calcium-Magnesium Ratio Geographically Weighted Regression

The model for the CaMgR had a significantly lower R-squared value of 0.257 and an adjusted R-squared of 0.236. The Ca and Mg variables directly determine the CaMgR as it is a ratio of the two, but the multicollinearity observable in the initial design of the regression model linked free Ca and Mg to the CEC. There was a strong observable multicollinearity between the CEC and the free Ca and Mg variable values. This extent of similarity dictated that the CEC be excluded from the analysis in lieu of the free Ca and Mg. The CEC was also excluded because the inclusion of all three variables would have fatally flawed the analysis due to high levels of correlation between independent variables.

There is a very low amount of the data explained using this design, roughly only 23.6% using the R-squared value as a measure of model parsimony. The CaMgR is the ratio of free Ca and Mg in the soil which could point to a multicollinear model design flaw, but the free amounts should instead explain the spatial variation of CaMgR as a dependent variable. The soil moisture has a higher correlation than that of both the free Ca and Mg in the initial OLS exploratory regression models which would indicate that the CaMgR is impacted by underlying soil processes. Soil moisture could potentially be leaching free minerals and nutrients in the soil to areas with increased run off, but the elevation values for the soil were not correlated to a high extent.

The final GWR model created in this study uses the yield variable as the dependent variable with CEC and CaMgR set as the independent variables. This model has an adjusted R-squared value of 0.194 which means that there was very low model performance in determining the changes in yield relative to changes in the values for CEC and CaMgR. This model fit outcome was expected after the initial OLS exploratory regression variable correlations. The importance of this regression model is that the low parsimony directly points to underlying soil processes that effect the yield in terms of both the CEC and CaMgR values.

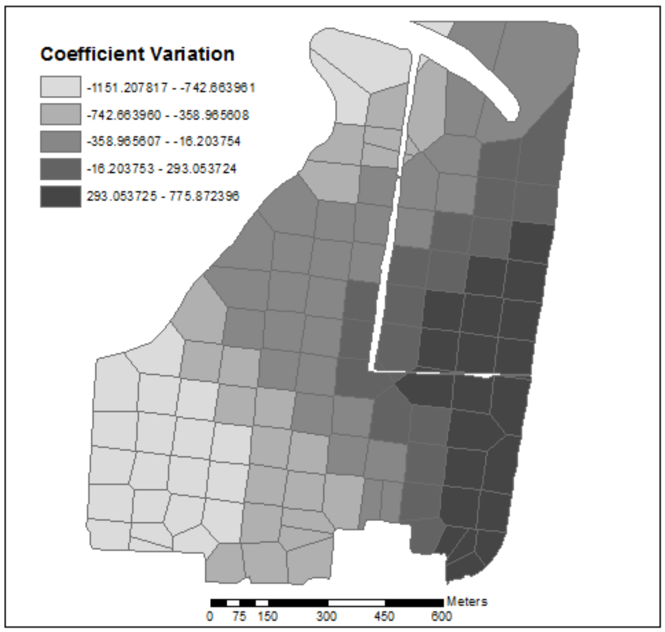

Figure 4.17 Calcium-Magnesium Ratio Coefficient Variation from Yield Geographically Weighted Regression

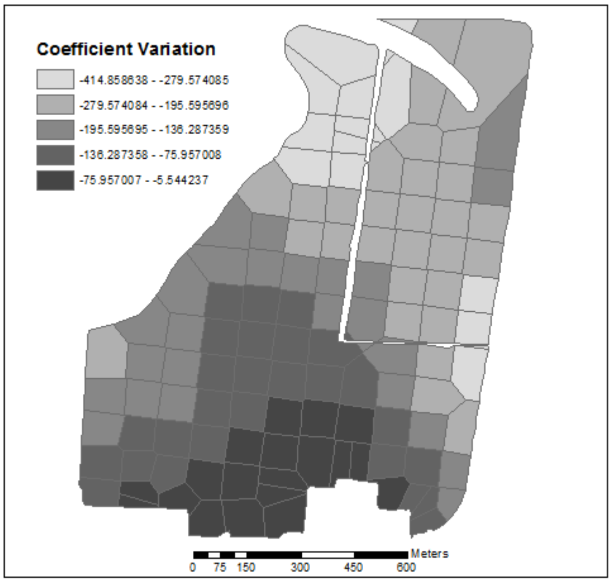

Figure 4.18 Cation Exchange Capacity Coefficient Variation from Yield Geographically Weighted Regression

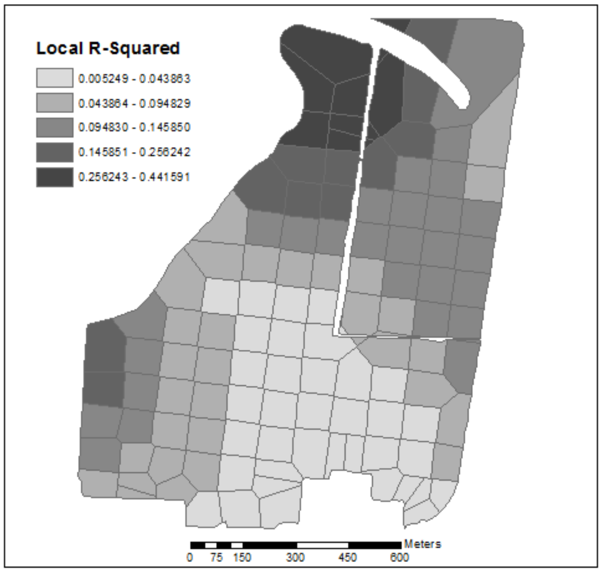

Figure 4.19 Local R^2 of Yield Geographically Weighted Regression

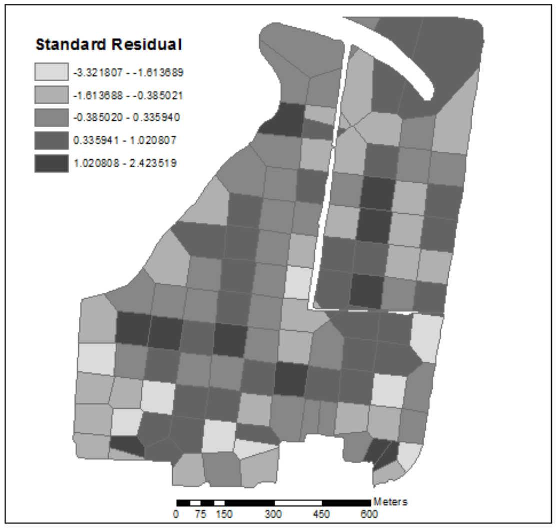

Figure 4.20 Standard Residual of Yield GWR

The hot spot analysis results can be used as a reference when looking at the local R-squared results of the yield GWR model. The highest localized measure of model fit was present where there is both low soil moisture and low elevation for the target area in the northwest corner. The large coefficient standard errors for both the CaMgR and the CEC pose the question of localized collinearity and to what extent that affects this regression model. Both explanatory variables exhibit a very high coefficient standard error in the northwest corner where the localized R-squared is between 0.41 and 0.44. This area presents the highest measure of model fit with one of the largest amounts of coefficient standard error.

CHAPTER V

DISCUSSION

The GWR results correlate the CEC with both the Ca and Mg values present in the soil samples for the study area. The CEC is comprised of base percentages of H, P, Ca, and Mg so the correlation is logical. Soil sample values for the base percentages of Ca and Mg are different from the amounts of those present in the soil individually. Another contributing variable to the CEC that was identified in the model design process using exploratory regression is the soil moisture. The properties of the soil nutrients to leech and transport to different areas based on soil moisture and varying elevation over the field surface could pair with the soil moisture values present in the soil samples. The trend of the GWR results being oriented from the Northern section of the field to the Southern section could be in part due to lower elevation values present at the Northern most sections of the field. This could explain the CEC values of the field being affected by the soil moisture because of nutrients being able to be more easily transported through soils experiencing a lower sloping observable elevation.

The CaMgR is a direct reflection of the ratio between the free Ca and Mg values present within the soil. This is problematic in terms of the low correlation between Ca and Mg as explanatory variables for the CaMgR in the GWR function. A potential explanation for this is the multicollinearity that had to be addressed during the design process of the model and determining which variables were necessary for inclusion. The

Ca and Mg variables conflict with CEC to the point of fatally flawing the GWR model. Since CEC is a measure of the ability of the soil to maintain nutrients, the multicollinearity of the variables could be skewing the regression model for this particular dependent variable.

The relationship between the soil sample values for free Ca, free Mg, soil moisture, and the CEC can be used to determine how the optimal individual levels of Ca and Mg affect the CEC. Optimal soil levels of Ca and Mg have previously been defined in terms of value ranges which can then be used to determine what impact those particular ranges have on the CEC over an entire field plot. Maximizing the ability of field soils to maintain nutrients in terms of the CEC could increase yield at a field plot level. This optimization would necessitate precision agricultural practices to address individual soil variables to a greater degree than currently in place.

The final GWR model using yield as a dependent variable maintained a poor measure of model fit. Changes in the yield are to a large degree not dependent on the CEC and CaMgR. This result shows that there are underlying soil processes or potentially plant dependent processes that impact the yield or the CEC and CaMgR. There are limitations for his study and uncovering complex underlying soil processes was anticipated from the beginning. There is also a limitation of the study due to the lack of soil samples from more than one date over the course of the growing season. Precipitation over the entire growing season would also have eliminated potential skewing of the results due to a single soil moisture reading taken at the time of harvest. There is further research required to determine what impact underlying soil processes such as leaching, temporal changes in the soil composition, and varying amounts of precipitation.

The findings of these models can be used as a starting point for future research towards providing optimized levels for these soil properties. The limitations of this study necessitate further research with numerous sampling dates throughout the growing season and additional knowledge of chemical applications to the soil. In terms of the findings of this study, the correlation between the CEC and freely available Ca and Mg in the soil is useable to provide insight for precision agricultural management decisions. As the CEC is a measure of the soil's ability to maintain nutrients (Corwin 2005), the relationship between changes in the CEC and changes in the values for soil moisture, Ca, and Mg is a reflection of nutrient availability. This correlation between the CEC and the explanatory variables could provide a much more accurate assessment of the nutrient availability of the soil during the growing season.

CHAPTER VI

CONCLUSION

This study seeks to identify the relationship between multiple field soil properties and increase knowledge of how they are correlated. The methods used to accomplish this involved first performing a multi-distance spatial cluster analysis to determine the dispersion of the variables present within the soil samples. In order to determine the degree to which each of the individual variables followed the overall trend of dispersion a spatial autocorrelation analysis was necessary. After the spatial orientation of the variable values was defined in terms of spatial autocorrelation, a local cluster and outlier analysis was used to readdress the clustering present in each of the variables. This is necessary because the initial Ripley's K function was used to display clustering over the entire set of data values for each variable at once, whereas the local cluster and outlier analysis was used for individual variables.

These variables were then processed using an ordinary least-squares based exploratory regression tool. This tool ran specified independent variables through regression functions in order to determine which of those variables expressed fatal amounts of multicollinearity. After the explanatory variables were decided for both the CEC and CaMgR regression models were defined, the selected model parameters were modeled using the GWR model. Correlation between the variables was then defined using determinants of model fit such as the adjusted R-squared values.

The GWR model for CEC maintained an adjusted R-squared of 0.979 while the CaMgR model exhibited an adjusted R-squared value of 0.236. The CEC model performed to a much more successful degree than that of the CaMgR model. The correlation between CEC and the specified explanatory variables could be used to optimize the amounts of the explanatory variables present in field soils. With these amounts optimized, the CEC could then be adjusted to increase the amounts of nutrients that the soil is able to maintain, which could in turn lead to increased yield. The CaMgR model requires further study as there could be additional independent variables necessary.

REFERENCES

Anderson-Cook, C. M., Alley, M. M., Roygard, J. K. F., Khosla, R., Noble, R. B., & Doolittle, J. a. (2002). Differentiating Soil Types Using Electromagnetic Conductivity and Crop Yield Maps. *Soil Science Society of America Journal*, *66*, 1562.

Aziz, S. A. (2008). Development of digital elevation models (DEMs) for agricultural applications, 130.

Black, J. R., & Thompson, S. R. (1978). Some Evidence on Weather-Crop-Yield Interaction. *American Journal of Agricultural Economics*, *60*(3), 540–543.

Cambardella, C. A., et al. (1994). Field-scale variability of soil properties in central Iowa soils. *Soil Science Society Of America,* 58(5), 1501-1511.

Charlton, M. (2009). Geographically Weighted Regression White Paper. *Science Foundation Ireland*.

Corwin, D. L., & Lesch, S. M. (2005). Apparent soil electrical conductivity measurements in agriculture. *Computers and Electronics in Agriculture*, *46*, 11–43.

Dawson, I. J., & Johnson, J. V. (2014). Growing pains: How risk perception and risk communication research can help to manage the challenges of global population growth. Risk Analysis: An Official Publication Of The Society For Risk Analysis, 34(8), 1378-1390. doi:10.1111/risa.12180

Dimitrakopoulos, R., Mustapha, H., & Gloaguen, E. (2010). High-order Statistics of Spatial Random Fields: Exploring Spatial Cumulants for Modeling Complex Non-Gaussian and Non-linear Phenomena. *Mathematical Geosciences*, *42*(1), 65–99.

Drewniak, B., Song, J., Prell, J., Kotamarthi, V. R., & Jacob, R. (2013). Modeling agriculture in the Community Land Model. *Geoscientific Model Development*, *6*, 495–515.

Duque, J. C., Aldstadt, J., Velasquez, E., Franco, J. L., & Betancourt, A. (2011). A computationally efficient method for delineating irregularly shaped spatial clusters. *Journal of Geographical Systems*, *13*(4), 355–372.

Fotheringham, a. S., Charlton, M. E., & Brunsdon, C. (1998). Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. *Environment and Planning A*, *30*(11), 1905–1927.

Gatrell, a. C., & Bailey, T. C. (1996). Interactive spatial data analysis in medical geography. *Social Science & Medicine*, *42*(6), 843–855.

Goovaerts, P. (1998). Geostatistical tools for characterizing the spatial variability of microbiological and physico-chemical soil properties. *Biology and Fertility of Soils* 27.4: 315-34.

Green, T. R., & Erskine, R. H. (2004). Measurement, scaling, and topographic analyses of spatial crop yield and soil water content. *Hydrological Processes*, *18*(May 2002), 1447–1465.

Green, T. R., Salas, J. D., Martinez, A., & Erskine, R. H. (2007). Relating crop yield to topographic attributes using Spatial Analysis Neural Networks and regression. *Geoderma*, *139*, 23–37.

Holmes, K. W., Chadwick, O. a., & Kyriakidis, P. C. (2000). Error in a USGS 30-meter digital elevation model and its impact on terrain modeling. *Journal of Hydrology*, *233*, 154–173.

Hu, Q., & Buyanovsky, G. (2003). Climate Effects on Corn Yield in Missouri*. *Journal of Applied Meteorology*, *42*, 1626–1635.

Huo, X.-N., Zhang, W.-W., Sun, D.-F., Li, H., Zhou, L.-D., & Li, B.-G. (2011). Spatial pattern analysis of heavy metals in Beijing agricultural soils based on spatial autocorrelation statistics. *International Journal of Environmental Research and Public Health*, *8*(6), 2074–2089.

Karlen, D. L., Cambardella, C. a., Kovar, J. L., & Colvin, T. S. (2013). Soil quality response to long-term tillage and crop rotation practices. *Soil and Tillage Research*, *133*, 54–64.

Kaufmann, R. K., & Snell, S. E. (1997). Biophysical Model Integrating Climatic of Corn Yield: and Determinants. *American Journal of Agricultural Economics*, *79*(1), 178–190.

Kravchenko, A. N., & Bullock, D. G. (2000). Correlation of corn and soybean grain yield with topography and soil properties. *Agronomy Journal*, *92*, 75–83.

Kroll, C. N., & Song, P. (2013). Impact of multicollinearity on small sample hydrologic regression models. *Water Resources Research*, *49*(6), 3756–3769.

Kühn, J., Brenning, A., Wehrhan, M., Koszinski, S., & Sommer, M. (2009). Interpretation of electrical conductivity patterns by soil properties and geological maps for precision agriculture. *Precision Agriculture*, *10*, 490–507.

Lauren, B., & Pratt, M. (2011). Finding a Meaningful Model: This checklist will help you evaluate regression models. *ArcUser*, *Winter*, 40–45.

Lobell, D. B., & Asner, G. P. (2003). Climate and management contributions to recent trends in U.S. agricultural yields. *Science*, *300*(February), 2003.

Marques da Silva, J. R., & Silva, L. L. (2008). Evaluation of the relationship between maize yield spatial and temporal variability and different topographic attributes. *Biosystems Engineering*, *101*(2), 183–190.

Mendelsohn, R. (2007). What causes crop failure? *Climatic Change*, *81*, 61–70.

Moore, I. D., Gessler, P. E., Nielsen, G. A., & Peterson, G. A. (1993). Soil Attribute Prediction Using Terrain Analysis. *Soil Science Society of America Journal*.

Officer, S. J., Kravchenko, a., Bollero, G. a., Sudduth, K. a., Kitchen, N. R., Wiebold, W. J., … Bullock, D. G. (2004). Relationships between soil bulk electrical conductivity and the principal component analysis of topography and soil fertility values. *Plant and Soil*, *258*, 269–280.

Olness, a., Palmquist, D., & Rinke, J. (2001). Ionic ratios and crop performance: II. Effects of interactions amongst vanadium, phosphorus, magnesium and calcium on soybean yield. *Journal of Agronomy and Crop Science*, *187*, 47–52.

Olness, A., Archer, D. W., Gesch, R. W., & Rinke, J. (2002). Resin-Extractable Phosphorus , Vanadium , Calcium and Magnesium as Factors in Maize (Zea mays L .) Yield, *101*, 94–101.

Pfeiffer, D. (1996). Issues related to handling of spatial data. *Massey University, Palmerston North*, (June), 23–28.

Robertson, A. W., Ines, a. V. M., & Hansen, J. W. (2007). Downscaling of seasonal precipitation for crop simulation. *Journal of Applied Meteorology and Climatology*, *46*, 677–693.

Rodrigues, M., de la Riva, J., & Fotheringham, S. (2014). Modeling the spatial variation of the explanatory factors of human-caused wildfires in Spain using geographically weightedlogistic regression. *Applied Geography*, *48*, 52–63.

Rosenshein, L., Scott, L., Pratt, M., & Esri. (2011). Exploratory Regression - A tool for modeling complex phenomena. *ArcUser Winter 2011*, 7. Retrieved from

Schlenker, W., & Roberts, M. J. (2009). Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(37), 15594–15598.

Schmidt, F., & Persson, A. (2003). Comparison of DEM data capture and topographic wetness indices. *Precision Agriculture*, *4*(1979), 179–192.

Seo, S. N. (2013). An essay on the impact of climate change on US agriculture: Weather fluctuations, climatic shifts, and adaptation strategies. *Climatic Change*, *121*, 115–124.

Thomson, A., Brown, R., & Ghan, S. (2002). Elevation dependence of winter wheat production in eastern Washington State with climate change: A methodological study. *Climatic Change*, *54*, 141–164.

Trangmar, B. B. (1985). Application of geostatistics to spatial studies of soil properties. *Advances in Agronomy*. 38: 45-92.

Villamil, M. B., Davis, V. M., & Nafziger, E. D. (2012). Estimating factor contributions to soybean yield from farm field data. *Agronomy Journal*, *104*, 881–887.

Vrindts, E., Reyniers, M., Darius, P., De Baerdemaeker, J., Gilot, M., Sadaoui, Y., … Destain, M. F. (2003). Analysis of soil and crop properties for precision agriculture for winter wheat. *Biosystems Engineering*, *85*, 141–152.

Wechsler, S. (2006). Uncertainties associated with digital elevation models for hydrologic applications: a review. *Hydrology and Earth System Sciences Discussions*, *3*, 2343–2384.