2009

Spatial Clustering Using the Likelihood Function

April Kerby
University of Nebraska at Lincoln, k2girl2000@yahoo.com

SPATIAL CLUSTERING USING THE LIKELIHOOD FUNCTION

by

April T. Kerby

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Statistics

Under the Supervision of Professor David B. Marx

Lincoln, Nebraska

August 2009

SPATIAL CLUSTERING USING THE LIKELIHOOD FUNCTION

April T. Kerby, Ph.D.

University of Nebraska, 2009

Advisor: David B. Marx

Researchers have been using clustering algorithms for many years to group similar observations based on a set of recorded characteristics. The majority of these algorithms maximize the similarity of the observations within a cluster, while at the same time maximize the dissimilarity with observations in other clusters. However, nearly all of the current clustering algorithms do not take into account the actual geographic location of the observation during the clustering process. This dissertation consists of three papers which propose a method to incorporate the geographical location of an observation into the clustering algorithm, known as spatial clustering.

The first paper examines spatial clustering when only one numeric response has been recorded for each observation. The geographic or spatial location is incorporated into the likelihood of the multivariate normal distribution through the variance-covariance matrix. The variance-covariance matrix is computed using any appropriate spatial covariance function, although the spherical covariance function was used for this research. The second paper extends the clustering algorithm to the multivariate case, i.e. when more than one response has been recorded on each observation. Again, the spatial location is incorporated through the variance-covariance matrix of the multivariate

normal distribution. However, the actual construction of the variance-covariance matrix must take into account the cross-covariance between the variates. Oliver's (2003) approach for modeling the cross-covariance is incorporated into the clustering algorithm.

Since not all recorded variables of interest are numeric, the third paper investigates incorporating categorical (non-numeric) responses into the spatial clustering algorithm. This paper looks first at the case where only categorical responses are recorded on the observations. After this has been implemented, the final step is to spatially cluster observations which contain both numeric and categorical responses. The algorithm must account for the spatial pattern of the data, the actual numeric responses and the categorical responses, and an appropriate weighting of the spatial component is determined. The final clustering algorithm clusters both numeric and categorical data while incorporating the geographic location of the observations.

ACKNOWLEDGEMENTS

This dissertation could not have been written without the help of my advisor, Dr. David Marx. I am grateful for your guidance and encouragement throughout this process. I would also like to thank my other committee members, Dr. Erin Blankenship, Dr. Stephen Kachman, Dr. Ashok Samal, and Dr. Viacheslav Adamchuk for your guidance and support throughout my academic program. This research was supported in part by the Hatch Act and through the Channing B. and Katherine W. Baker Fund #3424 grant provided by the University of Nebraska Agricultural Research Division. I would especially like to thank Bruce, my parents and my friends for their patience, support and encouragement during this time.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

Researchers have been using clustering algorithms for many years to group similar observations based on a set of recorded characteristics. The majority of these algorithms maximize the similarity of the observations within a cluster, while at the same time maximize the dissimilarity with observations in other clusters. However, nearly all of the current clustering algorithms do not take into account the actual geographical location of the observation during the clustering process. Those that do are relatively ad hoc and do not account for the underlying spatial structure of the variables measured (Cuzick & Edwards 2006, Lee 2005, Ng & Han 1994, Simbahan & Dobermann 2006).

This dissertation consists of three papers which propose a method to incorporate the geographical location of an observation into the clustering algorithm, known as spatial clustering. Thus, groups of observations that are formed will not only have similar characteristics but will also be similar in location. That is, observations may only be grouped if they have similar characteristics and are located in the same "neighborhood." Earlier work in this area has shown that the likelihood is one way to allow spatial structure to be incorporated into the clustering algorithm (Kerby et al. 2007, 2008 & 2009).

1.1 Spatial Clustering in the Univariate Case

Chapter 2 examines the case when only one numeric response variable is provided or is of interest. The geographic or spatial location of the observations can be incorporated into the likelihood of the multivariate normal distribution through the variance-covariance matrix. The variance-covariance matrix is computed using any appropriate spatial covariance function, although the spherical covariance function was used for this research since it is the most popular case in natural resources. However, if the numeric responses are not spatial in nature, a simple covariance function, possibly the linear, should be chosen when computing the variance-covariance matrix. The likelihood function will be larger when the observations in the clusters are spatially close to one another rather than spread apart or noncontiguous.

In addition to the algorithm itself, the likelihoods computed using a specific covariance function can be used to evaluate different clustering schemes created based upon expert opinion to determine which scheme best clusters the data. Since there are numerous clustering schemes for a given set of data, Chapter 2 also discusses methods of choosing the optimal clustering of the data based on Akaike's Information Criterion (AIC) and the likelihood itself.

An example using one numeric response variable was carried out to see which clustering scheme best suited the simulated data. A second example was presented which utilizes data from a Kansas field study. pH readings from a 23-ha field were analyzed to determine if groupings of pH levels existed in the field. Again, experts used their knowledge of precision agriculture, as well as the nature of the field itself, to create various clustering schemes that are evaluated using the likelihood approach.

Since the spatial clustering algorithm is specifically incorporating the geographical location of the observations, it should have more emphasis on the results of the analysis. Therefore, weighting the purely spatial component of the multivariate normal distribution was investigated. This allowed the spatial component of an observation to play a larger factor in the clustering process. Various combinations of the spatial parameters were used to get a better idea of how much weighting is needed to ensure a spatial component in the clustering algorithm.

1.2 Spatial Clustering in the Multivariate Case

If one is performing a cluster analysis, usually more than one numeric response has been recorded on an observation. Thus, Chapter 3 extends the clustering algorithm to account for more than one numeric response variable, i.e. the multivariate case. This chapter focuses on the ability to model the cross-covariance matrix between the response variables while still taking into account the spatial location of the observations. The spatial component is still incorporated into the variance-covariance matrix of the multivariate normal distribution. However, the actual construction of the variance-covariance matrix must take into account the cross-covariance between the response variables. Oliver's (2003) approach for modeling the cross-covariance is incorporated into the clustering algorithm. Again, a simulated data set with two numeric response variables was used to demonstrate this method.

1.3 Spatial Clustering Incorporating Categorical Data

Since not all recorded variables of interest are numeric, Chapter 4 investigates incorporating categorical (non-numeric) responses into the spatial clustering algorithm. This paper looks first at the case when only categorical responses have been recorded on the observations. After this has been implemented, the final step is to spatially cluster observations which possess both numeric and categorical responses. The algorithm must account for the spatial pattern of the data, the numeric responses and the categorical responses, and an appropriate weighting of the spatial component is determined. The final clustering algorithm clusters both numeric and categorical data while taking into account the actual geographical location of the observations.

Chapter 2

Spatial Clustering in the Univariate Case

2.1 Introduction

Cluster analysis is a tool used to place similar observations in groups or clusters based on measures of similarity or dissimilarity. Observations are placed in clusters that maximize the similarity among observations within a cluster while at the same time maximize the dissimilarity to observations in other clusters (Everitt 1974, Hartigan 1975, Johnson 1998, Johnson & Wichern 2002, Kaufman & Rousseeuw 1990).

Most of the current clustering methods group observations based upon a distance calculation and the three most prominent are Euclidean distance,

$$d_{rs} = \sqrt{(\mathbf{x}_r - \mathbf{x}_s)'(\mathbf{x}_r - \mathbf{x}_s)} \tag{2.1}$$

standardized Euclidean distance,

$$d_{rs} = \sqrt{(\mathbf{z}_r - \mathbf{z}_s)'(\mathbf{z}_r - \mathbf{z}_s)} \tag{2.2}$$

and Mahalanobis distance

$$d_{rs} = \sqrt{(\mathbf{x}_r - \mathbf{x}_s)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_r - \mathbf{x}_s)}\,. \tag{2.3}$$

In Equations (2.1) and (2.3) above, $\mathbf{x}_r$ and $\mathbf{x}_s$ are multivariate observations. In Equation (2.2) $\mathbf{z}_r$ and $\mathbf{z}_s$ are the standardized observation values, and Equation (2.3) uses $\boldsymbol{\Sigma}$, the variance-covariance matrix between pairs of observations (Johnson 1998). These distances may be used in a variety of hierarchical or nonhierarchical clustering methods. Hierarchical clustering methods place observations together in a nested sequence of

clusterings. Nearest Neighbor and Hierarchical Tree Dendograms are popular tools used in hierarchical clustering (Johnson 1998, Johnson & Wichern 2002).

These clustering methods do not allow one to account for the spatial structure of the observations. However, there are cases for which spatial location is both known (e.g. encoded as latitude and longitude) and relevant to the goals of the data analysis. One example is site-specific crop management, which has become an important aspect of agriculture production in recent years. Precision agriculture methods use multiple data layers within spatially variable observations to fine-tune crop management decisions. Since conventional coarse (approximately 1-ha) grid sampling fails to provide adequate representation of spatial variability in soils, alternative high-density sensor data have been used in many operations.

One of the major challenges in the data analysis process is to delineate field areas with potential for differentiated treatments that are frequently called “management zones.” Initially, a relatively inexpensive set of data such as on-the-go soil sensing maps and/or remote sensing imagery are collected. These data are very dense and can be used to define areas for targeted (guided) sampling which will provide detailed information about the agronomic quality of land through the analysis of soil samples run in a commercial lab. Since only a limited number of these costly samples can be afforded, they should come from homogenous areas of the field, away from boundaries or locations where sensor data change significantly over short distances, and spread across the entire landscape. These samples should also uniformly cover the entire range of measurements, indicating spots of high, medium or low readings (Adamchuk et al. 2007, Frogbrook &

Oliver 2007). Certain agronomic properties could be related to a linear or other combination of multiple sensor data layers where the area of applicability of such relationships may be limited to a series of spatial clusters with relative homogeneity. Therefore, a proper clustering method should be developed to delineate relatively homogeneous field areas while accounting for the physical values of high-density observations as well as their spatial distribution.

Oliver and Webster (1989) proposed a clustering method based upon a modified dissimilarity matrix. First, the similarities between all pairs of observations are calculated using Gower's (1971) similarity coefficient which takes into account the values of the observations, as well as a weighting factor attributed to each specific property. The similarities calculated are then transformed into measures of dissimilarity. The dissimilarities are modified to take into account the geographic distance as shown in Equation (2.4):

$$d_{ij}^{*} = d_{ij} f(\mathbf{x_i} - \mathbf{x_j}). \tag{2.4}$$

The d_{ij} are the dissimilarity values while $\mathbf{x_i}$ and $\mathbf{x_j}$ denote the i^{th} and j^{th} locations. $f(\mathbf{x_i} - \mathbf{x_j})$ can be computed using any specific covariance function. Webster and Oliver (1989) chose the exponential function where d_{ij}^{*} becomes

$$d_{ij}^{*} = d_{ij} \frac{c}{c_0 + c}\left(1 - e^{\frac{-u_{ij}}{W}}\right) + d_{ij} \frac{c_0}{c_0 + c}. \tag{2.5}$$

The spatial parameters are c, c_0 and W, while u_{ij} is the distance between the i^{th} and j^{th} locations. Equation (2.5) consists of two parts, one of which can be modified

depending on the spatial structure and the other which cannot. Therefore, the appropriate spatial structure may be incorporated for each situation. The modified dissimilarity matrix $\mathbf{D}^*$ can then be used in a variety of clustering strategies. If the data are structured in such a way that hierarchical clustering is applicable, then the operations shall be performed directly on $\mathbf{D}^*$. However, if the data do not warrant hierarchical clustering, $\mathbf{D}^*$ shall be transformed into a new set of variables. These transformed variables are used in the clustering process (Webster & Oliver, 1989).

In this paper a clustering method is proposed to explicitly incorporate the spatial structure by using the likelihood values to form the clusters. That is, if two points are located far apart, their likelihood will be smaller than if the points were closer together. The spatial structure is present as part of the variance-covariance matrix in the likelihood.

2.2 Clustering Univariate Observations Using the Likelihood Function

The procedure proposed maximizes the likelihood for the multivariate normal distribution at every step (hierarchical clustering). Initially, each observation is considered to form its own cluster, resulting in n clusters. The likelihood is computed for each possible pairing of two "clusters." The pairing which yields the largest likelihood is merged together to form a new cluster. After one step there are $n-1$ clusters (one cluster has two observations and the remaining $n-2$ clusters consist of only one observation each).

During step 2 all possible pairwise groupings of the $n-1$ clusters are evaluated. The pairing which gives the largest likelihood is selected as the new merged cluster. This continues until all the data are in one cluster.

To account for the spatial structure in the likelihood, the variance-covariance matrix is computed using any specific covariance function from which exponential, Gaussian and spherical are the most common. The frequently used spherical covariance function is given by,

$$C(d) = \begin{cases} \sigma^2 \left\{ 1 - \frac{3}{2}\left(\frac{d}{a}\right) + \frac{1}{2}\left(\frac{d}{a}\right)^3 \right\} & if\ d \le a \\ 0 & if\ d > a \end{cases} \tag{2.6}$$

where d is the distance between two points and a is the range of the variogram (Cressie 1991, Isaaks & Srivastava 1989, Schabenberger & Gotway 2005). The range is the separation distance at which an increase in distance no longer produces an increase in the average squared difference between pairs of values (Isaaks & Srivastava 1989). The Gaussian covariance function which works well with a small scale spatial structure is

$$C(d) = \sigma^2 e^{-\frac{3d^2}{a^2}}, \tag{2.7}$$

and the exponential covariance function is

$$C(d) = \sigma^2 e^{-\frac{3d}{a}} \tag{2.8}$$

which works best when there is less spatial structure at small distances. The Gaussian and exponential covariance functions have a similar range a, but they are not strictly identical, as the range refers to the rate at which the covariance function approaches the

sill (Cressie 1991, Isaaks & Srivastava 1989, Schabenberger & Gotway 2005). Figure 2.1 compares these covariance functions.

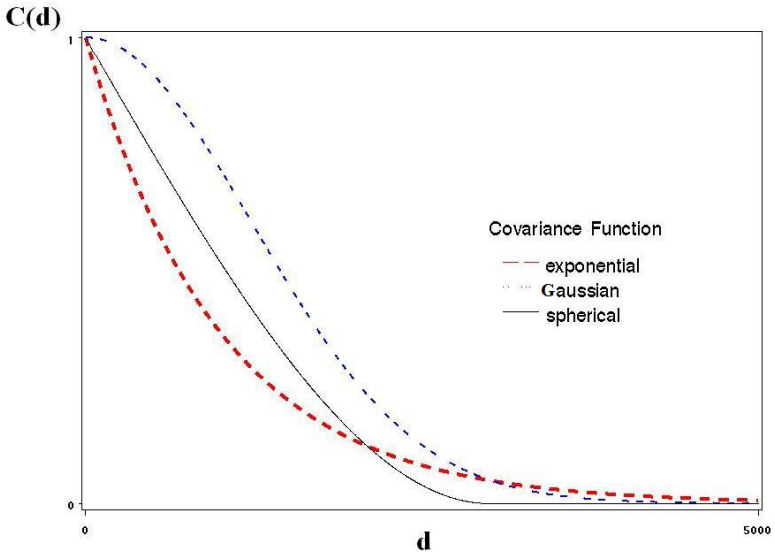

Figure 2.1: Comparison of covariance functions

The value of the variogram for a distance of zero is zero, however, due to sampling error and scale variability the values recorded at extremely small separation distances may be rather dissimilar causing discontinuity at the origin. The vertical jump from zero to these values is referred to as "the nugget effect" (Isaaks & Srivastava 1989), and must also be considered during spatial analyses. Since the spherical covariance function is most common when it comes to agronomic quality of soils, the examples provided in this paper use the spherical covariance function and assume no nugget effect.

The likelihood of the multivariate normal distribution can be written as

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{Nv/2}\left|\boldsymbol{\Sigma}\right|^{1/2}} e^{-1/2(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \tag{2.9}$$

where v is the number of clustering variates and $N = n_1 + n_2 + ... + n_c$, the sum of the number of observations which fall into each cluster, where c is the number of clusters. Under the univariate case (i.e. $v = 1$):

$\mathbf{x}' = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n_1} & x_{21} & \cdots & x_{cn_c} \end{pmatrix}$ where x_{ik} is the variate value of the k^{th} observation in the i^{th} cluster

i = 1, …, c where c is the number of clusters

k = 1, …, n_i where n_i is the total number of observations in the i^{th} cluster

$\boldsymbol{\mu}' = \begin{pmatrix} \mu_1 & \cdots & \mu_1 & \mu_2 & \cdots & \mu_c \end{pmatrix}$ where μ_i is the mean of the i^{th} cluster – there are n_i μ's in the i^{th} cluster

The variance-covariance matrix in equation (2.9) is given by $\boldsymbol{\Sigma} = \bigoplus_{i=1}^{c} \boldsymbol{\Sigma}_i$ where $\boldsymbol{\Sigma}_i$ is computed using the spherical covariance function ($sph(d_{kk'})$) from Equation (2.6):

$$\boldsymbol{\Sigma}_i = \begin{bmatrix} \sigma_i^2 & sph\left(\sigma_i^2, a_i, d_{i12}\right) & \cdots & sph\left(\sigma_i^2, a_i, d_{i1n_i}\right) \\ & \sigma_i^2 & \cdots & sph\left(\sigma_i^2, a_i, d_{i2n_i}\right) \\ & & \ddots & \vdots \\ & & & \sigma_i^2 \end{bmatrix}. \quad (2.10)$$

This is a symmetric matrix because $d_{ikk'}$ is the actual physical distance between observation units k and k', so sph(σ_i^2, a_i, d_{i12}) = sph(σ_i^2, a_i, d_{i21}) (Isaaks & Srivastava 1989).

2.3 Choosing the Optimal Number of Clusters

The likelihood function can be used to determine the optimal clustering scheme for a given set of data. A sharp increase in the plot of the likelihood against the number of clusters would indicate an appropriate number of delineated clusters. Since the likelihood is maximized at every step in the clustering process, an increase in the plot shows what clustering scheme(s) may be best.

An improvement over plotting the likelihood against the number of clusters would be to use Akaike's Information Criterion (AIC) (Akaike 1974). This criterion also uses the likelihood computed using a covariance function, while penalizing for the number of parameters being estimated. Since the ultimate goal is to maximize the likelihood, the parameter estimates are computed using maximum likelihood estimation (MLEs). The AIC is given by,

$$\text{AIC} = -2\log\left\{L\left(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}} \mid \mathbf{x}\right)\right\} + 2k \tag{2.11}$$

where k is the number of parameters estimated and $L(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}} \mid \mathbf{x})$ is the estimated likelihood given the data. For each cluster there are three parameters to estimate: sill, range and mean (assuming no nugget effect). Therefore, a penalty is imposed for having more clusters, i.e. more parameters to estimate. Thus, smaller AIC values are better.

The AIC is used as one of our deciding factors to determine the appropriate number of clusters for the data. A penalization for having a large number of clusters is important and is not taken into account when looking solely at the likelihood. Thus, both the likelihood and AIC values are summarized in the examples below, so both may be used in the decision making process.

2.4 Example 1: Simulated Data

The data for this example have been simulated to have a sill of 1, a range of 20 and no nugget effect. Without loss of generality, a nugget effect may be added, but the results presented are simply more dramatic without considering a nugget effect. A 10×10 grid was generated to ensure a strong spatial floor and the center 6×6 grid of the data was used. The smallest number of clusters is one indicating that all the data fall into just one cluster, and the largest number of clusters occurs when each point is its own cluster. Therefore, the largest number of clusters for this data set was 36.

Once the data were generated, random values from a normal distribution, with a mean of 25 were added to the middle diagonal of values. Similarly, random values from a normal distribution, with a mean of 10 were added to the top left and bottom right corners of the data grid to create a second cluster in the data. The data values are shown in Figure 2.2 representing the smallest possible clustering of the data, i.e. when all the points fall into one cluster.

20.78	**19.84**	**18.88**	**34.56**	**32.62**	**33.01**
20.85	**16.77**	**33.98**	**33.95**	**34.09**	**34.29**
18.88	**34.66**	**33.37**	**33.19**	**35.13**	**33.02**
37.33	**33.57**	**34.65**	**33.79**	**31.21**	**18.11**
34.13	**34.49**	**34.06**	**32.60**	**19.43**	**17.82**
35.43	**34.00**	**33.88**	**17.63**	**18.40**	**17.96**

Figure 2.2: Data values in one cluster scheme

Since the data values along the middle diagonal of the data grid were much larger than the values in the top left and bottom right corners of the grid, the observations were separated into two different clusters resulting in the two cluster scheme shown in Figure 2.3.

20.78	**19.84**	**18.88**	**34.56**	**32.62**	**33.01**
20.85	**16.77**	**33.98**	**33.95**	**34.09**	**34.29**
18.88	**34.66**	**33.37**	**33.19**	**35.13**	**33.02**
37.33	**33.57**	**34.65**	**33.79**	**31.21**	**18.11**
34.13	**34.49**	**34.06**	**32.60**	**19.43**	**17.82**
35.43	**34.00**	**33.88**	**17.63**	**18.40**	**17.96**

Figure 2.3: Data values in two cluster scheme (blue and red)

However, the goal of spatial clustering is to create spatially contiguous clusters. Therefore, the cluster which included the data values from the top left and bottom right corners of the data grid were broken into two clusters creating in the three cluster scheme found in Figure 2.4.

20.78	**19.84**	**18.88**	**34.56**	**32.62**	**33.01**
20.85	**16.77**	**33.98**	**33.95**	**34.09**	**34.29**
18.88	**34.66**	**33.37**	**33.19**	**35.13**	**33.02**
37.33	**33.57**	**34.65**	**33.79**	**31.21**	**18.11**
34.13	**34.49**	**34.06**	**32.60**	**19.43**	**17.82**
35.43	**34.00**	**33.88**	**17.63**	**18.40**	**17.96**

Figure 2.4: Data values in three cluster scheme (blue, red and orange)

Finally, the four cluster scheme was created by separating the three observations in the bottom left corner into their own cluster producing the four cluster scheme in Figure 2.5.

20.78	**19.84**	**18.88**	**34.56**	**32.62**	**33.01**
20.85	**16.77**	**33.98**	**33.95**	**34.09**	**34.29**
18.88	**34.66**	**33.37**	**33.19**	**35.13**	**33.02**
37.33	**33.57**	**34.65**	**33.79**	**31.21**	**18.11**
34.13	**34.49**	**34.06**	**32.60**	**19.43**	**17.82**
35.43	**34.00**	**33.88**	**17.63**	**18.40**	**17.96**

Figure 2.5: Data values in four cluster scheme (blue, red, orange, and green)

When the number of clusters was greater than the ability to adequately estimate the spatial parameters and the mean, the estimates were derived using the entire data set.

Three parameters need to be estimated in each cluster. These parameter estimates should be reasonable if there are at least six observations in a cluster. In this analysis, there seemed to be an adequate number of observations present to estimate the spatial parameters and the mean using the data for the one, two and three cluster schemes (Figures 2.2, 2.3 and 2.4). Therefore, the mean and spatial parameters were estimated using the data in these clustering schemes.

However, in the four cluster scheme (Figure 2.5) there were not enough observations in the fourth cluster (as shown in green in Figure 2.5) to estimate the spatial parameters and the mean. Therefore, the mean for the fourth cluster was estimated using the data in the cluster and the spatial parameters were estimated using the entire data set (Figure 2.2). There seemed to be an adequate number of observations in the remaining three clusters to estimate the mean and spatial parameters using the data itself in this clustering scheme. Figures 2.6 and 2.7 show plots of the log-likelihood and AIC values for the analysis.

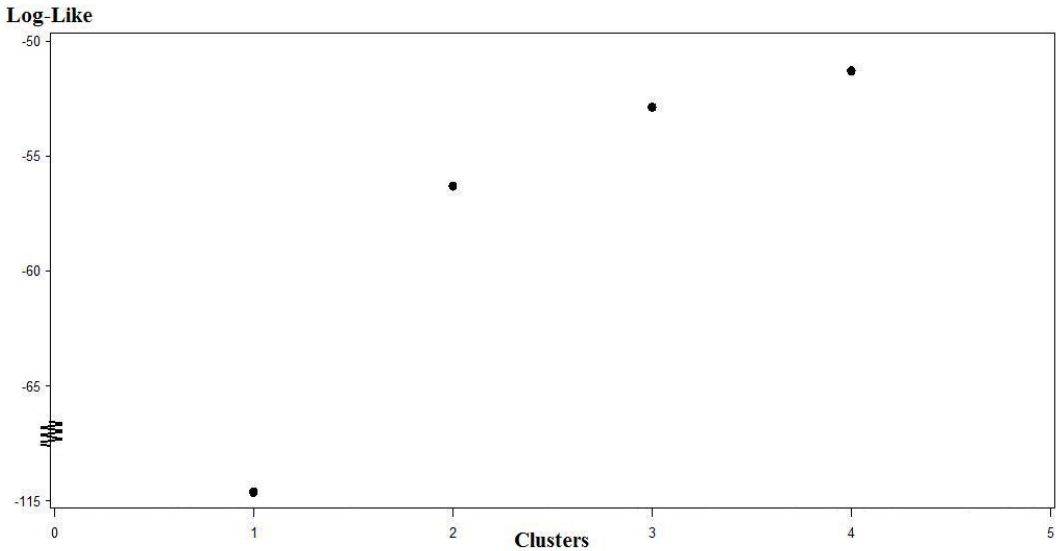

Figure 2.6: Plot of the log-likelihood values against the number of clusters

As shown in Figure 2.6 there is a sharp increase in the plot at two clusters. However, four clusters looks to have the largest log-likelihood value. Therefore, based solely on the likelihood, four clusters would be appropriate for these data.

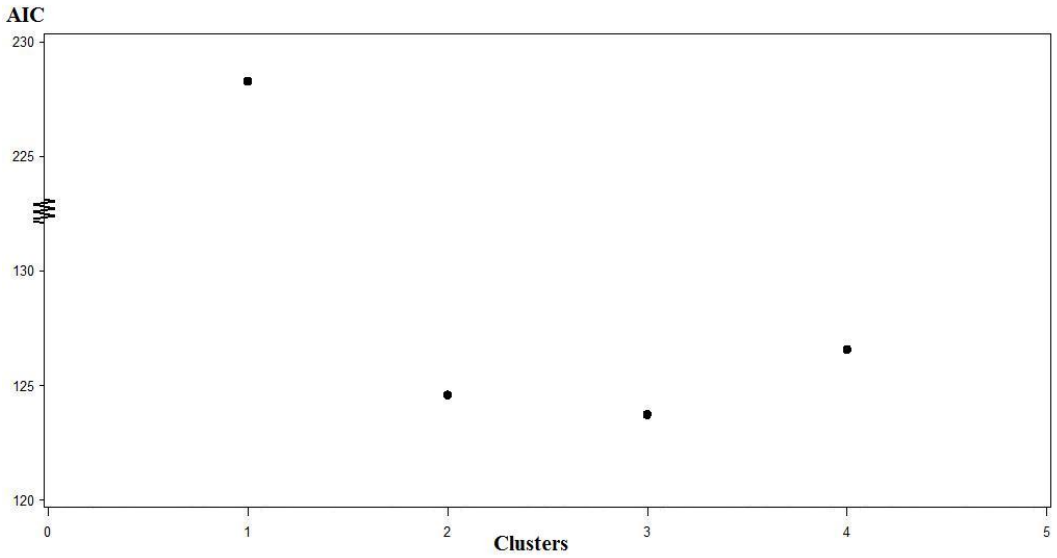

Figure 2.7: Plot of the AIC values against the number of clusters

Looking at Figure 2.7, the smallest AIC is produced when the data are clustered into three clusters, however the AIC for two clusters is only larger by 0.85. The actual log-likelihood and AIC values can be found in Table 2.1.

Number of Clusters	**Log-Likelihood**	**AIC**
1	-111.15	228.29
2	-56.30	**124.59**
3	**-52.87**	**123.74**
4	**-51.29**	126.58

Table 2.1: Clustering results from simulation study

The four cluster scheme produced the largest log-likelihood value. However, the log-likelihood for the three cluster scheme was only smaller than that of the four cluster

scheme by 1.58. The three cluster scheme produced a smaller AIC than the four cluster scheme since an additional penalty was assessed for a larger number of estimated parameters. Since the log-likelihoods from the three and four cluster schemes were close, the penalty for the additional cluster was enough to result in three clusters as the optimal clustering scheme to summarize the data.

2.5 Example 2: Kansas Field Study

The following example used a random subset of data (101 measurements) from a 23-ha field in Kansas which consisted of 598 soil pH measurements obtained using Mobile Sensor Platform (Veris Technologies, Inc., Salina, Kansas, USA) (Adamchuk et al. 2007). The data layer used in this research was univariate (soil pH only) as shown in Figure 2.8. This analysis was performed assuming no nugget effect, therefore only three parameters were estimated for each cluster: sill, range and mean.

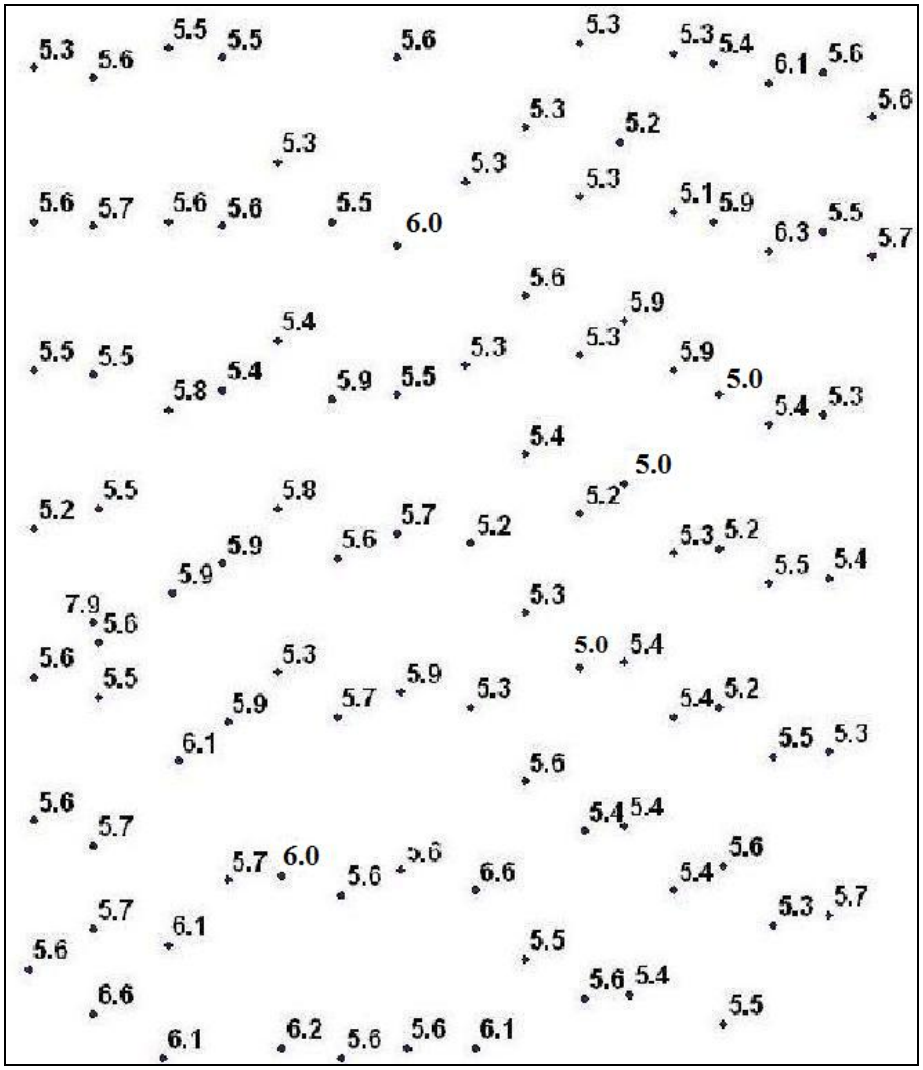

Figure 2.8: Data from Kansas field study

If there is no preliminary knowledge of the appropriate clustering arrangement, the hierarchical clustering process described in Section 2.2 shall be used. However, in this case experts not only used knowledge of the response variable, but other qualitative information as well to cluster the data. The clusters were assigned on the perceptions of what four expert individuals thought to be appropriate management zones of the data in regards to pH and spatial location. Therefore, not only can this spatial clustering approach be used to cluster the data in a hierarchical manner, but is can also be used to

evaluate and determine the optimal clustering scheme proposed by experts in a given subject area.

Each of the four experts consulted produced a three cluster and a four cluster scheme of the pH data. Therefore, four variations of each of two cluster sizes were analyzed. Figures 2.9, 2.10, 2.11, and 2.12 show the three cluster schemes provided by the experts.

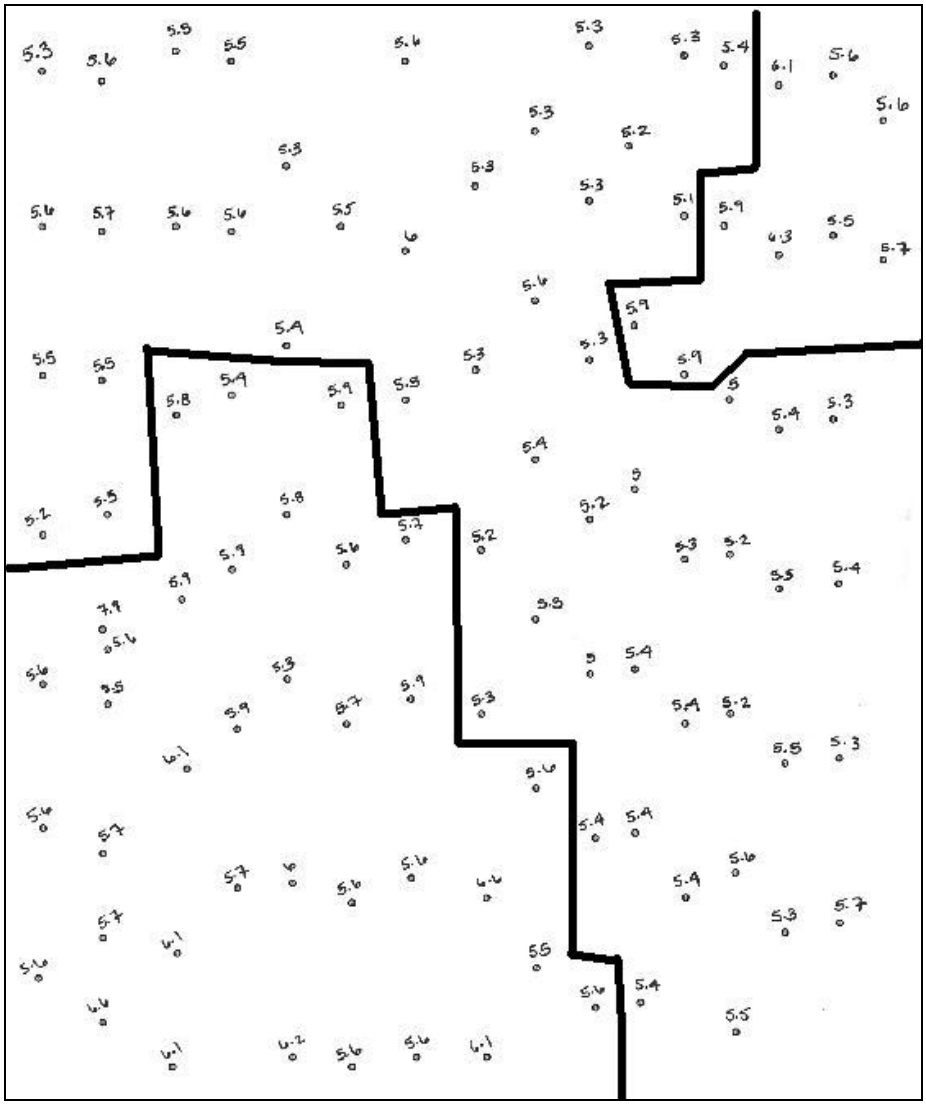

Figure 2.9: Three cluster scheme variation 1

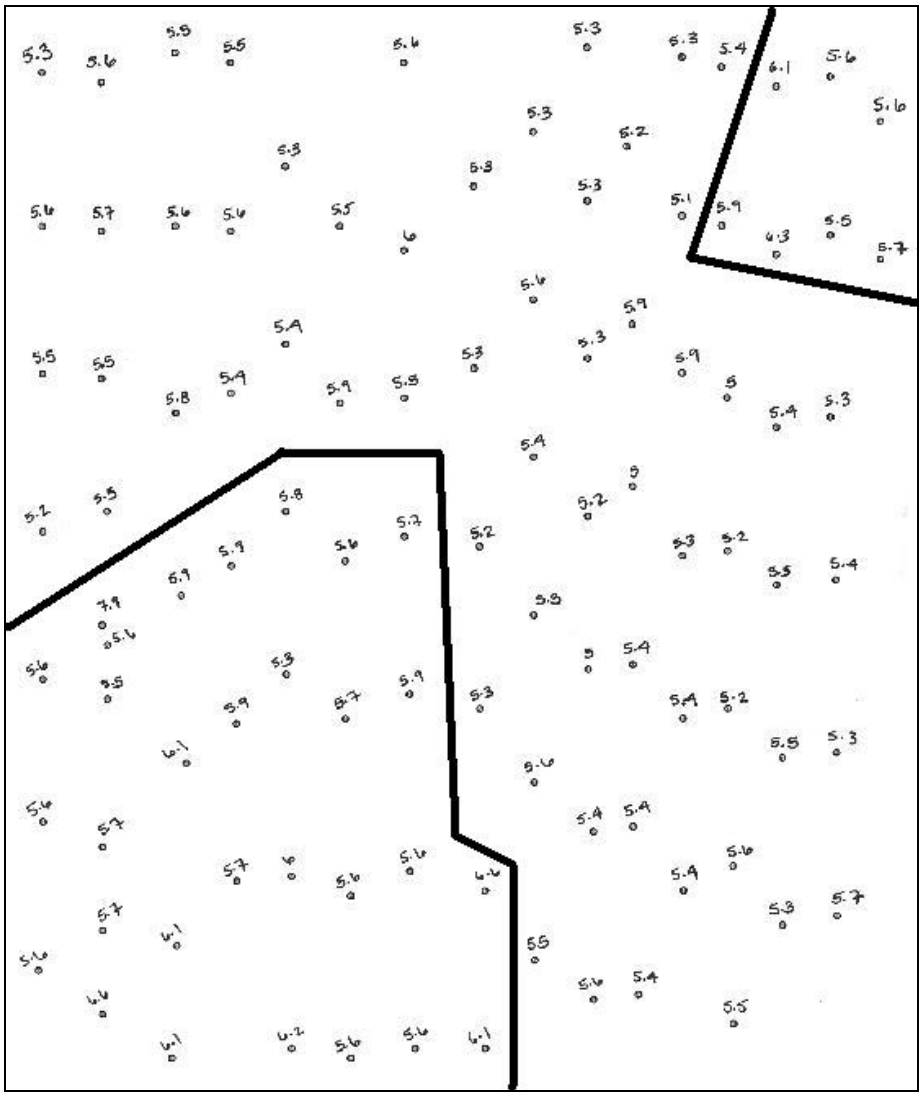

Figure 2.10: Three cluster scheme variation 2

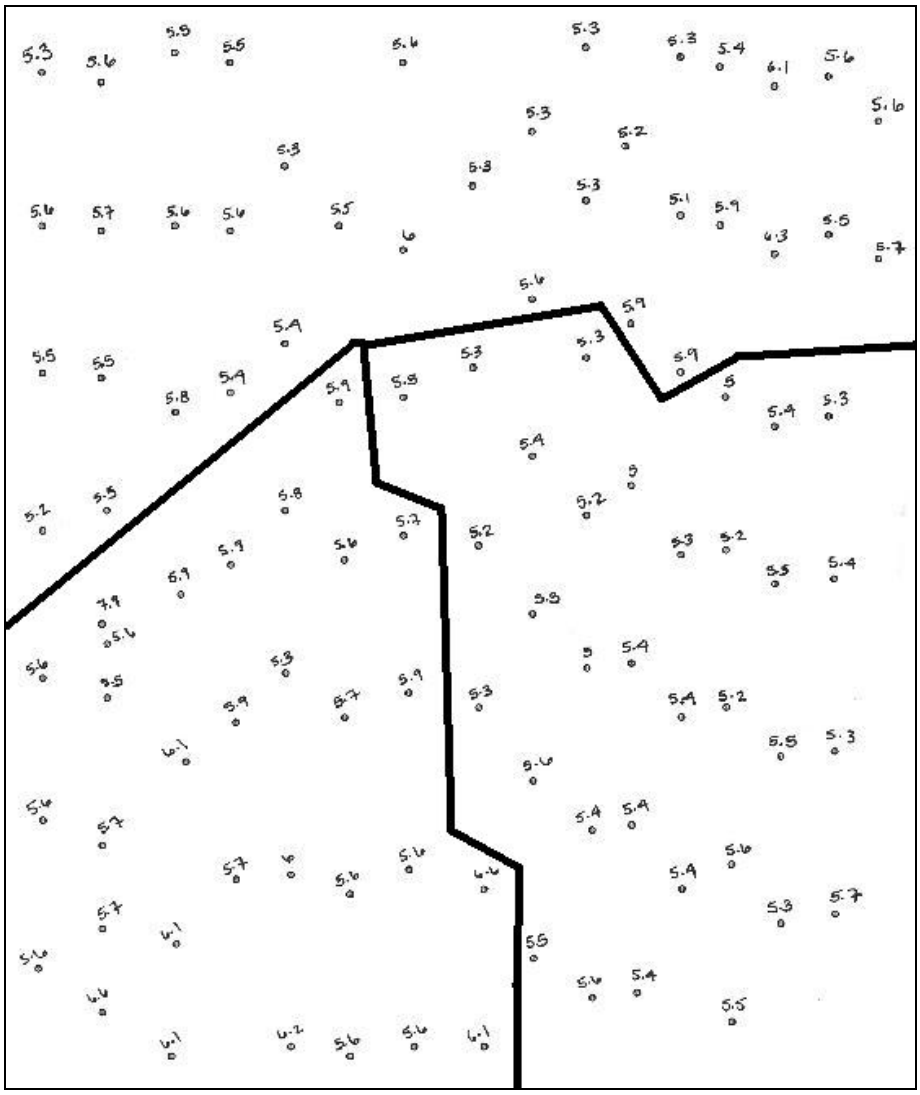

Figure 2.11: Three cluster scheme variation 3

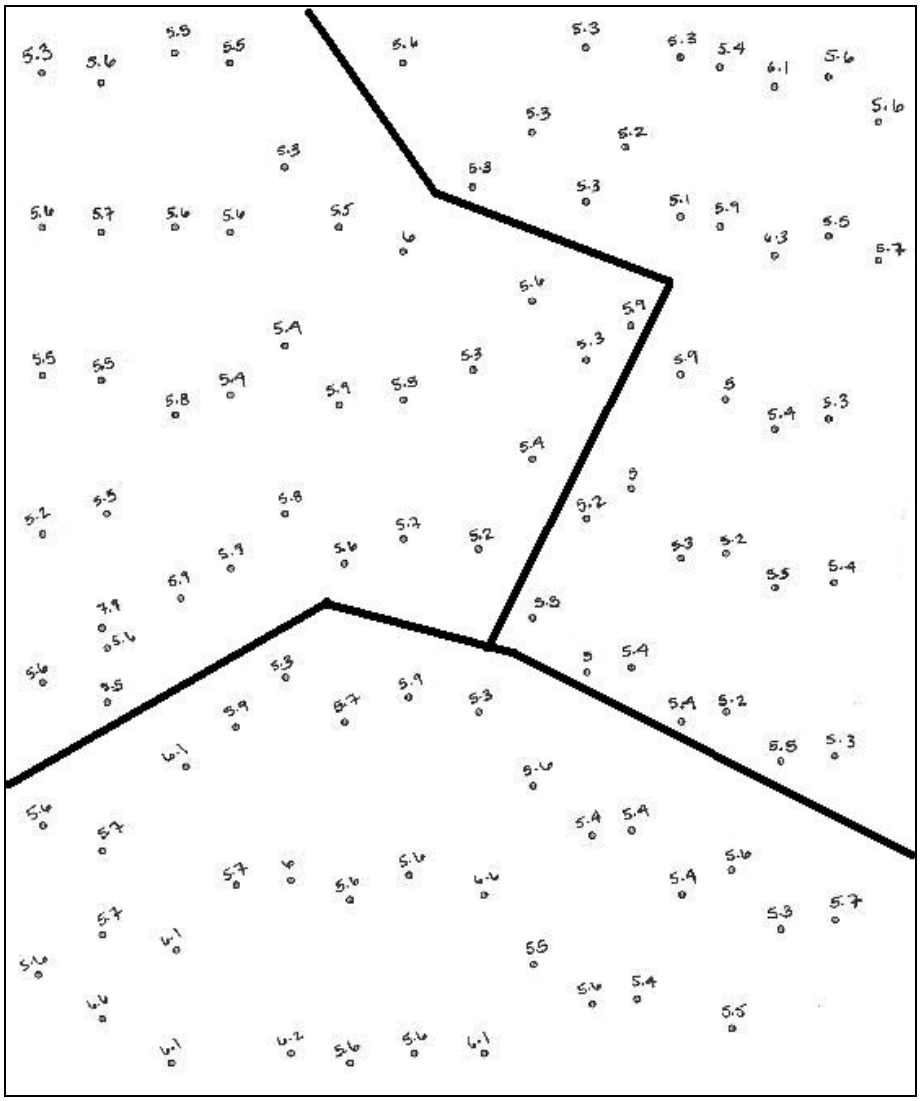

Figure 2.12: Three cluster scheme variation 4

As seen in Figures 2.9 – 2.12 there were enough observations present in each cluster to adequately estimate the spatial parameters and the mean for each cluster. The MLEs were used in the likelihood calculations. The different three cluster variations for the

data were compared and the best was chosen based on the likelihood, as well as the AIC summarized in Table 2.2.

Variation	Log-Likelihood	AIC
1	-12.63	43.26
2	-26.41	70.83
3	**-11.09**	**40.17**
4	-37.85	93.70

Table 2.2: Kansas field study three cluster results

The results show that the variation with the largest likelihood, as well as the smallest AIC, is variation 3. Therefore, the three cluster scheme in Figure 2.11 was the best for the given data.

Figures 2.13, 2.14, 2.15, and 2.16 show the four cluster schemes provided by the experts.

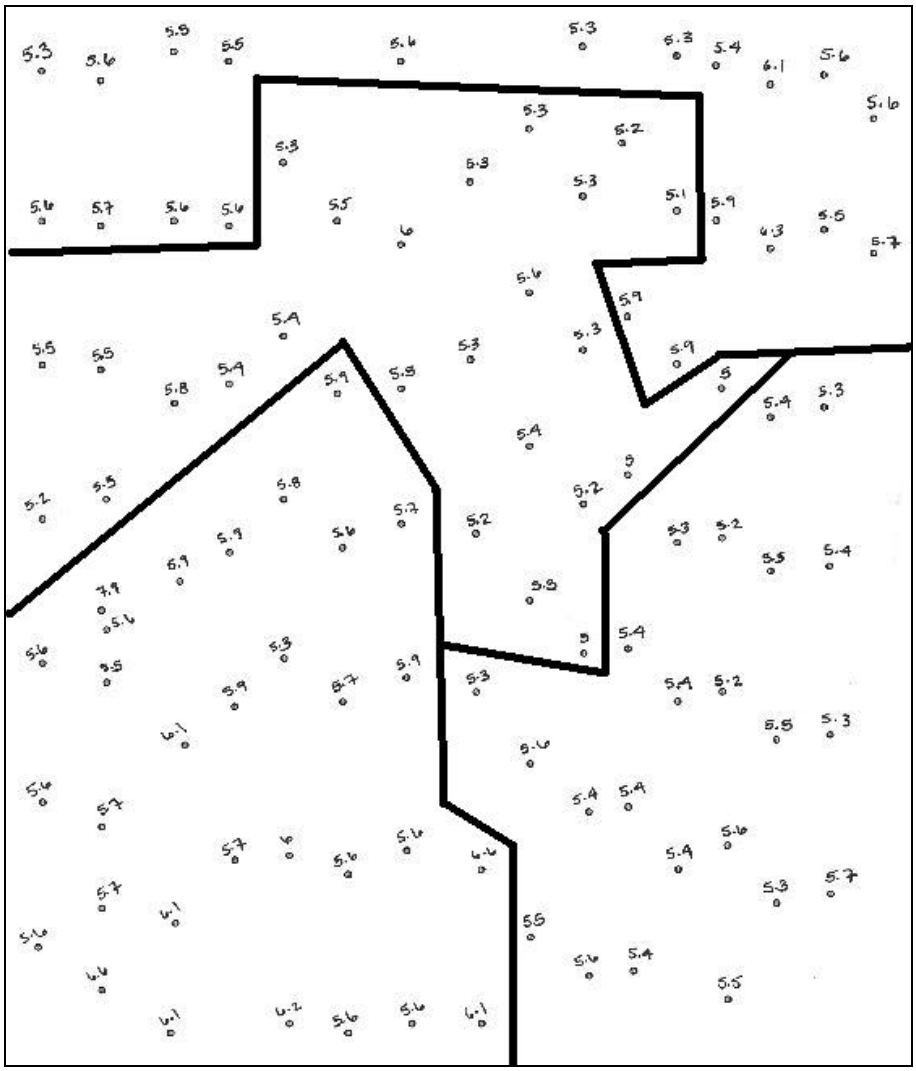

Figure 2.13: Four cluster scheme variation 1

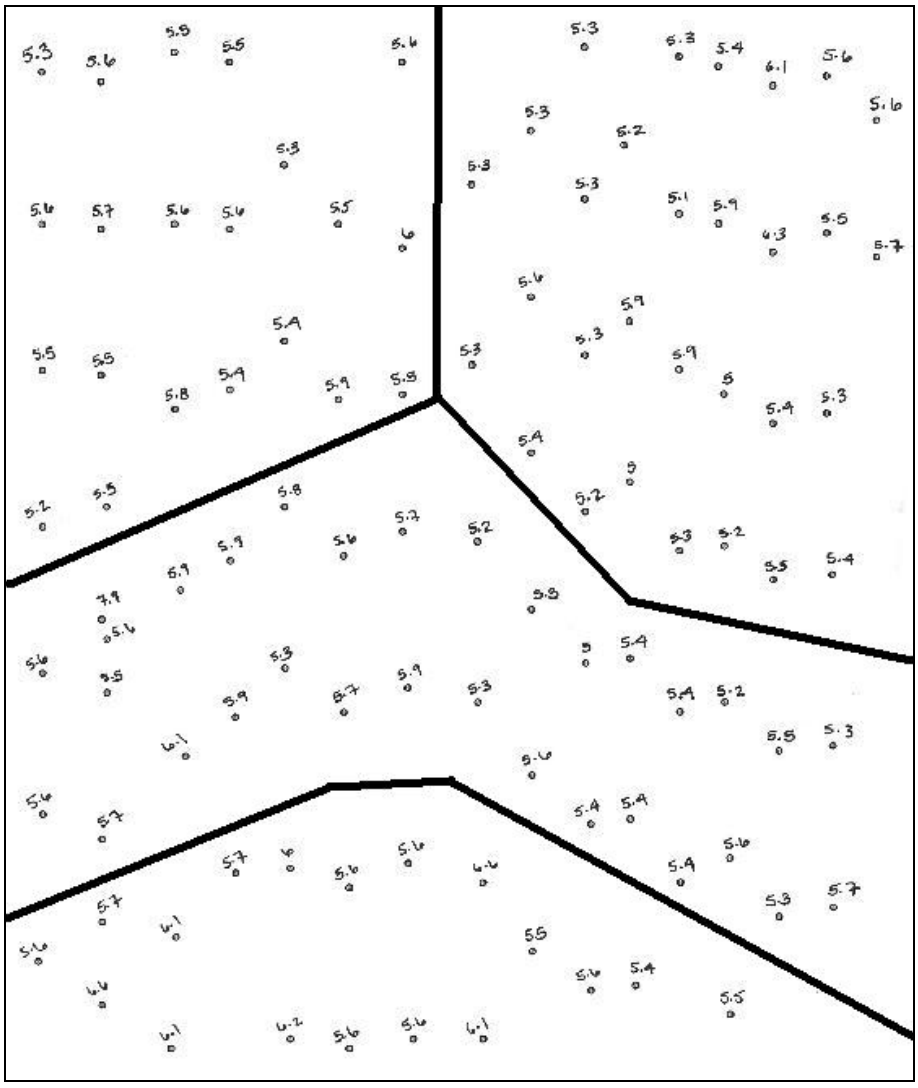

Figure 2.14: Four cluster scheme variation 2

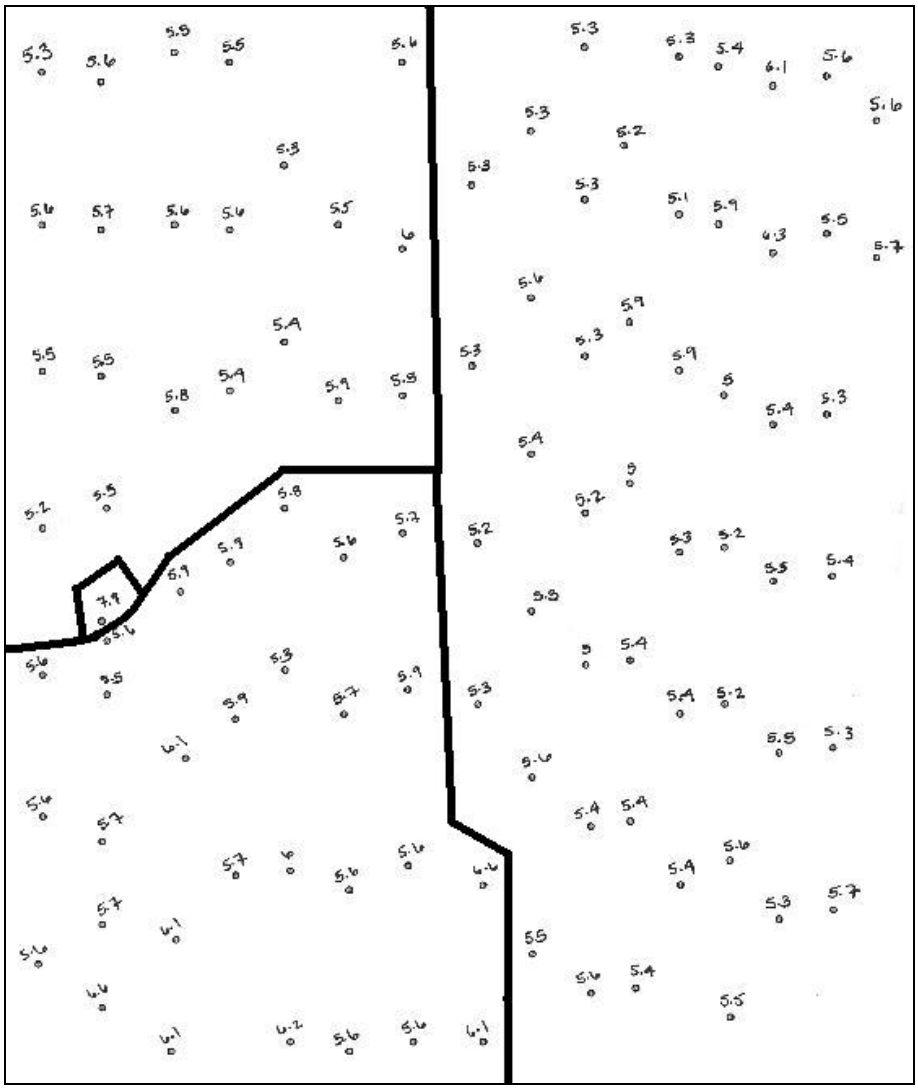

Figure 2.15: Four cluster scheme variation 3

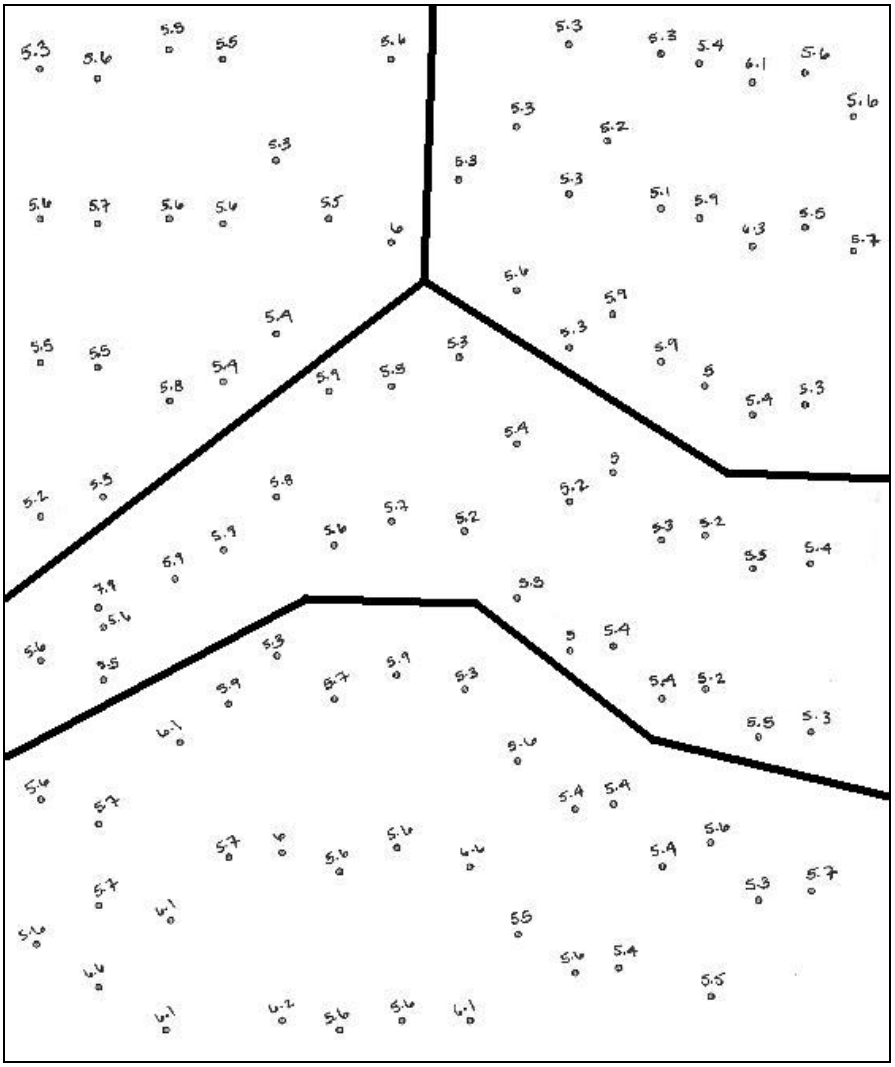

Figure 2.16: Four cluster scheme variation 4

The clustering schemes in Figures 2.13, 2.14 and 2.16 provided enough observations in each cluster to estimate the MLEs using the data alone. However, only three of the four clusters shown in Figure 2.15 had enough observations to estimate the mean and spatial parameters using the data in the clusters. For these clusters, the MLEs were calculated using the data in each cluster. The remaining cluster in this scheme had just one

observation. Therefore, only the mean of that cluster could be estimated using the data. The MLEs for the spatial parameters were calculated using the entire data (Figure 2.8). Table 2.3 summarizes the likelihoods and AICs from the four cluster analysis.

Variation	Log-Likelihood	AIC
1	**-4.10**	**32.21**
2	-30.18	84.36
3	-5.52	35.04
4	-32.21	88.43

Table 2.3: Kansas field study four cluster results

The results show that the variation with the largest likelihood, as well as the smallest AIC, is variation 1. Therefore, the four cluster scheme in Figure 2.13 was the best for the given data.

Finally, the best three cluster (Variation 3) and four cluster (Variation 1) schemes were compared to see which best suited the data overall. When determining whether three or four clusters would be more appropriate for the data, it appeared that the four cluster scheme was better. The likelihood computed with four clusters (1.65×10^{-2}) was larger than the likelihood for three clusters (1.53×10^{-5}). Also, the AIC was smaller for four clusters; 32.21 compared to 40.17.

2.6 Weighting the Spatial Component

Since one of the goals of precision agriculture is to define areas with potential for differentiated treatments, targeted (guided) samples are taken to provide detailed information about the agronomic properties of the land. Due to the high cost of obtaining these samples, only a limited number of them may be taken and should come from relatively homogenous spatially contiguous areas of the field. Also, small patches of

similar observations may not be suitable for the application of lime, fertilizers or other agriculture inputs. Therefore, the clusters should be formed to produce the most spatially contiguous clustering of the data as possible. Thus, it would be more beneficial to include small patches of dissimilar values into the surrounding larger clusters of similar value to minimize the cost associated with site-specific management. Since the differences in the response values and the cluster means are squared in the likelihood calculation, the spatial location of the observations does not have a strong effect on the likelihood function. Therefore, weighting the purely spatial component of the likelihood $\frac{1}{|\mathbf{\Sigma}|^{1/2}}$ will increase the spatial information used in the clustering process to produce more spatially contiguous clusters for management purposes. Thus, the multivariate normal distribution will become

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{Nv/2}|\mathbf{\Sigma}|^{W/2}} e^{-1/2(\mathbf{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \qquad (2.12)$$

where W is the weighting factor.

One might suspect that as the difference in the responses between the observations in the clusters increases, the weight needed to ensure spatial location emphasis would also increase. Therefore, five differences in the responses were investigated: 1, 3, 5, 7, and 10. The data are shown in Figures 2.17, 2.18, 2.19, 2.20, and 2.21.

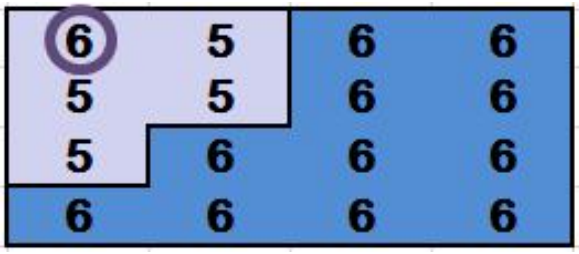

Figure 2.17: Difference in response variable of 1

Figure 2.18: Difference in response variable of 3

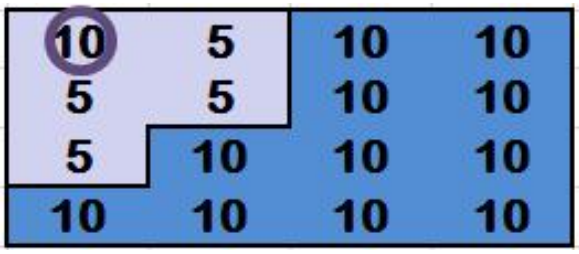

Figure 2.19: Difference in response variable of 5

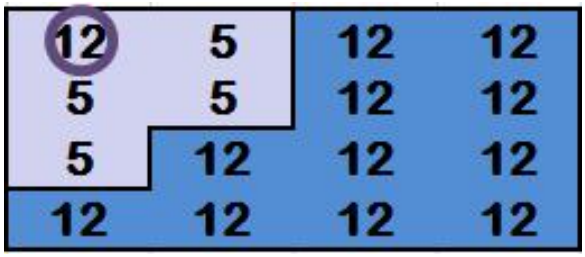

Figure 2.20: Difference in response variable of 7

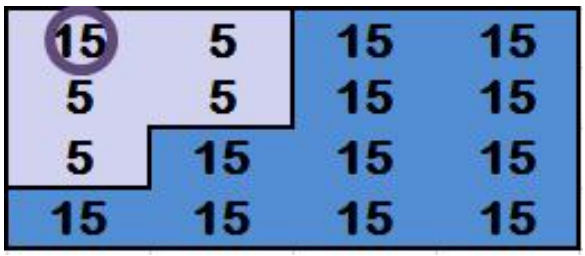

Figure 2.21: Difference in response variable of 10

Notice that the values in Cluster 1 (light blue) are all the same except for observation 1 (circled in the top left corner) which would appear to belong in Cluster 2 (dark blue). Clustering algorithms which do not account for the spatial location of the observations would place observation 1 with those in Cluster 2. However, when taking into account the actual location of observation 1, it should remain in Cluster 1 because that is where it is spatially contiguous to its neighbors.

To get an idea of how the spatial parameters, the range and sill, affect the behavior of the variance-covariance matrix, six range and sill combinations were investigated.

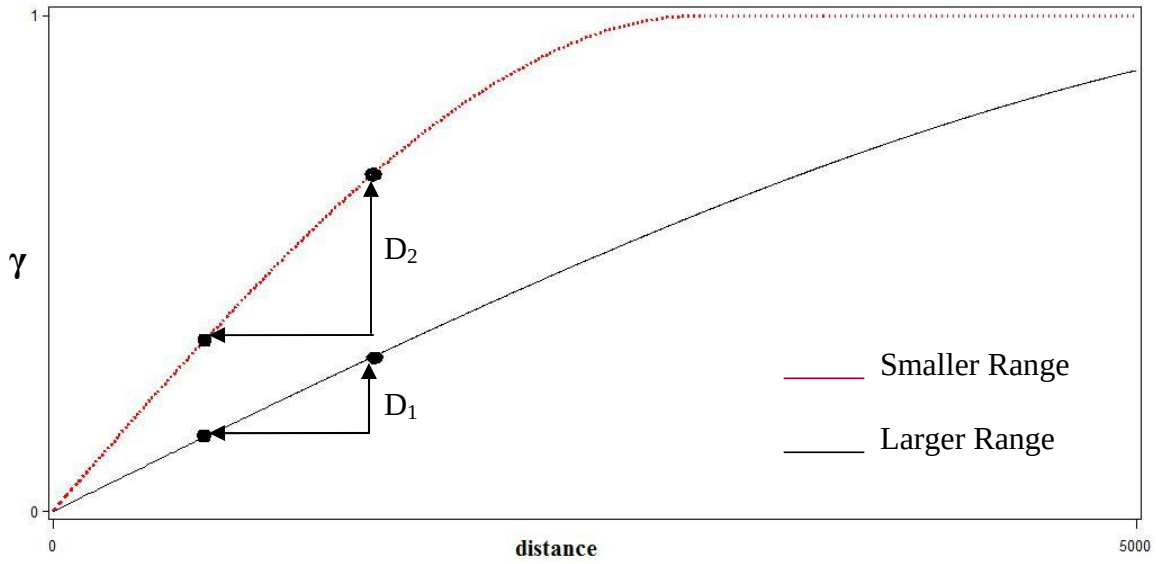

Figure 2.22: Effects of increasing the range

In Figure 2.22 the points represent the change in spatial variability between allowing an observation to remain in the cluster where it is similar in response value to moving the observation to the cluster where it is spatially contiguous, i.e. further in distance from observations with similar response values. As shown in Figure 2.22, there is less spatial variability (D_1) between the change in distance of two points which have a larger range, than the spatial variability (D_2) for a change in distance between two points which have a smaller range. Therefore, as the range increases it is harder to move an observation from the cluster where it is similar in response value to the cluster where it is spatially contiguous with the other observations.

For the above data, it should require a larger weight to keep observation 1 as a member of Cluster 1, versus letting it belong to Cluster 2 as the range increases. Holding the sill constant at a value of 1, the effects of the range were examined at values of 5, 10

and 15. Tables 2.4, 2.5 and 2.6 summarize the weights necessary to keep observation 1 in Cluster 1 where it is spatially contiguous where it would be most beneficial for management practices.

Response Difference	Weight
1	2.07
3	18.64
5	51.79
7	101.50
10	207.15

Table 2.4: Weighting results for a range = 5 & sill = 1

Response Difference	Weight
1	4.06
3	36.56
5	101.57
7	199.07
10	406.27

Table 2.5: Weighting results for a range = 10 & sill = 1

Response Difference	Weight
1	5.93
3	53.39
5	148.30
7	290.67
10	593.21

Table 2.6: Weighting results for a range = 15 & sill = 1

As expected, as the response difference increased, the weighting needed to maintain observation 1 was a member of Cluster 1 increased in each range and sill combination. Similarly, as the range increased the weight needed to ensure observation 1 was a

member of Cluster 1 also increased. Thus, if a large difference in the responses is present in the data, a larger weighting will be needed to sufficiently incorporate spatial location into the analysis. Also, as the spatial range of the data increases, larger weights are needed to ensure that spatial location plays a key role in the clustering process.

One could anticipate that as the sill increases the spatial component of the likelihood would get stronger, resulting in a lower weight needed to sufficiently incorporate the spatial location into the clustering process. Such a relationship is shown in Figure 2.23.

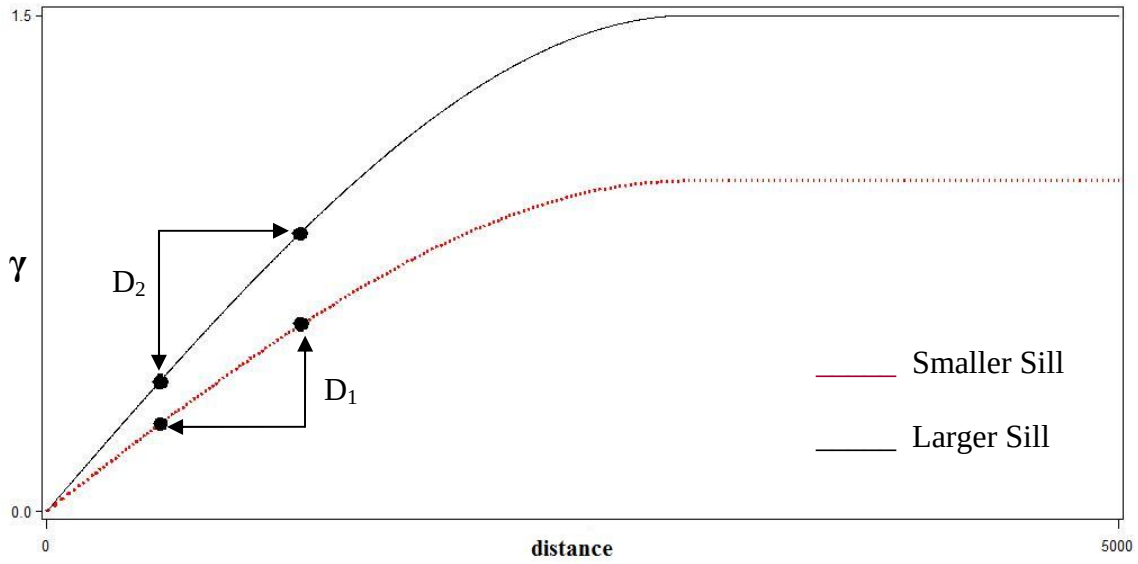

Figure 2.23: Effects of increasing the sill

In Figure 2.23 the points again represent the change in spatial variability between allowing an observation to remain in the cluster where it is similar in response value to moving the observation to the cluster where it is spatially contiguous, i.e. further in distance from observations with similar response values. As shown in Figure 2.23, there

is less spatial variability (D_1) between the change in distance of two points which have a smaller sill, than the spatial variability (D_2) for a change in distance between two points which have a larger sill. Therefore, as the sill increases it is easier to move an observation from the cluster where it is similar in response value to the cluster where it is spatially contiguous with the other observations.

For the above data, it should require a smaller weight to keep observation 1 in Cluster 1, versus letting it belong to Cluster 2 as the sill increases. Holding the range at a constant value of 5, the effect of the sill was examined at 5, 10 and 15 to see what outcome it has on the clustering process. Tables 2.7, 2.8 and 2.9 summarize the weights necessary to keep observation 1 in Cluster 1 where it is spatially contiguous which is most beneficial for management practices.

Response Difference	Weight
1	0.41
3	3.73
5	10.36
7	20.30
10	41.43

Table 2.7: Weighting results for a range = 5 & sill = 5

Response Difference	Weight
1	0.21
3	1.86
5	5.18
7	10.15
10	20.71

Table 2.8: Weighting results for a range = 5 & sill = 10

Response Difference	Weight
1	0.14
3	1.24
5	3.45
7	6.77
10	13.81

Table 2.9: Weighting results for a range = 5 & sill = 15

As can be seen in Tables 2.7 – 2.9, as the difference in the responses increased, the weighting needed to keep observation 1 in Cluster 1 also increased. This was to be expected as was seen previously when examining the effects of the range. As the sill increased, the weights needed to keep observation 1 as a member of Cluster 1 decreased. Specifically, when the difference in response is 3 and the sill value is 5, a weight of 3.73 was needed to ensure observation 1 remained a member of Cluster 1. However, when the sill was 15, a weight of only 1.24 was required to keep observation 1 in Cluster 1. Therefore, weighting the spatial component of the multivariate normal distribution will enhance the spatial clustering process.

2.7 Conclusions

The actual geographic location of an observation can be incorporated into the variance-covariance matrix of the multivariate normal distribution used in the clustering algorithm. The variance-covariance matrix can be computed using any covariance function and spherical was chosen for this research. In addition to the clustering algorithm itself, the likelihood can also be used to evaluate clustering schemes created from expert opinions.

Since the clustering algorithm is specifically incorporating the geographical location of the observations, it should be emphasized during the analysis. This chapter also looked at weighting the spatial component of the multivariate normal distribution to incorporate a larger spatial component in the clustering algorithm.

This chapter showed how to determine which clustering variation is more appropriate based on the likelihood and AIC, while taking into account the spatial distribution of the observations. Other information criteria could be explored, including Schwartz's Bayesian Information Criterion (SBC) which provides a larger penalty for more clusters (Schwarz 1978).

2.8 References

Adamchuk, V.I., D.B. Marx, A.T. Kerby, A.K. Samal, L.K. Soh, R.B. Ferguson, and C.S. Wortmann. 2007. Guided soil sampling for enhanced analysis of georefernced sensor-based data. In: Proceedings of the Ninth International Conference on Geocomputation 2007 Conference, Maynooth, Ireland, 3-5 September 2007, U. Demsar, ed. Maynooth, Ireland: NCG - National University of Ireland (E-proceedings, 4 pages).

Akaike, H. 1974 A New Look at the Statistical Model Identification. *IEEE Transaction on Automatic Control*, AC 19, 716-723.

Cressie, N. 1991 *Spatial Statistics for Spatial Data*. New York: John Wiley & Sons, Inc.

Everitt, B. 1974 *Cluster Analysis*. Toronto: Heinemann Educational Books Ltd.

Frogbrook, Z. L. & Oliver, M. A. 2007 Identifying management zones in agricultural fields using spatially constrained classification of soil and ancillary data. *Soil Use and Management*. 23, 40 – 51.

Gower, J. C. 1971 A General Coefficient of Similarity and Some of its Properties. *Biometric*. Vol. 27, No. 4, 857 - 871

Hartigan, J. A. 1975 *Clustering Algorithms*. New York: John Wiley & Sons, Inc.

Isaaks, E. H. & Srivastava, R. M. 1989 *An Introduction to Applied Geostatistics*. New York: Oxford University Press, Inc.

Johnson, D. E. 1998 *Applied Multivariate Methods for Data Analysis*. Pacific Grove: Brooks/Cole Publishing Company.

Johnson, R. A. & Wichern, D. W. 2002 *Applied Multivariate Statistical Analysis*. Upper Saddle River: Prentice-Hall, Inc.

Kaufman, L. & Rousseeuw, P. J. 1990 *Finding Groups in Data An Introduction to Cluster Analysis*. New York: John Wiley & Sons, Inc.

Oliver, M. A. & Webster, R. 1989 A Geostatistical Basis for Spatial Weighting in Multivariate Classification. *Mathematical Geology*, 21, 1, 15-35.

Schabenberger, O. & Gotway, C. A. 2005 *Spatial Methods for Spatial Data Analysis*. New York: Chapman & Hall/CRC Press.

Schwarz, G. 1978 Estimating Dimensions of a Model. *Annals of Statistics*, 6, 461-464.

Chapter 3

Spatial Clustering in the Multivariate Case

3.1 Introduction

Cluster analysis is used as a tool to place similar multivariate observations into groups or clusters. These clusters are based on measures of similarity or dissimilarity so that the similarity of observations within a cluster is maximized and the dissimilarity with observations in other clusters is also maximized. These measures of similarity (or dissimilarity) are usually based upon the Euclidean, standardized Euclidean, or Mahalanobis distance calculations. Once calculated, these distances can be used in a variety of hierarchical or nonhierarchical clustering methods (Everitt 1974, Hartigan 1975, Johnson 1998, Johnson & Wichern 2002, Kaufman & Rousseeuw).

Most of the current clustering methods do not allow one to account for the underlying spatial structure of the data. However, there are cases for which the spatial location is both known (e.g. encoded as latitude and longitude) and relevant to the goals of the analysis. One example is site-specific crop management, which has become an important aspect of agriculture production in recent years. Precision agriculture methods use multiple data layers within spatially variable observations to fine-tune crop management decisions. Since conventional coarse (approximately 1-ha) grid sampling fails to provide adequate representation of spatial variability in soils, alternative high-density sensor data have been used in many operations.

One of the major challenges in the data analysis process is to delineate field areas with potential for differentiated treatments that are frequently called “management zones.” Initially, a relatively inexpensive set of data such as on-the-go soil sensing maps and/or remote sensing imagery are collected. These data are very dense and can be used to define areas for targeted (guided) sampling which will provide detailed information about the agronomic quality of land through the analysis of soil samples run in a commercial lab. Since only a limited number of these costly samples can be afforded, they should come from homogenous areas of the field, away from boundaries or locations where sensor data changes significantly over short distances, and spread across the entire landscape. These samples should also uniformly cover the entire range of measurements, indicating spots of high, medium, or low readings. Some of the measurements collected might be pH, potassium, nitrate, moisture, or sodium content (Adamchuk 2006, Adamchuk et al. 2007, Frogbrook & Oliver 2007). Certain agronomic properties could be related to a linear or other combination of multiple sensor data layers where the area of applicability of such relationships may be limited to a series of spatial clusters with relative homogeneity. Therefore, a proper clustering method should delineate relatively homogeneous field areas while accounting for the physical values of high-density observations and their spatial distribution.

Bourgault, Marcotte and Legendre (1992) proposed a method to perform spatial clustering which uses either the multivariate variogram or multivariate covariogram. In the analysis, the multivariate variogram represents the multivariate dissimilarity between observations while the multivariate covariogram represents the multivariate similarity

between the observations. If the similarities between the observations are of utmost interest, the multivariate covariogram would be used to calculate the similarities between the observations as seen in Equation (3.1):

$$S_{ij}^{*^2} = S_{ij}^2 K(\mathbf{h}). \tag{3.1}$$

In Equation (3.1) S_{ij}^2 represents the similarities between observations at the i^{th} and j^{th} locations and $K(\mathbf{h})$ is the multivariate covariogram as defined in Equation (3.2):

$$K(\mathbf{h}) = E\left[\left(\mathbf{Z}(\mathbf{x}) - \boldsymbol{\mu}\right)\mathbf{M}\left(\mathbf{Z}(\mathbf{x}+\mathbf{h}) - \boldsymbol{\mu}\right)'\right] \tag{3.2}$$

where $\mathbf{h}$ is the geographical displacement, $\mathbf{Z}(\mathbf{x})$ is a row vector of p second-order stationary random functions, $\boldsymbol{\mu} = E[\mathbf{Z}(\mathbf{x})]$, and $\mathbf{M}$ is a $p \times p$ positive definite symmetric matrix used as a metric in the calculation of the similarities (Bourgault et al. 1992). The similarities may be calculated using any of the measures mentioned above. Bourgault et al. (1992) chose to use the Mahalanobis distance calculation and a spherical spatial structure in the multivariate covariogram to compute the similarities.

If dissimilarities between observations are more important, the multivariate variogram may be used to calculate a dissimilarity value which will then be used to cluster the observations. Equation (3.3) displays the dissimilarity calculation:

$$d_{ij}^{*^2} = d_{ij}^2 G(\mathbf{h}). \tag{3.3}$$

$G(\mathbf{h})$ represents the multivariate variogram

$$2G(\mathbf{h}) = E\left[\left(\mathbf{Z}(\mathbf{x}) - \mathbf{Z}(\mathbf{x}+\mathbf{h})\right)\mathbf{M}\left(\mathbf{Z}(\mathbf{x}) - \mathbf{Z}(\mathbf{x}+\mathbf{h})\right)'\right] \tag{3.4}$$

and d_{ij}^2 represents the dissimilarities calculated between observations at the i^{th} and j^{th} locations. $\mathbf{Z}(\mathbf{x})$ is a row vector of p second-order stationary random functions, $\boldsymbol{\mu} = E[\mathbf{Z}(\mathbf{x})]$ and $\mathbf{M}$ is a $p \times p$ positive definite symmetric matrix used as a metric in the calculation of the dissimilarities (Bourgault et al. 1992). Equation (3.1) tends to have a stronger impact on the spatial component of the observations and tends to produce groups which are spatially homogenous (Bourgault et al. 1992).

The similarities and dissimilarities computed using Equations (3.1) and (3.3) can be used as the starting point for many hierarchical or nonhierarchical clustering algorithms. Bourgault et al. (1992) used a nonhierarchical clustering algorithm as outlined below.

Step 0: An initial partition with *k* groups is performed

Step 1: For each sample, the modified similarities (dissimilarities) (Equation (3.1) or (3.3)) are calculated with all other samples. The average similarity (dissimilarity) is computed for each of the *k* groups. The sample is assigned to the group with the smallest average dissimilarity or greatest average similarity.

Step 2: If no samples changed assignation in Step 1, then the algorithm is stopped. Otherwise, Step 1 is repeated as the next iteration. This will continue until no change occurs in group assignment or a fixed stopping point has been reached.

Bourgault et al. (1992) found that the groups formed using a measure of similarity were not drastically different from the groupings formed using a measure of dissimilarity. As

suspected, it was also found that when Equation (3.1) was used, the groups formed were more spatially homogenous than when using Equation (3.3).

In this paper a clustering method is proposed to cluster multivariate observations which explicitly takes into account the spatial structure by using the multivariate normal distribution. The spatial structure is present as part of the variance-covariance matrix and it is assumed that each variate has the same spatial structure. However, this method can be generalized to the case where the variates have different spatial structures.

3.2 Clustering Multivariate Observations Using the Likelihood Function

The clustering algorithm proposed here maximizes the likelihood of the multivariate normal distribution at every step (hierarchical clustering). To start, each observation is considered to form its own cluster, resulting in n initial clusters. The likelihood is then computed for each possible pairing of two "clusters." The pairing which produces the largest likelihood is merged to form a new cluster. After the first pairing (step 1) there are now $n-1$ clusters (one cluster will have two observations and the remaining $n-2$ clusters will each have one observation).

During step 2 all possible pairwise groupings of the $n-1$ clusters are evaluated. The pairing which produces the largest likelihood is selected as the new merged cluster. This process will continue until all the observations are placed into a single cluster.

To account for the spatial structure in the likelihood, the variance-covariance matrix is computed using any specific covariance function from which exponential,

Gaussian and spherical are the most common. The frequently used spherical covariance function is given by,

$$C(d) = \begin{cases} \sigma^2 \left\{ 1 - \frac{3}{2}\left(\frac{d}{a}\right) + \frac{1}{2}\left(\frac{d}{a}\right)^3 \right\} & if\ d \le a \\ 0 & if\ d > a \end{cases} \tag{3.5}$$

where d is the distance between two points and a is the range of the variogram (Cressie 1991, Isaaks & Srivastava 1989, Schabenberger & Gotway 2005). The range is the separation distance at which an increase in distance no longer produces an increase in the average squared difference between pairs of values (Isaaks & Srivastava 1989). The Gaussian covariance function which works well with a small scale spatial structure is

$$C(d) = \sigma^2 e^{-\frac{3d^2}{a^2}}, \tag{3.6}$$

and the exponential covariance function is

$$C(d) = \sigma^2 e^{-\frac{3d}{a}} \tag{3.7}$$

which works best when there is less spatial structure at small distances. The Gaussian and exponential covariance functions have a similar range a, but they are not strictly identical, as the range refers to the rate at which the covariance function approaches the sill (Cressie 1991, Isaaks & Srivastava 1989, Schabenberger & Gotway 2005). Figure 3.1 compares these covariance functions.

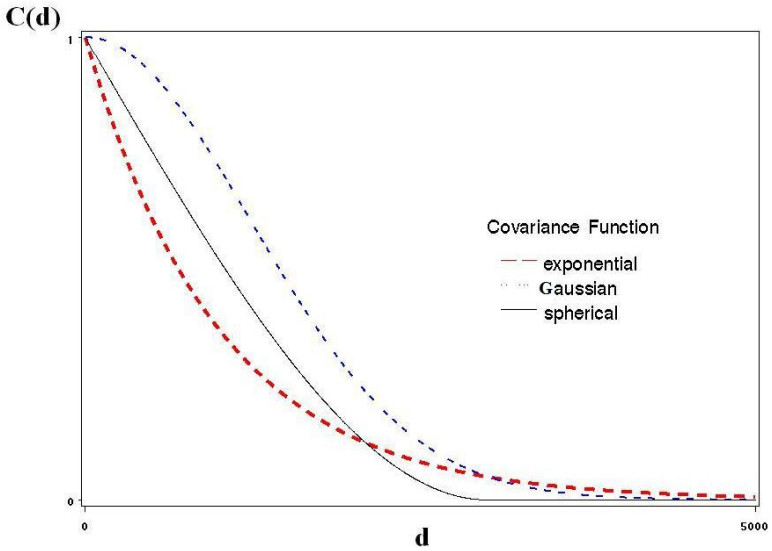

Figure 3.1: Comparison of covariance functions

The value of the variogram for a distance of zero is zero, however, due to sampling error and scale variability the values recorded at extremely small distances may be rather dissimilar causing discontinuity at the origin. The vertical jump from zero to these values is the nugget effect (Isaaks & Srivastava 1989), which must also be considered in spatial analyses. Since the spherical covariance function is most common, the examples provided in this paper use the spherical covariance function and assume there is no nugget effect.

The likelihood of the multivariate normal distribution can be written as

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{Nv/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-1/2(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \tag{3.8}$$

where v is the number of clustering variates and $N = n_1 + n_2 + \ldots + n_c$, the sum of the number of observations which fall into each cluster, where c is the number of clusters.

$$\mathbf{x}' = \left(x_{111} \quad \cdots \quad x_{11n_1} \quad x_{121} \quad \cdots \quad x_{12n_1} \quad \cdots \quad x_{211} \quad \cdots \quad x_{21n_2} \quad x_{221} \quad \cdots \quad x_{cvn_c} \right)$$

where x_{ijk} is the variate value of the k^{th} observation for the j^{th} variate in the i^{th} cluster

$i = 1, \ldots, c$ where c is the number of clusters

$j = 1, \ldots, v$ where v is the number of variates observed

$k = 1, \ldots, n_i$ where n_i is the total number of observations in the i^{th} cluster

$$\boldsymbol{\mu}' = \left(\mu_{11} \quad \cdots \quad \mu_{11} \quad \mu_{12} \quad \cdots \quad \mu_{12} \quad \cdots \quad \mu_{21} \quad \cdots \quad \mu_{21} \quad \mu_{22} \quad \cdots \quad \mu_{cv} \right)$$

where μ_{ij} is the mean for each cluster variate combination - there are n_i μ_{ij}'s in the i^{th} cluster of the j^{th} variate

The variance-covariance matrix in Equation (3.8) is given by $\boldsymbol{\Sigma} = \bigoplus_{i=1}^{c} \boldsymbol{\Sigma}_i^*$ which assumes that each variable is uncorrelated with itself between clusters. $\boldsymbol{\Sigma}_i^*$ is the cross-covariance matrix between variates and is computed using the spherical covariance function from Equation (3.5):

$$\boldsymbol{\Sigma}_{\mathbf{i}}^* = \begin{bmatrix} \boldsymbol{\Sigma}_{i11} & \boldsymbol{\Sigma}_{i12} & \cdots & \boldsymbol{\Sigma}_{i1v} \\ \boldsymbol{\Sigma}_{i21} & \boldsymbol{\Sigma}_{i22} & \cdots & \boldsymbol{\Sigma}_{i2v} \\ \vdots & & \ddots & \vdots \\ \boldsymbol{\Sigma}_{iv1} & \boldsymbol{\Sigma}_{iv2} & \cdots & \boldsymbol{\Sigma}_{ivv} \end{bmatrix} = \left[\boldsymbol{\Sigma}_{ijj'} \right]. \tag{3.9}$$

When $j = j'$ the cross-covariance matrix $\boldsymbol{\Sigma}_{ijj'}$ in Equation (3.9) is of the form

$$\boldsymbol{\Sigma}_{ijj} = \begin{bmatrix} \sigma^2_{ijj} & sph\left(\sigma^2_{ij}, a_{ij}, d_{i12}\right) & \cdots & sph\left(\sigma^2_{ij}, a_{ij}, d_{i1n_i}\right) \\ & \sigma^2_{ijj} & \cdots & sph\left(\sigma^2_{ij}, a_{ij}, d_{i2n_i}\right) \\ & & \ddots & \vdots \\ & & & \sigma^2_{ijj} \end{bmatrix}. \tag{3.10}$$

This matrix is symmetric because $d_{ikk'}$ is the actual physical distance between observations at locations k and k', so sph($\sigma^2_{ij}, a_{ij}, d_{i12}$) = sph($\sigma^2_{ij}, a_{ij}, d_{i21}$) (Isaaks & Srivastava 1989).

When $j \neq j'$ the cross-covariance matrix $\boldsymbol{\Sigma}_{ijj'}$ must account for the correlation between variates j and j'. In addition to the correlation of the variates, the variability from each variate, j and j', must also be taken into account. Oliver (2003) proposed a method to find the cross-covariance between variates. First, the variance-covariance matrix for each variate must be calculated, that is $\boldsymbol{\Sigma}_{\mathbf{ijj}}$ and $\boldsymbol{\Sigma}_{\mathbf{ij'j'}}$. Once $\boldsymbol{\Sigma}_{\mathbf{ijj}}$ and $\boldsymbol{\Sigma}_{\mathbf{ij'j'}}$ have been calculated a Cholesky decomposition is performed so the variance-covariance matrices can be defined as in Equations (3.11) and (3.12):

$$\boldsymbol{\Sigma}_{\mathbf{ijj}} = \mathbf{L}_{\mathbf{ij}}\mathbf{L}'_{\mathbf{ij}} \tag{3.11}$$

$$\boldsymbol{\Sigma}_{\mathbf{ij'j'}} = \mathbf{L}_{\mathbf{ij'}}\mathbf{L}'_{\mathbf{ij'}} \quad . \tag{3.12}$$

After the variance-covariance matrices have been decomposed $\mathbf{L}_{\mathbf{ij}}$ and $\mathbf{L}'_{\mathbf{ij'}}$ will be used to compute the cross-covariance matrix, $\boldsymbol{\Sigma}_{\mathbf{ijj'}}$, between variates j and j'. Equation (3.13) shows how the cross-covariance matrix $\boldsymbol{\Sigma}_{ijj'}$ incorporates the variability from variates j and j', as well as the correlation:

$$\mathbf{\Sigma_{ijj'}} = \rho \mathbf{L_{ij}} \mathbf{L_{ij'}}' . \tag{3.13}$$

For example, suppose two variates are of interest. Then $\mathbf{\Sigma}_i^*$ will be the 2×2 matrix shown in Equation (3.14):

$$\mathbf{\Sigma_i^*} = \begin{bmatrix} \mathbf{\Sigma_{i11}} & \mathbf{\Sigma_{i12}} \\ \mathbf{\Sigma_{i21}} & \mathbf{\Sigma_{i22}} \end{bmatrix} = \begin{bmatrix} \mathbf{L_{i1}} \mathbf{L_{i1}}' & \rho \mathbf{L_{i1}} \mathbf{L_{i2}}' \\ \rho \mathbf{L_{i2}} \mathbf{L_{i1}}' & \mathbf{L_{i2}} \mathbf{L_{i2}}' \end{bmatrix}. \tag{3.14}$$

Since $\mathbf{\Sigma_{ijj'}}$ is comprised of the vectors from the decomposed variance matrices for each variate $\left(\mathbf{L_{ij}}, \mathbf{L_{ij'}}\right)$, $\mathbf{\Sigma_{ijj}}$ contains the sill values from each variate, σ_{ij}^2 and $\sigma_{ij'}^2$, as well as the range values, a_{ij} and $a_{ij'}$. To ensure that $\mathbf{\Sigma_{ijj'}}$ is positive definite the sill of $\mathbf{\Sigma_{ijj'}}$ can be no larger than $\sqrt{\sigma_{ij}^2 \sigma_{ij'}^2}$ and the range can be no larger than $\sqrt{a_{ij} a_{ij'}}$. To simplify the cross-covariance function it is assumed that $a_{ij} = a_{ij'}$ in this research. If $a_{ij} \neq a_{ij'}$ the spherical cross-covariance function changes depending on the relationship between the ranges.

3.3 Choosing an Optimal Number of Clusters

The likelihood function can be used to determine the optimal clustering scheme for a given set of data. A sharp increase in the plot of the likelihood against the number of clusters indicates an appropriate number of clusters. Since the likelihood is maximized at every step in the clustering process, an increase in the plot shows what clustering scheme(s) may be best.

An improvement over plotting the likelihood against the number of clusters is to use Akaike's Information Criterion (AIC) (Akaike 1974). This criterion also uses the

likelihood computed using a covariance function, while penalizing for the number of estimated parameters. Since the ultimate goal is to maximize the likelihood, the parameter estimates are computed using maximum likelihood estimation (MLEs). The AIC is given by,

$$\text{AIC} = -2\log\left\{L\left(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}} \mid \mathbf{x}\right)\right\} + 2k \tag{3.15}$$

where k is the number of parameters estimated and $L(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}} \mid \mathbf{x})$ is the estimated likelihood given the data. Therefore, a penalty will be imposed for having more clusters, i.e. more parameters, to estimate. Thus, smaller AIC values are better.

Within each cluster the range, sill and mean (assuming no nugget effect) must be estimated for each variate. Therefore, 3ν parameters must be estimated where ν is the number of variates. In addition to those parameters, the correlation between each pair of variates must also be estimated, resulting in $\binom{\nu}{2}$ additional parameters for each cluster. Thus, there are a total of $3\nu + \binom{\nu}{2}$ parameters to estimate for each cluster. It would seem that if there were at least $2*\left(3\nu + \binom{\nu}{2}\right)$ observations in a cluster the parameter estimates computed using the current data configuration would seem reasonable for the analysis. Therefore, when the number of observations in a cluster is smaller than $2*\left(3\nu + \binom{\nu}{2}\right)$ the MLEs for the spatial parameters (range and sill) are derived using the entire data set, but the MLE for the mean is still calculated from the data in each cluster.

When both large and small clusters are present in the data, the MLEs for the large clusters will be estimated using the data in the cluster and for the small clusters the entire data set is used to estimate the MLEs for the spatial parameters and the data in the cluster are used to estimate the MLE for the mean.

The AIC is used as one of our deciding factors to determine the appropriate number of clusters for the data. A penalization for having a large number of clusters is important and is not taken into account when looking solely at the likelihood. Thus, both the likelihood and AIC values are given in the example, so both may be used in the decision making process.

3.4 Example: Simulated Data

The data for the example were simulated to have two variates with a correlation of 0.50. The first variate has a sill of 1 and a range of 5, while the second variate has a sill of 5 and a range of 5 and neither variate had a nugget effect. Without loss of generality, a nugget effect may be added, but the results here are simply more dramatic without a nugget effect. A 10×10 grid of data was generated in order to ensure a strong spatial floor and the center 6×6 grid was extracted for the analysis. The smallest number of clusters possible is when all the observations fall into just one cluster, and the largest number of possible clusters occurs when each observation is its own cluster. Therefore, the largest number of clusters for the data was 36.

Once the data were generated, random values from a normal distribution, with a mean of 50 for variate 1 and a mean of 70 for variate 2, were added to the middle

diagonal of values (minus the three observations in the bottom left corner) to create a cluster. Similarly, random values from a normal distribution, with a mean of 15 for variate 1 and a mean of 25 for variate 2, were added to the top left and bottom right corners of the data grid to create a second cluster in the data. Finally, three random values from a normal distribution, with a mean of 55 for variate 1 and a mean of 75 for variate 2, were added to create a third cluster in the bottom left corner of the data grid. The final data values are shown in Figure 3.2 representing the smallest possible clustering of the data, i.e. when all the points fall into one cluster.

23.40	21.90	25.33	59.91	60.71	59.79
21.46	24.14	61.00	61.19	61.55	58.54
22.93	58.93	57.89	57.86	60.54	61.68
62.53	59.29	58.39	60.95	60.27	22.35
65.39	59.80	60.17	58.85	25.30	22.31
69.91	61.95	61.20	25.96	23.39	23.39

(a) Variate 1

36.43	33.69	38.49	85.59	85.74	85.84
36.98	38.26	85.82	87.55	87.82	86.00
37.85	83.47	83.15	85.23	88.82	89.85
85.91	83.28	84.09	87.17	87.21	39.28
89.96	85.06	85.08	82.36	39.93	38.77
94.68	86.25	83.38	37.87	36.53	38.44

(b) Variate 2

Figure 3.2: Data values in one cluster scheme

Since the data values along the middle diagonal of the data grid were much larger than the values in the top left and bottom right corners of the grid for both variates, the data values were separated into different clusters resulting in the two cluster scheme in Figure 3.3.

23.40	21.90	25.33	59.91	60.71	59.79
21.46	24.14	61.00	61.19	61.55	58.54
22.93	58.93	57.89	57.86	60.54	61.68
62.53	59.29	58.39	60.95	60.27	22.35
65.39	59.80	60.17	58.85	25.30	22.31
69.91	61.95	61.20	25.96	23.39	23.39

(a) Variate 1

36.43	33.69	38.49	85.59	85.74	85.84
36.98	38.26	85.82	87.55	87.82	86.00
37.85	83.47	83.15	85.23	88.82	89.85
85.91	83.28	84.09	87.17	87.21	39.28
89.96	85.06	85.08	82.36	39.93	38.77
94.68	86.25	83.38	37.87	36.53	38.44

(b) Variate 2

Figure 3.3: Data values in two cluster scheme (blue and green)

However, the goal of spatial clustering is to create spatially contiguous clusters so the cluster which included the data values from both the top left and bottom right corners of the data grid were broken into two clusters producing in the three cluster scheme in Figure 3.4.

23.40	21.90	25.33	59.91	60.71	59.79
21.46	24.14	61.00	61.19	61.55	58.54
22.93	58.93	57.89	57.86	60.54	61.68
62.53	59.29	58.39	60.95	60.27	22.35
65.39	59.80	60.17	58.85	25.30	22.31
69.91	61.95	61.20	25.96	23.39	23.39

(a) Variate 1

36.43	33.69	38.49	85.59	85.74	85.84
36.98	38.26	85.82	87.55	87.82	86.00
37.85	83.47	83.15	85.23	88.82	89.85
85.91	83.28	84.09	87.17	87.21	39.28
89.96	85.06	85.08	82.36	39.93	38.77
94.68	86.25	83.38	37.87	36.53	38.44

(b) Variate 2

Figure 3.4: Data values in three cluster scheme (blue, green and orange)

Finally, the four cluster scheme was created by separating the three observations in the bottom left corner into their own cluster giving the four cluster scheme in Figure 3.5.

23.40	21.90	25.33	59.91	60.71	59.79
21.46	24.14	61.00	61.19	61.55	58.54
22.93	58.93	57.89	57.86	60.54	61.68
62.53	59.29	58.39	60.95	60.27	22.35
65.39	59.80	60.17	58.85	25.30	22.31
69.91	61.95	61.20	25.96	23.39	23.39

(a) Variate 1

36.43	33.69	38.49	85.59	85.74	85.84
36.98	38.26	85.82	87.55	87.82	86.00
37.85	83.47	83.15	85.23	88.82	89.85
85.91	83.28	84.09	87.17	87.21	39.28
89.96	85.06	85.08	82.36	39.93	38.77
94.68	86.25	83.38	37.87	36.53	38.44

(b) Variate 2

Figure 3.5: Data values in four cluster scheme (blue, green, orange, and purple)

Since there are two variates in this example, seven parameters need to be estimated for each cluster: range, sill and mean for each variate and the correlation between the variates $\left(3*2+\binom{2}{2}=7\right)$. Therefore, it is recommended that there are at least 14 observations in each cluster to adequately estimate the parameters. This only occurs when all the data are in one cluster (Figure 3.2). Hence, all seven parameters cannot be estimated in this analysis.

In this case it seems logical to estimate the range and sill using the entire data set. However, to simplify this analysis the range and sill estimates used in the analysis are those used in the data generation process (variate 1: sill = 1, range = 5 and variate 2: sill = 5, range = 5). Therefore, only the means and the correlation between variates are

estimated for each cluster, resulting in three estimated parameters for each. The MLE estimates for the means are used in the computation of the bivariate log-likelihood.

Initially, the estimate of the correlation between variate 1 and variate 2 within a cluster was estimated using the SAS® **PROC CORR** (SAS Institute 2008) procedure. The correlation estimate for the one cluster scheme seemed a bit disconcerting since the data were simulated to have a correlation of 0.50 and a value of 0.9969 was estimated. However, SAS® **PROC CORR** produces estimates of the Pearson correlation coefficient which assumes the observations are independent of one another. This is not the case for the data since they were simulated to have a spatial structure which makes them dependent observations. Also, the Pearson correlation coefficient assumes that $\boldsymbol{\mu}_{ij} = \boldsymbol{\mu}$ for all cluster variate combinations which is not the case since random normal values with different means were added to create clusters. Lastly, the correlation estimates were computed assuming a variance structure of $\mathbf{I}\sigma^2$ which does not incorporate the spatial component into the clustering process. The correlation estimates outputted from SAS ® **PROC CORR** can be found in Table 3.1.

The **optimize** function in R version 2.5 (R Development Team 2007) was used to find the correlation estimate which maximized the log-likelihood where the variance-covariance matrix (Equation (3.9)) was computed using the spherical covariance function. The **optimize** function utilizes the golden section search (Press et al. 1988) to find the value which optimizes (in regards to the minimum or maximum) a function with respect to the specified variable. The correlation estimate for the one cluster scheme was examined and found to be much closer to 0.50. Therefore, the **optimize** function was

used in this analysis to estimate the correlations. The correlation estimates from R can be found in Table 3.1 and the code for the R correlation estimation program can be found in the Appendix.

	Estimation Process	
Cluster Size	Pearson	Optimization
1	r = 0.9969	r = 0.5722
2	r_1 = 0.4281 r_2 = 0.8357	r_1 = 0.3912 r_2 = 0.4354
3	r_1 = 0.6572 r_2 = 0.8357 r_3 = 0.0122	r_1 = 0.4777 r_2 = 0.4354 r_3 = 0.2752
4	r_1 = 0.6572 r_2 = 0.6267 r_3 = 0.0122 r_4 = 0.9963	r_1 = 0.4777 r_2 = 0.4347 r_3 = 0.2752 r_4 = 0.5427

Table 3.1: Comparison of correlation estimates

Figures 3.6 and 3.7 show plots of the log-likelihood and AIC values for the one, two, three, and four cluster schemes analyzed.

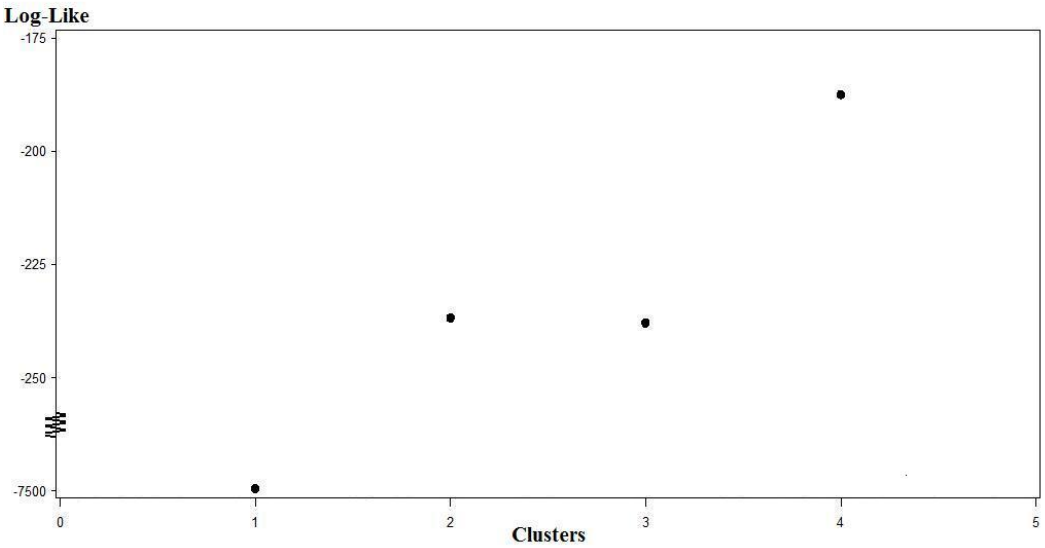

Figure 3.6: Plot of the log-likelihood values against the number of clusters

Figure 3.6 shows that the sharp increase in the plot of the log-likelihood occurs at two clusters. It appears that three clusters may have a similar log-likelihood as two clusters. However, the four cluster scheme has the largest log-likelihood. Thus, based solely on the log-likelihood it is determined that at least two clusters are appropriate for the data.

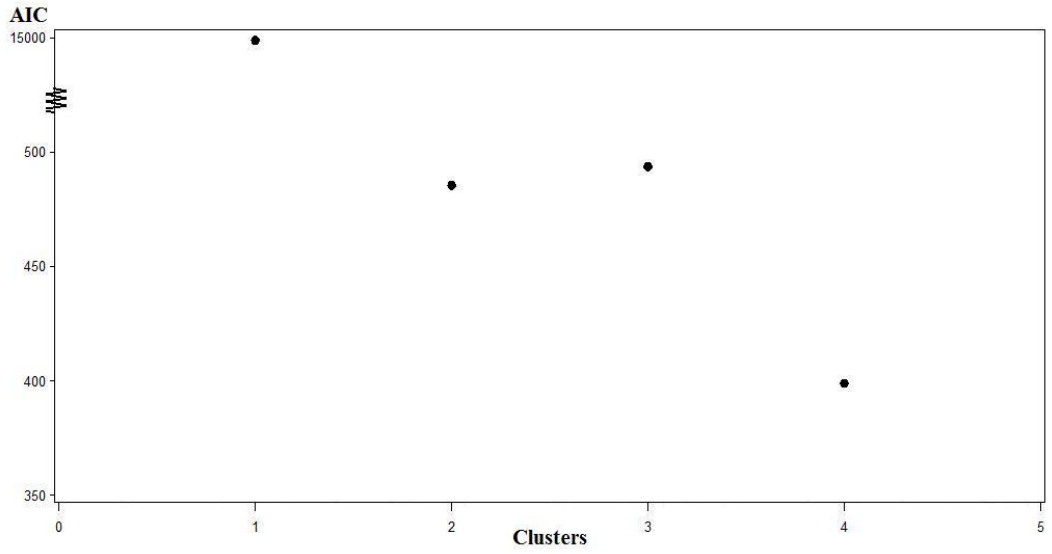

Figure 3.7: Plot of the AIC values against the number of clusters

Looking at Figure 3.7, the smallest AIC occurs when the data are grouped into four clusters. The actual log-likelihood and AIC values can be found in Table 3.2.

Number of Clusters	Log-Likelihood	AIC
1	-7339.91	14,685.81
2	-236.77	485.55
3	-237.89	493.78
4	**-187.47**	**398.94**

Table 3.2: Clustering results from simulation study

The simulation results show that when the data are grouped in a four cluster arrangement the log-likelihood value is largest. Additionally, the four cluster

arrangement gave the smallest AIC value. Thus, the four cluster arrangement of the data seems to be best from those considered.

The key to this analysis was incorporating the cross-covariance into the variance-covariance matrix for the multivariate normal distribution while taking into account the spatial location of the observations. Since the four cluster arrangement best suited the data, the variance-covariance matrix in Equation (3.8) for the arrangement would be $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_1^* & 0 & \cdots & 0 \\ 0 & \boldsymbol{\Sigma}_2^* & \ddots & \vdots \\ \vdots & \ddots & \boldsymbol{\Sigma}_3^* & 0 \\ 0 & \cdots & 0 & \boldsymbol{\Sigma}_4^* \end{bmatrix}$ where $\boldsymbol{\Sigma}_i^* = \begin{bmatrix} \boldsymbol{\Sigma}_{i11} & \boldsymbol{\Sigma}_{i12} \\ \boldsymbol{\Sigma}_{i21} & \boldsymbol{\Sigma}_{i22} \end{bmatrix}$. As seen in Figure 3.5, cluster 1 is shown in blue, cluster 2 in green, cluster 3 in orange and cluster 4 in purple. The estimated variance-covariance matrix used in the analysis is broken down by cluster. The cross-covariance matrix for cluster one is below:

$$\boldsymbol{\Sigma}_1^* = \left[\begin{array}{cccccc|cccccc} 1 & 0.704 & 0.432 & 0.704 & 0.587 & 0.432 & 1.068 & 0.752 & 0.461 & 0.752 & 0.627 & 0.461 \\ 0.704 & 1 & 0.704 & 0.587 & 0.704 & 0.374 & 0.752 & 1.068 & 0.752 & 0.627 & 0.752 & 0.399 \\ 0.432 & 0.704 & 1 & 0.374 & 0.587 & 0.242 & 0.461 & 0.752 & 1.068 & 0.399 & 0.627 & 0.258 \\ 0.704 & 0.587 & 0.374 & 1 & 0.704 & 0.704 & 0.752 & 0.627 & 0.399 & 1.068 & 0.752 & 0.752 \\ 0.587 & 0.704 & 0.587 & 0.704 & 1 & 0.587 & 0.627 & 0.752 & 0.627 & 0.752 & 1.068 & 0.627 \\ 0.432 & 0.374 & 0.242 & 0.704 & 0.587 & 1 & 0.461 & 0.399 & 0.258 & 0.752 & 0.627 & 1.068 \\ \hline 1.068 & 0.752 & 0.461 & 0.752 & 0.627 & 0.461 & 5 & 3.520 & 2.160 & 3.520 & 2.935 & 2.160 \\ 0.752 & 1.068 & 0.752 & 0.627 & 0.752 & 0.399 & 3.520 & 5 & 3.520 & 2.935 & 3.520 & 1.870 \\ 0.461 & 0.752 & 1.068 & 0.399 & 0.627 & 0.258 & 2.160 & 3.520 & 5 & 1.870 & 2.935 & 1.210 \\ 0.752 & 0.627 & 0.399 & 1.068 & 0.752 & 0.752 & 3.520 & 2.935 & 1.870 & 5 & 3.520 & 3.520 \\ 0.627 & 0.752 & 0.627 & 0.752 & 1.068 & 0.627 & 2.935 & 3.520 & 2.935 & 3.520 & 5 & 2.935 \\ 0.461 & 0.399 & 0.258 & 0.752 & 0.627 & 1.068 & 2.160 & 1.870 & 1.210 & 3.520 & 2.935 & 5 \end{array}\right].$$

Since $\boldsymbol{\Sigma}_2^*$ is a 42×42 matrix, it is further broken down into its components $\boldsymbol{\Sigma}_{211}$ (the variance-covariance matrix for the first variate), $\boldsymbol{\Sigma}_{212} = \boldsymbol{\Sigma}_{221}$ (the cross-covariance matrix

between variate 1 and variate 2), and $\mathbf{\Sigma}_{222}$ (the variance-covariance matrix for the second variate).

Recall, the first variate has a range = 5 and a sill = 1. Therefore, the spherical covariance function for this variate would be $sph(d) = 1*\left\{1 - \frac{3}{2}\left(\frac{d}{5}\right) + \frac{1}{2}\left(\frac{d}{5}\right)^3\right\}$. Since the distance between observations (1, 4) and (1, 4) is 0, the spherical covariance function has a value of $sph(0) = 1*\left\{1 - \frac{3}{2}\left(\frac{0}{5}\right) + \frac{1}{2}\left(\frac{0}{5}\right)^3\right\} = 1$. Thus, the diagonal elements of $\mathbf{\Sigma}_{211}$ represent the spherical covariance function of each observation with itself (a distance of zero). The value in the first row and second column of $\mathbf{\Sigma}_{211}$ represents the spherical covariance function value between observations (1,4) and (1,5). The distance between these two observations is 1, resulting in a spherical covariance function value of 0.704.

$$\Sigma_{211} = \begin{bmatrix}
1 & 0.704 & 0.432 & 0.587 & 0.704 & 0.587 & 0.374 & 0.242 & 0.374 & 0.432 & 0.374 & 0.242 & 0.106 & 0.178 & 0.208 & 0.178 & 0.016 & 0.043 & 0.056 & 0 & 0 \\
0.704 & 1 & 0.704 & 0.374 & 0.587 & 0.704 & 0.587 & 0.106 & 0.242 & 0.379 & 0.432 & 0.374 & 0.033 & 0.106 & 0.178 & 0.208 & 0 & 0.016 & 0.434 & 0 & 0 \\
0.432 & 0.704 & 1 & 0.178 & 0.374 & 0.587 & 0.704 & 0.016 & 0.106 & 0.242 & 0.374 & 0.432 & 0 & 0.033 & 0.106 & 0.178 & 0 & 0 & 0.016 & 0 & 0 \\
0.587 & 0.374 & 0.178 & 1 & 0.704 & 0.432 & 0.208 & 0.587 & 0.704 & 0.587 & 0.374 & 0.178 & 0.374 & 0.432 & 0.374 & 0.242 & 0.178 & 0.208 & 0.178 & 0.043 & 0.056 \\
0.704 & 0.587 & 0.374 & 0.704 & 1 & 0.704 & 0.432 & 0.374 & 0.587 & 0.704 & 0.587 & 0.374 & 0.242 & 0.374 & 0.432 & 0.374 & 0.106 & 0.178 & 0.208 & 0.016 & 0.043 \\
0.587 & 0.704 & 0.587 & 0.432 & 0.704 & 1 & 0.704 & 0.178 & 0.374 & 0.587 & 0.704 & 0.587 & 0.106 & 0.242 & 0.374 & 0.432 & 0.033 & 0.106 & 0.178 & 0 & 0.016 \\
0.374 & 0.587 & 0.704 & 0.208 & 0.432 & 0.704 & 1 & 0.043 & 0.178 & 0.374 & 0.587 & 0.704 & 0.016 & 0.106 & 0.242 & 0.374 & 0 & 0.033 & 0.106 & 0 & 0 \\
0.242 & 0.106 & 0.016 & 0.587 & 0.374 & 0.178 & 0.043 & 1 & 0.704 & 0.432 & 0.208 & 0.056 & 0.704 & 0.587 & 0.374 & 0.178 & 0.432 & 0.374 & 0.242 & 0.208 & 0.178 \\
0.374 & 0.242 & 0.106 & 0.704 & 0.587 & 0.374 & 0.178 & 0.704 & 1 & 0.704 & 0.432 & 0.208 & 0.587 & 0.704 & 0.587 & 0.374 & 0.374 & 0.432 & 0.374 & 0.178 & 0.208 \\
0.432 & 0.374 & 0.242 & 0.587 & 0.704 & 0.587 & 0.374 & 0.432 & 0.704 & 1 & 0.704 & 0.432 & 0.374 & 0.587 & 0.704 & 0.587 & 0.242 & 0.374 & 0.432 & 0.106 & 0.178 \\
0.374 & 0.432 & 0.374 & 0.374 & 0.587 & 0.704 & 0.587 & 0.208 & 0.432 & 0.704 & 1 & 0.704 & 0.178 & 0.374 & 0.587 & 0.704 & 0.106 & 0.241 & 0.374 & 0.033 & 0.106 \\
0.242 & 0.374 & 0.432 & 0.178 & 0.374 & 0.587 & 0.704 & 0.056 & 0.208 & 0.432 & 0.704 & 1 & 0.043 & 0.178 & 0.374 & 0.587 & 0.016 & 0.106 & 0.242 & 0 & 0.033 \\
0.106 & 0.033 & 0 & 0.374 & 0.242 & 0.106 & 0.016 & 0.704 & 0.587 & 0.374 & 0.178 & 0.043 & 1 & 0.704 & 0.432 & 0.208 & 0.704 & 0.587 & 0.374 & 0.432 & 0.374 \\
0.178 & 0.106 & 0.033 & 0.432 & 0.374 & 0.242 & 0.106 & 0.587 & 0.704 & 0.587 & 0.374 & 0.178 & 0.704 & 1 & 0.704 & 0.432 & 0.587 & 0.704 & 0.587 & 0.374 & 0.432 \\
0.208 & 0.178 & 0.106 & 0.274 & 0.432 & 0.374 & 0.242 & 0.374 & 0.587 & 0.704 & 0.587 & 0.374 & 0.432 & 0.704 & 1 & 0.704 & 0.374 & 0.587 & 0.704 & 0.242 & 0.374 \\
0.178 & 0.208 & 0.178 & 0.242 & 0.374 & 0.432 & 0.374 & 0.178 & 0.374 & 0.587 & 0.704 & 0.587 & 0.208 & 0.432 & 0.704 & 1 & 0.178 & 0.374 & 0.587 & 0.106 & 0.242 \\
0.016 & 0 & 0 & 0.178 & 0.106 & 0.033 & 0 & 0.432 & 0.374 & 0.242 & 0.106 & 0.016 & 0.704 & 0.587 & 0.374 & 0.178 & 1 & 0.704 & 0.432 & 0.704 & 0.587 \\
0.043 & 0.016 & 0 & 0.208 & 0.178 & 0.106 & 0.033 & 0.374 & 0.432 & 0.374 & 0.242 & 0.106 & 0.587 & 0.704 & 0.587 & 0.374 & 0.704 & 1 & 0.704 & 0.587 & 0.704 \\
0.056 & 0.043 & 0.016 & 0.178 & 0.208 & 0.178 & 0.106 & 0.242 & 0.374 & 0.432 & 0.374 & 0.242 & 0.374 & 0.587 & 0.704 & 0.587 & 0.432 & 0.704 & 1 & 0.374 & 0.587 \\
0 & 0 & 0 & 0.043 & 0.016 & 0 & 0 & 0.208 & 0.178 & 0.106 & 0.033 & 0 & 0.432 & 0.374 & 0.242 & 0.106 & 0.704 & 0.587 & 0.374 & 1 & 0.704 \\
0 & 0 & 0 & 0.056 & 0.043 & 0.016 & 0 & 0.178 & 0.208 & 0.178 & 0.106 & 0.033 & 0.374 & 0.432 & 0.374 & 0.242 & 0.587 & 0.704 & 0.587 & 0.704 & 1
\end{bmatrix}$$

$\mathbf{\Sigma_{212}}$ represents the cross-covariance matrix between variate 1 and variate 2. The element in the first row and first column is the spherical covariance function value between variate 1 and variate 2 for the observation (1,4). Since the ranges are assumed to be equal, the spherical covariance function for the cross-covariance matrix becomes $\rho * sph(d) = 0.4347 * \left\{ 1 * \sqrt{5} \left(1 - \frac{3}{2}\left(\frac{d}{5}\right) + \frac{1}{2}\left(\frac{d}{5}\right)^3 \right) \right\}$ where $\sigma_{21} = 1$ and $\sigma_{22} = \sqrt{5}$ are the square roots of the sill values from the first and second variates respectively (Oliver 2003). Therefore, the value in the cross-covariance matrix in the first row and first column (i.e. the observation at the point (1,4) in the figures) is $0.4347 * \left\{ 1 * \sqrt{5} \left(1 - \frac{3}{2}\left(\frac{0}{5}\right) + \frac{1}{2}\left(\frac{0}{5}\right)^3 \right) \right\} = 0.972$. The element in the first row and the second column of $\mathbf{\Sigma_{212}}$ is the spherical covariance function between the observation from variate 1 at the point (1,4) and the observation from variate 2 at the point (1,5). The distance between these two observations is 1, and the value in $\mathbf{\Sigma_{212}}$ is 0.684.

$$\boldsymbol{\Sigma}_{212} = \begin{bmatrix}
0.972 & 0.684 & 0.420 & 0.571 & 0.684 & 0.571 & 0.363 & 0.235 & 0.363 & 0.420 & 0.363 & 0.235 & 0.103 & 0.173 & 0.202 & 0.173 & 0.016 & 0.042 & 0.54 & 0 & 0 \\
0.684 & 0.972 & 0.684 & 0.363 & 0.571 & 0.684 & 0.571 & 0.103 & 0.235 & 0.363 & 0.420 & 0.363 & 0.031 & 0.103 & 0.173 & 0.202 & 0 & 0.016 & 0.042 & 0 & 0 \\
0.420 & 0.684 & 0.972 & 0.173 & 0.363 & 0.571 & 0.684 & 0.016 & 0.103 & 0.235 & 0.363 & 0.420 & 0 & 0.031 & 0.103 & 0.173 & 0 & 0 & 0.016 & 0 & 0 \\
0.571 & 0.363 & 0.173 & 0.972 & 0.684 & 0.420 & 0.202 & 0.571 & 0.684 & 0.571 & 0.363 & 0.173 & 0.363 & 0.420 & 0.363 & 0.235 & 0.173 & 0.202 & 0.173 & 0.042 & 0.054 \\
0.684 & 0.571 & 0.363 & 0.684 & 0.972 & 0.684 & 0.420 & 0.363 & 0.571 & 0.684 & 0.571 & 0.363 & 0.235 & 0.363 & 0.420 & 0.363 & 0.103 & 0.173 & 0.202 & 0.016 & 0.042 \\
0.571 & 0.684 & 0.571 & 0.420 & 0.684 & 0.972 & 0.684 & 0.173 & 0.363 & 0.571 & 0.684 & 0.571 & 0.103 & 0.235 & 0.363 & 0.420 & 0.031 & 0.103 & 0.173 & 0 & 0.016 \\
0.363 & 0.571 & 0.684 & 0.202 & 0.420 & 0.684 & 0.972 & 0.042 & 0.173 & 0.363 & 0.571 & 0.684 & 0.016 & 0.103 & 0.235 & 0.363 & 0 & 0.031 & 0.103 & 0 & 0 \\
0.235 & 0.103 & 0.106 & 0.571 & 0.363 & 0.173 & 0.042 & 0.972 & 0.684 & 0.420 & 0.202 & 0.054 & 0.684 & 0.571 & 0.363 & 0.173 & 0.420 & 0.363 & 0.235 & 0.202 & 0.173 \\
0.363 & 0.235 & 0.103 & 0.684 & 0.571 & 0.363 & 0.173 & 0.684 & 0.972 & 0.684 & 0.420 & 0.202 & 0.571 & 0.684 & 0.571 & 0.363 & 0.363 & 0.420 & 0.363 & 0.173 & 0.202 \\
0.420 & 0.363 & 0.235 & 0.571 & 0.684 & 0.571 & 0.363 & 0.420 & 0.684 & 0.972 & 0.684 & 0.420 & 0.363 & 0.571 & 0.684 & 0.571 & 0.235 & 0.363 & 0.420 & 0.103 & 0.173 \\
0.363 & 0.420 & 0.363 & 0.363 & 0.571 & 0.684 & 0.571 & 0.202 & 0.420 & 0.684 & 0.972 & 0.684 & 0.173 & 0.363 & 0.571 & 0.684 & 0.103 & 0.235 & 0.363 & 0.031 & 0.103 \\
0.235 & 0.363 & 0.420 & 0.173 & 0.363 & 0.571 & 0.684 & 0.054 & 0.202 & 0.420 & 0.684 & 0.972 & 0.042 & 0.173 & 0.363 & 0.571 & 0.016 & 0.103 & 0.235 & 0 & 0.031 \\
0.103 & 0.032 & 0 & 0.363 & 0.235 & 0.102 & 0.016 & 0.684 & 0.571 & 0.363 & 0.173 & 0.042 & 0.972 & 0.684 & 0.420 & 0.2020 & 0.684 & 0.571 & 0.363 & 0.420 & 0.363 \\
0.173 & 0.103 & 0.031 & 0.420 & 0.363 & 0.235 & 0.103 & 0.571 & 0.684 & 0.581 & 0.363 & 0.173 & 0.684 & 0.972 & 0.684 & 0.420 & 0.571 & 0.684 & 0.571 & 0.363 & 0.420 \\
0.202 & 0.173 & 0.103 & 0.363 & 0.420 & 0.363 & 0.235 & 0.363 & 0.571 & 0.684 & 0.571 & 0.363 & 0.420 & 0.684 & 0.972 & 0.684 & 0.363 & 0.571 & 0.684 & 0.235 & 0.363 \\
0.173 & 0.202 & 0.173 & 0.235 & 0.363 & 0.420 & 0.363 & 0.173 & 0.363 & 0.571 & 0.684 & 0.571 & 0.202 & 0.420 & 0.684 & 0.972 & 0.173 & 0.363 & 0.571 & 0.103 & 0.235 \\
0.016 & 0 & 0 & 0.173 & 0.103 & 0.031 & 0 & 0.420 & 0.363 & 0.235 & 0.103 & 0.016 & 0.684 & 0.571 & 0.363 & 0.173 & 0.972 & 0.684 & 0.420 & 0.684 & 0.571 \\
0.042 & 0.016 & 0 & 0.202 & 0.173 & 0.103 & 0.031 & 0.363 & 0.420 & 0.363 & 0.235 & 0.103 & 0.571 & 0.684 & 0.571 & 0.363 & 0.684 & 0.972 & 0.684 & 0.571 & 0.684 \\
0.054 & 0.042 & 0.016 & 0.173 & 0.202 & 0.173 & 0.103 & 0.235 & 0.363 & 0.420 & 0.363 & 0.235 & 0.363 & 0.571 & 0.684 & 0.571 & 0.420 & 0.684 & 0.972 & 0.363 & 0.571 \\
0 & 0 & 0 & 0.042 & 0.016 & 0 & 0 & 0.202 & 0.173 & 0.103 & 0.031 & 0 & 0.420 & 0.363 & 0.235 & 0.103 & 0.684 & 0.571 & 0.363 & 0.972 & 0.684 \\
0 & 0 & 0 & 0.054 & 0.042 & 0.016 & 0 & 0.173 & 0.202 & 0.173 & 0.103 & 0.031 & 0.363 & 0.420 & 0.363 & 0.235 & 0.571 & 0.684 & 0.571 & 0.684 & 0.972
\end{bmatrix}$$

$\boldsymbol{\Sigma}_{222}$ is the variance-covariance matrix for the second variate. Since variate 2 has a range = 5 and a sill = 5, the spherical covariance becomes $sph(d) = 5*\left\{1-\frac{3}{2}\left(\frac{d}{5}\right)+\frac{1}{2}\left(\frac{d}{5}\right)^3\right\}$. Therefore, along the diagonal when the distance between observations is 0, the spherical covariance function value is $sph(0) = 5*\left\{1-\frac{3}{2}\left(\frac{0}{5}\right)+\frac{1}{2}\left(\frac{0}{5}\right)^3\right\} = 5$. The element in the first row and the second column of $\boldsymbol{\Sigma}_{222}$ is the spherical covariance function value between observation (1,4) and observation (1,5). These observations are at a distance of 1 and their spherical covariance function value is 3.520.

$$\Sigma_{222} = \begin{bmatrix}
5 & 3.520 & 0.2.160 & 2.935 & 3.520 & 2.935 & 1.870 & 1.210 & 1.870 & 2.160 & 1.870 & 1.210 & 0.529 & 0.889 & 1.040 & 0.889 & 0.081 & 0.217 & 0.280 & 0 & 0 \\
3.520 & 5 & 3.520 & 1.870 & 2.935 & 3.520 & 2.935 & 0.529 & 1.210 & 1.870 & 2.160 & 1.870 & 0.163 & 0.529 & 0.889 & 1.040 & 0 & 0.081 & 0.217 & 0 & 0 \\
2.160 & 3520 & 5 & 0.889 & 1.870 & 2.935 & 3.520 & 0.081 & 0.529 & 1.210 & 1.870 & 2.160 & 0 & 0.163 & 0.529 & 0.889 & 0 & 0 & 0.081 & 0 & 0 \\
2.935 & 1.870 & 0.889 & 5 & 3.520 & 2.160 & 1.040 & 2.935 & 3.520 & 2.935 & 1.870 & 0.889 & 1.870 & 2.160 & 1.870 & 1.210 & 0.889 & 1.040 & 0.889 & 0.217 & 0.280 \\
3.520 & 2.935 & 1.870 & 3.520 & 5 & 3.520 & 2.160 & 1.870 & 2.935 & 3.520 & 2.935 & 1.870 & 1.210 & 1.870 & 2.160 & 1.870 & 0.529 & 0.889 & 1.040 & 0.081 & 0.217 \\
2.935 & 3.520 & 2.935 & 2.160 & 3.520 & 5 & 3.520 & 0.889 & 1.870 & 2.935 & 3.520 & 2.935 & 0.529 & 1.210 & 1.870 & 2.160 & 0.163 & 0.529 & 0.889 & 0 & 0.081 \\
1.870 & 2.935 & 3.520 & 1.040 & 2.160 & 3.520 & 5 & 0.217 & 0.889 & 1.870 & 2.935 & 3.520 & 0.081 & 0.529 & 1.210 & 1.870 & 0 & 0.163 & 0.529 & 0 & 0 \\
1.210 & 0.529 & 0.081 & 2.935 & 1.870 & 0.889 & 0.217 & 5 & 3.520 & 2.160 & 1.040 & 0.280 & 3.520 & 2.935 & 1.870 & 0.889 & 2.160 & 1.870 & 1.210 & 1.040 & 0.889 \\
1.870 & 1.210 & 0.529 & 3.520 & 2.935 & 1.870 & 0.889 & 3.520 & 5 & 3.520 & 2.16 & 1.040 & 2.935 & 3.520 & 2.935 & 1.870 & 1.870 & 2.160 & 1.870 & 0.889 & 1.040 \\
2.160 & 1.870 & 1.210 & 2.935 & 3.520 & 2.935 & 1.870 & 2.160 & 3.520 & 5 & 3.520 & 2.160 & 1.870 & 2.935 & 3.520 & 2.935 & 1.210 & 1.870 & 2.160 & 0.529 & 0.889 \\
1.870 & 2.160 & 1.870 & 1.870 & 2.935 & 3.520 & 2.935 & 1.040 & 2.160 & 3.520 & 5 & 3.520 & 0.889 & 1.870 & 2.935 & 3.520 & 0.529 & 1.210 & 1.870 & 0.163 & 0.529 \\
1.210 & 1.870 & 2.160 & 0.889 & 1.870 & 2.935 & 3.520 & 0.281 & 1.040 & 2.160 & 3.520 & 5 & 0.217 & 0.889 & 1.870 & 2.935 & 0.081 & 0.529 & 1.210 & 0 & 0.163 \\
0.529 & 0.163 & 0 & 1.870 & 1.210 & 0.529 & 0.081 & 3.520 & 2.935 & 1.870 & 0.889 & 0.217 & 5 & 3.520 & 2.160 & 1.040 & 3.520 & 2.935 & 1.870 & 2.160 & 1.870 \\
0.889 & 0.529 & 0.163 & 2.160 & 1.870 & 1.210 & 0.529 & 2.935 & 3.520 & 2.935 & 1.870 & 0.889 & 3.520 & 5 & 3.520 & 2.160 & 2.935 & 3.520 & 2.935 & 1.870 & 2.160 \\
1.040 & 0.889 & 0.529 & 1.870 & 2.160 & 1.870 & 1.210 & 1.870 & 2.935 & 3.520 & 2.935 & 1.870 & 2.160 & 3.520 & 5 & 3.520 & 1.870 & 2.935 & 3.520 & 1.210 & 1.870 \\
0.889 & 1.040 & 0.889 & 1.210 & 1.870 & 2.160 & 1.870 & 0.889 & 1.870 & 2.935 & 3.520 & 2.935 & 1.040 & 2.160 & 3.520 & 5 & 0.889 & 1.870 & 2.935 & 0.529 & 1.210 \\
0.080 & 0 & 0 & 0.889 & 0.529 & 0.163 & 0 & 2.160 & 1.870 & 1.210 & 0.529 & 0.081 & 3.520 & 2.935 & 1.870 & 0.889 & 5 & 3.520 & 2.160 & 3.520 & 2.935 \\
0.217 & 0.081 & 0 & 1.040 & 0.889 & 0.529 & 0.163 & 1.870 & 2.160 & 1.870 & 1.210 & 0.529 & 2.935 & 3.520 & 2.935 & 1.870 & 3.520 & 5 & 3.520 & 2.935 & 3.520 \\
0.280 & 0.217 & 0.081 & 0.889 & 1.040 & 0.889 & 0.529 & 1.210 & 1.870 & 2.160 & 1.870 & 1.210 & 1.870 & 2.935 & 3.520 & 2.935 & 2.160 & 3.520 & 5 & 1.870 & 2.935 \\
0 & 0 & 0 & 0.217 & 0.081 & 0 & 0 & 1.040 & 0.889 & 0.529 & 0.163 & 0 & 2.160 & 1.870 & 1.210 & 0.529 & 3.520 & 2.935 & 1.870 & 5 & 3.520 \\
0 & 0 & 0 & 0.280 & 0.217 & 0.081 & 0 & 0.889 & 1.040 & 0.889 & 0.529 & 0.163 & 1.870 & 2.160 & 1.870 & 1.210 & 2.935 & 3.520 & 2.935 & 3.520 & 5
\end{bmatrix}$$

The cross-covariance matrices for the remaining clusters are below:

$$\boldsymbol{\Sigma}_3^* = \left[\begin{array}{cccccc:cccccc} 1 & 0.587 & 0.704 & 0.242 & 0.374 & 0.432 & 0.615 & 0.361 & 0.433 & 0.149 & 0.230 & 0.266 \\ 0.587 & 1 & 0.704 & 0.587 & 0.704 & 0.587 & 0.361 & 0.615 & 0.433 & 0.361 & 0.433 & 0.361 \\ 0.704 & 0.704 & 1 & 0.374 & 0.587 & 0.704 & 0.433 & 0.433 & 0.615 & 0.230 & 0.361 & 0.433 \\ 0.242 & 0.587 & 0.374 & 1 & 0.704 & 0.432 & 0.149 & 0.361 & 0.230 & 0.615 & 0.433 & 0.266 \\ 0.374 & 0.704 & 0.587 & 0.704 & 1 & 0.704 & 0.230 & 0.433 & 0.361 & 0.433 & 0.615 & 0.433 \\ 0.432 & 0.587 & 0.704 & 0.432 & 0.704 & 1 & 0.266 & 0.361 & 0.433 & 0.266 & 0.433 & 0.615 \\ \hdashline 0.615 & 0.361 & 0.433 & 0.149 & 0.230 & 0.266 & 5 & 2.935 & 3.520 & 1.210 & 1.870 & 2.160 \\ 0.361 & 0.615 & 0.433 & 0.361 & 0.433 & 0.361 & 2.935 & 5 & 3.520 & 2.935 & 3.520 & 2.935 \\ 0.433 & 0.433 & 0.615 & 0.230 & 0.361 & 0.433 & 3.520 & 3.520 & 5 & 1.870 & 2.935 & 3.520 \\ 0.149 & 0.361 & 0.230 & 0.615 & 0.433 & 0.266 & 1.210 & 2.935 & 1.870 & 5 & 3.520 & 2.160 \\ 0.230 & 0.433 & 0.361 & 0.433 & 0.615 & 0.433 & 1.870 & 3.520 & 2.935 & 3.520 & 5 & 3.520 \\ 0.266 & 0.361 & 0.433 & 0.266 & 0.433 & 0.615 & 2.160 & 2.935 & 3.520 & 2.160 & 3.520 & 5 \end{array}\right]$$

$$\boldsymbol{\Sigma}_4^* = \left[\begin{array}{ccc:ccc} 1 & 0.704 & 0.432 & 1.214 & 0.854 & 0.524 \\ 0.704 & 1 & 0.704 & 0.854 & 1.214 & 0.854 \\ 0.432 & 0.704 & 1 & 0.524 & 0.854 & 1.214 \\ \hdashline 1.214 & 0.854 & 0.524 & 5 & 3.520 & 2.160 \\ 0.854 & 1.214 & 0.854 & 3.520 & 5 & 3.520 \\ 0.524 & 0.854 & 1.214 & 2.160 & 3.520 & 5 \end{array}\right].$$

3.5 Conclusions

The actual geographic location of an observation can be incorporated into the variance-covariance matrix of the multivariate normal distribution used in the clustering algorithm. The variance-covariance matrix can be computed using any specific covariance function and the spherical covariance function was chosen for this research. The main challenge in the spatial clustering process was to adequately model the cross-covariance between the variates recorded while taking into account the spatial location of the observations. This was accomplished by incorporating the correlation between the variates, in addition to the variability within each variate. Oliver's (2003) method for

calculating the cross-covariance was implemented in the creation of the variance-covariance matrix used in the multivariate normal distribution.

There are many possible clustering schemes for a given data set. Thus, the most appropriate clustering scheme was determined using the likelihood itself, as well as the AIC. Since the AIC is computed using the estimated likelihood, it also accounts for the spatial variability within the observations.

3.6 References

Adamchuk, V. I. 2006 *Site-Specific Management Guidelines*. Potash & Phosphate Institute, in cooperation with the Foundation of Agronomic Research.

Adamchuk, V.I., D.B. Marx, A.T. Kerby, A.K. Samal, L.K. Soh, R.B. Ferguson, and C.S. Wortmann. 2007. Guided soil sampling for enhanced analysis of georefernced sensor-based data. In: Proceedings of the Ninth International Conference on Geocomputation 2007 Conference, Maynooth, Ireland, 3-5 September 2007, U. Demsar, ed. Maynooth, Ireland: NCG - National University of Ireland (E-proceedings, 4 pages).

Akaike, H. 1974 A New Look at the Statistical Model Identification. *IEEE Transaction on Automatic Control*, AC 19, 716-723.

Bourgault, Gilles, Marcotte, Denis, & Legendre Pierre 1992 The Multivariate (Co)Variogram as a Spatial Weighting Function in Classification Methods. *Mathematical Geology*. Vol. 24, No. 5, 463-478.

Cressie, N. 1991 *Spatial Statistics for Spatial Data*. New York: John Wiley & Sons, Inc.

Everitt, B. 1974 *Cluster Analysis*. Toronto: Heinemann Educational Books Ltd.

Frogbrook, Z. L. & Oliver, M. A. 2007 Identifying management zones in agricultural fields using spatially constrained classification of soil and ancillary data. *Soil Use and Management*. 23, 40 – 51.

Hartigan, J. A. 1975 *Clustering Algorithms*. New York: John Wiley & Sons, Inc.

Isaaks, E. H. & Srivastava, R. M. 1989 *An Introduction to Applied Geostatistics*. New York: Oxford University Press, Inc.

Johnson, D. E. 1998 *Applied Multivariate Methods for Data Analysis*. Pacific Grove: Brooks/Cole Publishing Company.

Johnson, R. A. & Wichern, D. W. 2002 *Applied Multivariate Statistical Analysis*. Upper Saddle River: Prentice-Hall, Inc.

Kaufman, L. & Rousseeuw, P. J. 1990 *Finding Groups in Data An Introduction to Cluster Analysis*. New York: John Wiley & Sons, Inc.

Oliver, D. S. 2003 Guassian Cosimulation: Modelling of the Cross-Covariance. *Mathematical Geology*. Vol. 35, No. 6, 681-698.

Press, W. Flannery, B. Teukolsky, S. & Vetterling, W. 1988 *Numerical Recipes in C The Art of Scientific Computing*. New York: Cambridge University Press.

R Development Core Team. 2007 *R: a language and environment for statistical computing*. Vienna, Austria. R Foundation for Statistical Computing.

SAS Institute. 2008 *SAS Online Doc.* Version 9.2. SAS Institute, Inc. Cary, NC.

Schabenberger, O. & Gotway, C. A. 2005 *Spatial Methods for Spatial Data Analysis*. New York: Chapman & Hall/CRC Press.

Chapter 4

Spatial Clustering Incorporating a Categorical Variable

4.1 Introduction

Cluster analysis is a tool used to place similar observations into groups or clusters. These clusters are based on measures of similarity or dissimilarity so that the similarity of observations within a cluster is maximized and the dissimilarity with other observations is also maximized. These measures of similarity (or dissimilarity) are usually based upon the Euclidean, standardized Euclidean or Mahalanobis distance calculations (Everitt 1974, Hartigan 1975, Johnson 1998, Johnson & Wichern 2002, Kaufman & Rousseeuw). However, not all recorded variables of interest are numeric. Therefore, clustering algorithms such as *k*-modes (Huang 1998), CACTUS (Ganti et al. 1999), COOLCAT (Barbará et al. 2002), Squeezer (Zengyou et al. 2002), and ROCK (Guha et al. 2000) have been created to cluster non-numeric (categorical) variables.

Similar to the clustering algorithms for numeric data, these algorithms do not account for the underlying spatial structure of the data. There are cases in which the spatial (geographical) location is known and relevant to the analysis. One example is site-specific crop management, which has become an important aspect of agriculture production in recent years. Precision agriculture methods use multiple data layers within spatially variable observations to fine-tune crop management decisions. Since conventional coarse (approximately 1-ha) grid sampling fails to provide adequate

representation of spatial variability in soils, alternative high-density sensor data have been used in many operations.

One of the major challenges in the data analysis process is to delineate field areas with potential for differentiated treatments that are frequently called “management zones.” Initially, a relatively inexpensive set of data such as on-the-go soil sensing maps and/or remote sensing imagery are collected. These data are very dense and can be used to define areas for targeted (guided) sampling which will provide detailed information about the agronomic quality of land through the analysis of soil samples run in a commercial lab. Since only a limited number of these costly samples can be afforded, they should come from homogenous areas of the field, away from boundaries or locations where sensor data changes significantly over short distances, and spread across the entire landscape. Some of the categorical responses recorded might be soil type, soil structure or soil texture (Adamchuk 2006, Adamchuk et al. 2007, Frogbrook & Oliver 2007). Thus, a proper clustering method is needed to delineate relatively homogeneous field areas while accounting for the physical values of high-density observations and their spatial distribution.

In this paper a clustering method to incorporate the spatial structure of the observations is proposed. The case when only categorical responses have been recorded on the observations is investigated first. After spatial clustering of solely categorical responses has been implemented, the algorithm is extended to include both numeric and categorical responses. The final algorithm accounts for the spatial pattern of the data, the actual numeric responses and the categorical responses.

4.2 Clustering Categorical Data Only

4.2.1 Background

There have been many different approaches to creating a clustering algorithm to handle categorical values. An entropy-based approach was utilized in the COOLCAT (Barbará et al. 2002) algorithm, whose authors define entropy as a "measure of "disorder" in a system." Therefore, observations are clustered based upon whether or not the disorder or entropy decreases by combining observations to create clusters.

CACTUS (Ganti et al. 1999) is an algorithm to cluster categorical variables in a three-stage process: summarizing, clustering, validation. CACTUS (Ganti et al. 1999) uses a summary of all the data to define clusters. The summary information is then used in the clustering stage to cluster the observations. The validation stage is where the best clustering of the data is determined from a "set of candidate clusters" (Ganti et al. 1999). The decision to form a cluster is determined by a threshold that is a function of the size of the data set, the size of the domains of the attributes and an α-level. If the number of connected pairs is larger than the threshold, a cluster is formed (Ganti et al. 1999).

An extension to the rather popular *k*-means clustering algorithm has been developed to cluster categorical variables (Huang 1998). This clustering algorithm uses the mode rather than the mean of the observations in the clustering process and uses a count of the number of mismatches between observations (Huang 1998). Therefore, those observations with the smallest number of mismatches are the most alike in the dataset. Cluster assignation is based upon the mode of the cluster rather than the mean of

the cluster as done in the *k*-means algorithm. The *k*-modes (Huang 1998) algorithm is outlined below:

Step 1: Select *k* initial modes, one for each cluster

Step 2: Allocate an object to the cluster whose mode is the nearest according to

$d_1(X,Y) = \sum_{j=1}^{m} \delta(x_j, y_j)$ where m is the number of attributes and

$\delta(x_j, y_j) = \begin{cases} 0 & x_j = y_j \\ 1 & x_j \neq y_j \end{cases}$ is the dissimilarity measure. The mode of each cluster is updated after each allocation so that the function $D(\mathbf{X}, Q)$ (where the mode of $\mathbf{X}$ is a vector Q that minimizes $D(\mathbf{X}, Q) = \sum_{i=1}^{n} d_1(X_i, Q))$ is minimized iff

$f_r(A_j = q_j \mid \mathbf{X}) \geq f_r(A_j = c_{k,j} \mid \mathbf{X})$ for $q_j \neq c_{k,j}$ for all j = 1, …, m.

Step 3: After all objects have been allocated to clusters, retest the dissimilarity of the objects against the current modes. If an object is found such that its nearest mode belongs to another cluster rather than its current one, reallocate the object and update the modes of both clusters.

Step 4: Repeat Step 3 until no object has changed clusters after a full cycle test of the whole data set.

Unlike the *k*-modes algorithm, the Squeezer (Zengyou et al. 2002) algorithm does not require an initial cluster size however the desired degree of similarity within a cluster, *s*, must be specified. Zengyou et al. (2002) denote the set of categorical attributes as A_1,

…, A_m with domains D_1, …, D_m, and D to be the set of tuples. Each tuple t is an element of $D_1 \times \ldots \times D_m$ and TID is the set of unique IDs for the tuples. Therefore, a tuple is represented by $val(tid, A_i)$ where $tid \in TID$ with attribute value A_i. A cluster is then defined as $C = \{tid \mid tid \in TID\}$ where the set of attribute values on A_i with respect to C are defined as $VAL_i(C) = \{val(tid, A_i) \mid tid \in C\}$. Zengyou et al. (2002) also define the support of a_i in C with respect to A_i as $Sup(a_i) = |\{tid \mid tid.A_i = a_i\}|$ given the cluster C and $a_i \in D_i$. Thus, the support of an attribute is the number of tuples in the cluster which contain the attribute value.

Once the data set D and the similarity threshold s have been inputted into the Squeezer algorithm the first tuple is read and placed into the first cluster. The next tuple is read and the similarity between the first tuple (cluster) and the new tuple is calculated using $Sim(C, tid) = \sum_{i=1}^{m} \left(\frac{Sup(a_i)}{\sum_j Sup(a_j)} \right)$ where $tid.A_i = a_i$ and $a_j \in VAL_i(C)$ (Zengyou et al. 2002). If the similarity between the new tuple and the first cluster is greater than s then the tuple is added to the cluster. If the similarity between the new tuple and the first cluster is not greater than s then a new cluster is formed. The algorithm then reads in another tuple and computes the similarity between the tuple and all established clusters. The new tuple either merges with the cluster with which it has the largest similarity above the threshold or it forms a new cluster. This process occurs until all the data have been read and clusters have been formed. (Zengyou et al. 2002)

Another categorical clustering algorithm that uses a similarity measure to cluster observations is ROCK (RObust Clustering using linKs) (Guha et al. 2000). Similar to the *k*-modes algorithm, an initial estimate for the number of clusters must be specified. This clustering approach uses the notion of neighbors and links to determine whether two observations should be placed into the same cluster. Only those observations which are deemed neighbors and have a large number of links between them will be merged to form the *k* clusters (Guha et al 2000).

In order to determine whether two observations are considered neighbors the similarity between them $\left(sim\left(p_i, p_j\right)\right)$ is calculated (Guha et al. 2000). The similarity is a normalized function which describes the closeness between the points p_i and p_j. It is assumed that *sim* is a function which takes on values between 0 and 1, with larger values indicating points which are more similar in nature. Therefore, points p_i and p_j are deemed neighbors if $sim\left(p_i, p_j\right) \geq \theta$ where θ is a user specified threshold (Guha et al. 2000). For example, let T_i represent the attributes for point p_i and T_j represent the attributes for point p_j. The similarity between points p_i and p_j could be defined as $sim\left(T_i, T_j\right) = \frac{\left|T_i \cap T_j\right|}{\left|T_i \cup T_j\right|}$.

Since $\left|T_i \cup T_j\right|$ is in the denominator, this assures the similarity is always between 0 and 1 (Guha et al. 2000).

The similarity can be used to determine which points are neighbors and once neighbors have been established, the number of links $\left(link\left(p_i, p_j\right)\right)$ between neighbors is found. Links are defined "to be the number of common neighbors between points p_i and

p_j" (Guha et al. 2000). The number of links will be used to determine how closely related points within a cluster are. Therefore Guha et al. (2000) propose a criterion function to maximize the number of links between pairs of points within a cluster. The criterion function $E_l = \sum_{i=1}^{k} n_i * \sum_{p_q, p_r \in C_i} \frac{link\left(p_q, p_r\right)}{n_i^{1+2f(\theta)}}$, takes into account the size of the clusters (n_i) and the number of clusters (k) (Guha et al. 2000). The denominator is used to weight the function by the estimated number of links for a cluster of size n_i. The function $f(\theta)$ is dependent on the data and assumes that every point in the cluster C_i has approximately $n_i^{f(\theta)}$ neighbors (Guha et al. 2000).

Guha et al. (2000) also define a similar criterion function used in practice to determine which clusters should be merged in the clustering process. The function $g\left(C_i, C_j\right) = \frac{link\left(C_i, C_j\right)}{\left(n_i + n_j\right)^{1+2f(\theta)} - n_i^{1+2f(\theta)} - n_j^{1+2f(\theta)}}$ is used to determine whether there is a sufficient number of cross links between the clusters to be merged (Guha et al. 2000). Again, the estimated number of cross links between the clusters is used in the denominator. Larger values for this criterion would mean there are more cross links (similar observations) between the two clusters and the best clustering of the data would maximize $g\left(C_i, C_j\right)$ (Guha et al. 2000).

The actual ROCK (Guha et al. 2000) clustering algorithm uses a combination of the above similarity, link and criterion functions to cluster a given set of data. Initially, each observation inputted is considered to be its own cluster. Each observation's neighbors are established and the number of links between clusters (observations) is

calculated. The clusters which have the largest $g(C_i, C_j)$ are merged to form a new cluster. The process then recalculates the number of links between the clusters and combines those which have the largest $g(C_i, C_j)$ between them. This process continues until the desired number of clusters (*k*) is reached or the number of links between the clusters is no longer non-zero (Guha et al. 2000).

Suppose sixteen samples were taken from a small plot in the corner of a field. From these samples, three variables were recorded: soil type (clay or silt), sunlight (sun or shade) and vegetation (whether there was any present or not, i.e. yes or no). The *k*-modes, Squeezer and ROCK algorithms are used to cluster the data given in Figure 4.1.

Clay	**Clay**	**Silt**	**Silt**
Clay	**Clay**	**Clay**	**Silt**
Clay	**Clay**	**Silt**	**Silt**
Silt	**Silt**	**Silt**	**Silt**

(a) Soil Type

Sun	**Sun**	**Sun**	**Shade**
Sun	**Sun**	**Sun**	**Shade**
Sun	**Shade**	**Shade**	**Shade**
Shade	**Shade**	**Shade**	**Shade**

(b) Sunlight

No	**No**	**No**	**No**
No	**No**	**No**	**No**
No	**Yes**	**Yes**	**Yes**
Yes	**Yes**	**Yes**	**Yes**

(c) Vegetation

=

Clay Sun No	**Clay Sun No**	**Silt Sun No**	**Silt Shade No**
Clay Sun No	**Clay Sun No**	**Clay Sun No**	**Silt Shade No**
Clay Sun No	**Clay Shade Yes**	**Silt Shade Yes**	**Silt Shade Yes**
Silt Shade Yes	**Silt Shade Yes**	**Silt Shade Yes**	**Silt Shade Yes**

Figure 4.1: Data to demonstrate *k*-modes, Squeezer & ROCK algorithms

The *k*-modes algorithm is used first to cluster the observations. In step 1 of the *k*-modes algorithm *k* initial modes must be chosen. Due to the size of the plot from which

the samples were taken, it seems reasonable that two clusters would sufficiently group the data. Thus, two initial modes must be selected from the sixteen samples taken. The frequencies for the three attributes were calculated and assigned to the initial modes Q_1 = {silt, sun, no} and Q_2 = {clay, shade, yes} ensuring that the categories with the most frequencies were evenly distributed between the two initial modes (Huang 1998). Next, the observations were compared to these initial modes and the most similar observations from the data replaced Q_1 and Q_2 as the initial modes. The observation (1,3) was an exact match to Q_1 which created the first initial cluster highlighted in blue in Figure 4.1. The second initial cluster consisted of the observation (3,2) which had an exact match to Q_2 highlighted in green in Figure 4.1 (Huang 1998).

The next step in the clustering process was to find the observations "closest" to the initial clusters. The "closest" cluster was determined by the smallest dissimilarity between the observation and the cluster as defined in Step 2 of the *k*-modes algorithm above. The observations with the smallest dissimilarity to the first cluster (blue cluster) were observations (1,1), (1,2), (1,4), (2,1), (2,2), (2,3), (2,4), and (3,1) with a dissimilarity value of 1. Therefore, these observations became a member of the first cluster. The observations (3,3), (3,4), (4,1), (4,2), (4,3), and (4,4) had a dissimilarity value equal to 1 with cluster two (green). Therefore, these observations became members of the second initial cluster. When the observations were tested against the mode in the cluster which it was not a member, no observations changed membership ending the clustering process. Thus, the final clustering of the observations is shown in Figure 4.2.

Clay Sun No	**Clay Sun No**	**Silt Sun No**	**Silt Shade No**
Clay Sun No	**Clay Sun No**	**Clay Sun No**	**Silt Shade No**
Clay Sun No	**Clay Shade Yes**	**Silt Shade Yes**	**Silt Shade Yes**
Silt Shade Yes	**Silt Shade Yes**	**Silt Shade Yes**	**Silt Shade Yes**

Figure 4.2: Final clustering of the data using the *k*-modes algorithm

The Squeezer (Zengyou 2002) algorithm is now used to cluster the observations, and an initial number of clusters does not need to be specified for this algorithm. However, the desired similarity within a cluster does need to be preset. A small similarity threshold would produce a smaller number of clusters since it would be easier for an observation to join an already established cluster. On the other hand, a larger similarity threshold results in more clusters, since a larger threshold necessitates more matches to join an established cluster. The maximum similarity possible for this example was 3 because there were three variables of interest. Therefore, a similarity of more than one half, but less than two thirds was desired for this analysis and 1.75 was used. The data table summarizing the attributes is found in Table 4.1 below.

Tuple ID	Soil Type	Sunlight	Vegetation
1	Clay	Sun	No
2	Clay	Sun	No
3	Silt	Sun	No
4	Silt	Shade	No
5	Clay	Sun	No
6	Clay	Sun	No
7	Clay	Sun	No
8	Silt	Shade	No
9	Clay	Sun	No
10	Clay	Shade	Yes
11	Silt	Shade	Yes
12	Silt	Shade	Yes
13	Silt	Shade	Yes
14	Silt	Shade	Yes
15	Silt	Shade	Yes
16	Silt	Shade	Yes

Table 4.1: Squeezer data table

Tuple 1 (Clay, Sun, No) was read into the algorithm and formed the first cluster C_1. The next step in the algorithm read in Tuple 2 (Clay, Sun, No) and computed the similarity between Tuple 2 and the existing clusters. In this case, Tuple 1 was the only cluster formed so the similarity between Tuple 1 and Tuple 2 was calculated using $Sim(C, tid) = \sum_{i=1}^{m} \left(\frac{Sup(a_i)}{\sum_j Sup(a_j)} \right)$ where $tid.A_i = a_i$ and $a_j \in VAL_i(C)$ (Zengyou et al. 2002). Tuple 1 and Tuple 2 have all three attributes in common (Clay, Sun, No) which resulted in a similarity value of 3 = $\left(\frac{1}{1}+\frac{1}{1}+\frac{1}{1}\right)$. Since 3 > 1.75, the tuples were combined and C_1 = {Tuple 1, Tuple 2}. Tuple 3 (Silt, Sun, No) was read and the similarity with C_1 was calculated. The resulting similarity was 2 = $\left(\frac{0}{2}+\frac{2}{2}+\frac{2}{2}\right)$ which is larger than the given threshold of 1.75 so Tuple 3 merged with C_1. At this point in the

algorithm C_1 = {Tuple 1, Tuple 2, Tuple 3}. Tuple 4 was read and the similarity with C_1 was calculated. The similarity between Tuple 4 and C_1 was 1.33 = $\left(\frac{1}{2+1}+\frac{0}{3}+\frac{3}{3}\right)$ which is not larger than 1.75. Therefore, Tuple 4 formed its own cluster and C_1 = {Tuple 1, Tuple 2, Tuple3} and C_2 = {Tuple 4}. After Tuples 5 – 9 were assigned to clusters, C_1= Tuple 1, Tuple 2, Tuple 3, Tuple 5, Tuple 6, Tuple 7, Tuple 9} and C_2 = {Tuple 4, Tuple 8}.

Tuple 10 was read and the similarity between the tuple and C_1 was calculated, as well as the similarity between the tuple and C_2. The similarity with C_1 was 0.86 = $\left(\frac{6}{6+1}+\frac{0}{7}+\frac{0}{7}\right)$ and the similarity with C_2 was 1 = $\left(\frac{0}{2}+\frac{2}{2}+\frac{0}{2}\right)$ which are not greater than the specified threshold. Therefore, Tuple 10 formed a new cluster and C_1 = {Tuple 1, Tuple 2, Tuple 3, Tuple 5, Tuple 6, Tuple 7, Tuple 9}, C_2 = {Tuple 4, Tuple 8} and C_3 = {Tuple 10}. Once Tuple 11 has been read and the similarities calculated, the similarity between Tuple 11 and C_1 was 0.14. The similarity between Tuple 11 and both C_2 = $\left(\frac{2}{2}+\frac{2}{2}+\frac{0}{2}\right)$ and C_3 = $\left(\frac{0}{1}+\frac{1}{1}+\frac{1}{1}\right)$ was 2. In this case, there was not only one, but two similarities which were maximum and larger than 1.75. Zengyou (2002) does not address what to do in the case of ties. It seemed reasonable to place Tuple 11 in the cluster in which it has more matches. In this case, C_2 had two observations and Tuple 11 had two attributes in common with both observations. C_3 only had one observation where Tuple 11 matched two of the attributes. Since there were more matches with C_2, Tuple 11 was merged with C_2. Thus, after this stage in the clustering algorithm, C_1 = {Tuple 1, Tuple

2, Tuple 3, Tuple 5, Tuple 6, Tuple 7, Tuple 9}, C_2 = {Tuple 4, Tuple 8, Tuple 11}, and C_3 = {Tuple 10}.

The last tuple was read into the algorithm and the following similarities were computed for clusters C_1, C_2 and C_3 respectively: 0.14 = $\left(\frac{1}{6+1}+\frac{0}{7}+\frac{0}{7}\right)$, 2.71 = $\left(\frac{7}{7}+\frac{7}{7}+\frac{5}{2+5}\right)$ and 2 = $\left(\frac{0}{1}+\frac{1}{1}+\frac{1}{1}\right)$. The largest similarity was extracted (2.71) and compared to the threshold. Since 2.71 > 1.75, Tuple 16 was merged with C_2. Hence, the final clustering of the data using the Squeezer algorithm (Zengyou 2002) was C_1 = {Tuple 1, Tuple 2, Tuple 3, Tuple 5, Tuple 6, Tuple 7. Tuple 9}, C_2 = {Tuple 4, Tuple 8, Tuple 11, Tuple 12, Tuple 13, Tuple 14, Tuple 15, Tuple 16} and C_3 = {Tuple 10} as shown in Figure 4.3.

Clay Sun No	**Clay Sun No**	**Silt Sun No**	**Silt Shade No**
Clay Sun No	**Clay Sun No**	**Clay Sun No**	**Silt Shade No**
Clay Sun No	**Clay Shade Yes**	**Silt Shade Yes**	**Silt Shade Yes**
Silt Shade Yes	**Silt Shade Yes**	**Silt Shade Yes**	**Silt Shade Yes**

Figure 4.3: Final clustering of the data using the Squeezer algorithm

The next method used to the cluster the data was the ROCK (Guha et al. 2000) clustering algorithm. An initial estimate of the number of clusters must be inputted for the algorithm, and as with the *k*-modes algorithm, an initial cluster estimate of 2 was

used. The first step in the algorithm was to define the neighbors of each observation. Two observations were considered neighbors if $sim(p_i, p_j) \geq \theta$ (Guha et al. 2000). Recall the similarity function, $sim(T_i, T_j) = \frac{|T_i \cap T_j|}{|T_i \cup T_j|}$, may only take on values between 0 and 1 so the threshold for the analysis is also restricted to be between 0 and 1 (Guha et al. 2000). The threshold was set at θ = 0.50 as was done by Guha et al. (2000). Therefore, if at least half the attributes between two observations match, the observations are deemed neighbors. The similarities were calculated between all pairs of observations in the data and the pairs with similarities larger than 0.50 were deemed neighbors. Table 4.2 gives each observation and its corresponding neighbors.

Observation	**Neighbors**
(1,1)	(1,2) (1,3) (2,1) (2,2) (2,3) (3,1)
(1,2)	(1,1) (1,3) (2,1) (2,2) (2,3) (3,1)
(1,3)	(1,1) (1,2) (1,4) (2,4) (3,1)
(1,4)	(1,3) (2,4) (3,3) (3,4) (4,1) (4,2) (4,3) (4,4)
(2,1)	(1,1) (1,2) (2,2) (2,3) (3,1)
(2,2)	(1,1) (1,2) (2,1) (2,3) (3,1)
(2,3)	(1,1) (1,2) (2,1) (2,2) (3,1)
(2,4)	(1,3) (1,4) (3,3) (3,4) (4,1) (4,2) (4,3) (4,4)
(3,1)	(1,1) (1,2) (1,3) (2,1) (2,2) (2,3)
(3,2)	(3,3) (3,4) (4,1) (4,2) (4,3) (4,4)
(3,3)	(1,4) (2,4) (3,2) (3,3) (4,1) (4,2) (4,3) (4,4)
(3,4)	(1,4) (2,4) (3,2) (3,3) (4,1) (4,2) (4,3) (4,4)
(4,1)	(1,4) (2,4) (3,2) (3,3) (3,4) (4,2) (4,3) (4,4)
(4,2)	(1,4) (2,4) (3,2) (3,3) (3,4) (4,1) (4,3) (4,4)
(4,3)	(1,4) (2,4) (3,2) (3,3) (3,4) (4,1) (4,2) (4,4)
(4,4)	(1,4) (2,4) (3,2) (3,3) (3,4) (4,1) (4,2) (4,3)

Table 4.2: Neighbors based on similarity

The number of links (i.e. common neighbors) between each observation was computed and $g\left(C_i, C_j\right) = \frac{link\left(C_i, C_j\right)}{\left(n_i + n_j\right)^{1+2f(\theta)} - n_i^{1+2f(\theta)} - n_j^{1+2f(\theta)}}$ was used to determine which observations to merge at each step in the algorithm (Guha et al. 2000). The pair of observations with the largest $g\left(C_i, C_j\right)$ was merged. Guha et al. (2000) defined $f\left(\theta\right) = \frac{1-\theta}{1+\theta}$ and since the threshold was set to be 0.50, $f\left(\theta\right) = 0.33$ in this analysis.

Initially, each observation started as its own cluster, i.e. 16 clusters. After $g\left(C_i, C_j\right)$ was calculated for each pair of points, the observations which produced the largest value for $g\left(C_i, C_j\right)$ were (1,4) and (2,4), as well as (3,3), (3,4), (4,1), (4,2), (4,3), and (4,4). Each of these points had 7 links between them so $g\left(C_i, C_j\right) = \frac{7}{(1+1)^{1.66} - 1^{1.66} - 1^{1.66}} = 6.03$. Therefore, two multi-observation clusters were formed at this stage in the clustering process. One cluster had the two points (1,4) and (2,4) and the second had the other six observations mentioned above. The remaining 8 observations each formed their own cluster, for a total of 10 clusters after the first iteration. After six iterations of the ROCK algorithm (Guha et al. 2000) the data were grouped into two clusters as shown in Figure 4.4.

Clay Sun No	**Clay Sun No**	**Silt Sun No**	**Silt Shade No**
Clay Sun No	**Clay Sun No**	**Clay Sun No**	**Silt Shade No**
Clay Sun No	**Clay Shade Yes**	**Silt Shade Yes**	**Silt Shade Yes**
Silt Shade Yes	**Silt Shade Yes**	**Silt Shade Yes**	**Silt Shade Yes**

Figure 4.4: Final clustering of the data using the ROCK algorithm

As can be seen in Figures 4.2 – 4.4, the *k*-modes (Huang 1998) and ROCK (Guha et al. 2000) algorithms clustered the data differently into two clusters. It appears the driving force of the *k*-modes algorithm was the vegetation variable. This could be due to the ordering of the variables when finding Q_1 and Q_2 to determine the initial modes. Therefore, the *k*-modes algorithm may produce different clustering results based upon the initial modes used. The Squeezer algorithm (Zengyou 2002) clustered the data into three clusters. However, the only difference between the Squeezer (Zengyou 2002) and ROCK (Guha et al. 2000) algorithms was that the observation (3,2) was placed into a cluster by itself when clustering with the Squeezer algorithm while the observation was combined with the second cluster when using the ROCK algorithm.

Although the ROCK (Guha et al. 2000) algorithm used the notion of neighbors in the clustering process, the actual geographic location of the observations was not taken into account. The next section of this paper proposes a method which explicitly incorporates the spatial or geographic location of the categorical observations into the

clustering algorithm. Therefore, observations which have the same attribute values and are similar in location are merged to form clusters.

4.2.2 Spatially Cluster Dichotomous Categorical Variables

This section discusses the clustering of dichotomous categorical variables. The recorded responses of the observations will either be 0 or 1 depending on whether a specific attribute is absent or present. The clustering algorithm proposed maximizes the likelihood of the multivariate normal distribution at every step in the clustering process (hierarchical clustering). Similar to the Squeezer (Zengyou 2002) algorithm, each observation is considered to form its own cluster, resulting in n initial clusters. The likelihood between each possible pairing of clusters is computed and the pair which yields the largest likelihood is merged to form a new cluster. The resulting arrangement at this stage in the clustering process has one cluster with two observations, and the remaining n-2 clusters each have a single observation. Again, the likelihood for each possible pairing of clusters is calculated and the pair of clusters which yields the largest likelihood is merged. This process is repeated until all the observations have been grouped into a single cluster.

The spatial or geographic location of the observations is taken into account in the variance-covariance matrix of the multivariate normal distribution. The variance-covariance matrix is computed using any specific covariance function from which exponential, Gaussian and spherical are most common. The frequently used spherical covariance function is given by,

$$C(d) = \begin{cases} \sigma^2 \left\{ 1 - \frac{3}{2}\left(\frac{d}{a}\right) + \frac{1}{2}\left(\frac{d}{a}\right)^3 \right\} & if\ d \le a \\ 0 & if\ d > a \end{cases} \tag{4.1}$$

where d is the distance between two observations and a is the range of the variogram, which is the separation distance where an increase in distance no longer produces an increase in the average squared difference between pairs of observations (Cressie 1991, Isaaks & Srivastava 1989, Schabenberger & Gotway 2005). The Gaussian covariance function which works well with a small scale spatial structure is

$$C(d) = \sigma^2 e^{-\frac{3d^2}{a^2}} \tag{4.2}$$

and the exponential covariance function is

$$C(d) = \sigma^2 e^{-\frac{3d}{a}} \tag{4.3}$$

which works best when there is less spatial structure at small distances. The Gaussian and exponential covariance functions have a similar range a, but they are not strictly identical, as the range refers to the rate at which the covariance function asymptotically approaches the sill (Cressie 1991, Isaaks & Srivastava 1989, Schabenberger & Gotway 2005). Since the spherical covariance function is most common, the examples provided in this paper use that covariance function. Also, the nugget effect which accounts for discontinuities (which may be due to sampling error and/or variability at extremely small distances) near the origin of the variogram are assumed to be zero (Isaaks & Srivastava 1989).

The likelihood of the multivariate normal distribution can be written as

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{N\nu/2}\,|\boldsymbol{\Sigma}|^{1/2}} e^{-1/2(\mathbf{x}-\mathbf{u})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{u})} \tag{4.4}$$

where ν is the number of clustering variates and $N = n_1 + n_2 + \ldots + n_c$ (the sum of the number of observations which fall into each cluster) where *c* is the number of clusters.

$$\mathbf{x}' = \left(x_{111} \quad \cdots \quad x_{11n_1} \quad x_{121} \quad \cdots \quad x_{12n_1} \quad \cdots \quad x_{211} \quad \cdots \quad x_{21n_2} \quad x_{221} \quad \cdots \quad x_{cvn_c} \right)$$

where x_{ijk} is the variate value of the k^{th} observation for the j^{th} variate in the i^{th} cluster

$i = 1, \ldots, c$ where c is the number of clusters

$j = 1, \ldots, \nu$ where ν is the number of categorical variates observed

$k = 1, \ldots, n_i$ where n_i is the total number of observations in the i^{th} cluster

$\boldsymbol{\mu}' = \left(\mu_{11} \quad \cdots \quad \mu_{11} \quad \mu_{12} \quad \cdots \quad \mu_{12} \quad \cdots \quad \mu_{21} \quad \cdots \quad \mu_{21} \quad \mu_{22} \quad \cdots \quad \mu_{cv} \right)$ where μ_{ij} is the mean for each cluster variate combination – there are n_i μ_{ij}'s in the i^{th} cluster of the j^{th} variate

The variance-covariance matrix in Equation (4.4) is given by $\boldsymbol{\Sigma} = \bigoplus_{i=1}^{c} \boldsymbol{\Sigma}_i^*$ which assumes the same variable is uncorrelated with itself between clusters. $\boldsymbol{\Sigma}_i^*$ is the cross-covariance matrix between variates within a cluster and is computed using the spherical covariance function from Equation (4.1). It is also general practice to assume the variates are uncorrelated when dealing with categorical attributes. Therefore,

$$\boldsymbol{\Sigma}_i^* = \begin{bmatrix} \boldsymbol{\Sigma}_{i11} & 0 & \cdots & 0 \\ 0 & \boldsymbol{\Sigma}_{i22} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \boldsymbol{\Sigma}_{ivv} \end{bmatrix} = \left[\boldsymbol{\Sigma}_{ijj} \right] \tag{4.5}$$

and

$$\boldsymbol{\Sigma}_{ijj} = \begin{bmatrix} \sigma_{ijj}^2 & sph\left(d_{12}\right) & \cdots & sph\left(d_{1n_i}\right) \\ & \sigma_{ijj}^2 & \cdots & sph\left(d_{2n_i}\right) \\ & & \ddots & \vdots \\ & & & \sigma_{ijj}^2 \end{bmatrix} \tag{4.6}$$

where σ_{ijj}^2 is the sill of the j^{th} variate in the i^{th} cluster. The sill $\left(\sigma_{ij}^2\right)$ may be approximated by $p\left(1-p\right)$ where p = the proportion of 1s. $\boldsymbol{\Sigma}_{ijj}$ is a symmetric matrix because $d_{kk'}$ is the actual physical distance between observations at locations k and k', so $sph\left(d_{12}\right) = sph\left(d_{21}\right)$ (Isaaks & Srivastava 1989).

4.2.3 Choosing an Optimal Number of Clusters

The likelihood itself can be used to determine the optimal clustering scheme for a given set of data. A sharp increase in the plot of the likelihood values against the number of clusters would indicate an appropriate clustering scheme. Since the likelihood of the multivariate normal distribution is maximized at every step in the clustering process, an increase in the plot would indicate which clustering scheme(s) may be best.

An improvement over solely plotting the likelihood values against the number of clusters would be to use Akaike's Information Criterion (AIC) (Akaike 1974). This criterion also uses the likelihood which is computed using a covariance function, but penalizes for the number of parameters being estimated. Since the ultimate goal is to

maximize the likelihood, the parameter estimates are computed using maximum likelihood estimation (MLEs). The AIC is given by,

$$\text{AIC} = -2\log\left\{L\left(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}} \mid \mathbf{x}\right)\right\} + 2k \tag{4.7}$$

where k is the number of parameters estimated and $L\left(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}} \mid \mathbf{x}\right)$ is the estimated likelihood given the data. Therefore, a penalty is imposed for having more clusters, i.e. more parameters to estimate. Thus, smaller values for the AIC are better.

Within each cluster the range, sill and mean (assuming no nugget effect) must be estimated for each variate. Since the sill and mean are functions of p, only the range and p must be estimated for each variable resulting in 2ν estimated parameters, where ν is the number of dichotomous categorical variables. It is recommended that at least $2*2\nu$ observations are present in each cluster in order to estimate all the parameters. However, if there are not at least $2*2\nu$ observations in a cluster, the range is estimated by variate using the entire data set. Therefore, all the variate 1 observations (regardless of cluster) are used to estimate the range for variate 1, all the observations for variate 2 (regardless of cluster) are used to estimate the range for variate 2, etc. This way, all the clusters will have the same range for each variate, but the variance-covariance matrices for each cluster will still differ since different observations are used to calculate each. If there are at least five 1s and five 0s for each variate within a cluster p is estimated using the data in the cluster. If not, then p is estimated by variate using the entire data set.

4.2.4 Example

The spatial clustering algorithm described in Section 4.2.2 may not only be used to cluster the observations in a hierarchical manner, but also to evaluate various clustering schemes to determine which best fits the data. In this example, the multivariate normal distribution was used to evaluate the clustering schemes produced by the *k*-modes (Huang 1998), Squeezer (Zengyou 2002) and ROCK (Guha et al. 2000) algorithms, in Section 4.2.1 above, to determine which best fits the data when spatial location is explicitly taken into account in the clustering process. Since the responses used in the multivariate normal distribution are numeric (0s and 1s), the responses recorded on the observations needed to be converted. The soil type attributes were re-coded as clay = 1 and silt = 0. For the sunlight variable the attributes were re-coded as sun = 1 and shade = 0, and for vegetation no = 1 and yes = 0. Figure 4.5 summarizes the clustering results of the *k*-modes (Huang 1998) algorithm and the re-coded data.

Clay Sun No	Clay Sun No	Silt Sun No	Silt Shade No
Clay Sun No	Clay Sun No	Clay Sun No	Silt Shade No
Clay Sun No	Clay Shade Yes	Silt Shade Yes	Silt Shade Yes
Silt Shade Yes	Silt Shade Yes	Silt Shade Yes	Silt Shade Yes

(a) Original variables

1 1 1	1 1 1	0 1 1	0 0 1
1 1 1	1 1 1	1 1 1	0 0 1
1 1 1	1 0 0	0 0 0	0 0 0
0 0 0	0 0 0	0 0 0	0 0 0

(b) Re-coded variables

Figure 4.5: *k*-modes clustering results and the re-coded data

Similarly, Figures 4.6 and 4.7 summarize the clustering results from the Squeezer (Zengyou 2002) and ROCK (Guha et al. 2000) algorithms and the re-coded data.

Clay Sun No	Clay Sun No	Silt Sun No	Silt Shade No
Clay Sun No	Clay Sun No	Clay Sun No	Silt Shade No
Clay Sun No	Clay Shade Yes	Silt Shade Yes	Silt Shade Yes
Silt Shade Yes	Silt Shade Yes	Silt Shade Yes	Silt Shade Yes

(a) Original variables

1 1 1	1 1 1	0 1 1	0 0 1
1 1 1	1 1 1	1 1 1	0 0 1
1 1 1	1 0 0	0 0 0	0 0 0
0 0 0	0 0 0	0 0 0	0 0 0

(b) Re-coded variable

Figure 4.6: Squeezer clustering results and the re-coded data

Clay Sun No	Clay Sun No	Silt Sun No	Silt Shade No
Clay Sun No	Clay Sun No	Clay Sun No	Silt Shade No
Clay Sun No	Clay Shade Yes	Silt Shade Yes	Silt Shade Yes
Silt Shade Yes	Silt Shade Yes	Silt Shade Yes	Silt Shade Yes

(a) Original variables

1 1 1	1 1 1	0 1 1	0 0 1
1 1 1	1 1 1	1 1 1	0 0 1
1 1 1	1 0 0	0 0 0	0 0 0
0 0 0	0 0 0	0 0 0	0 0 0

(c) Re-coded variables

Figure 4.7: ROCK clustering results and the re-coded data

In Section 4.2.3 it was recommended that at least 12 $\left(2*2\nu = 2*\left(2*3\right)\right)$ observations should be present in each cluster to adequately estimate the range, sill and mean. Since this is not the case, all sixteen observations are used to estimate the range

and sill for each variate using the kriging function in ArcGIS (ESRI 2006). For example, the responses used to estimate the range and sill for soil type are $\mathbf{x}' = (1 \quad 1 \quad 0 \quad 0 \quad 1 \quad 1 \quad 1 \quad 0 \quad 1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)$. The mean is estimated using the data in each cluster since ArcGIS (ESRI 2006) was used to obtain the spatial parameter estimates and p is not estimated alone. Estimating the parameters in this fashion is also a valid technique, however increases the number of estimated parameters and requires a larger number of observations in each cluster.

The three clustering schemes were evaluated and the log-likelihood and AIC values are found in Table 4.3.

Algorithm	**Log-likelihood**	**AIC**
k-modes	-15.80	55.59
Squeezer	-15.08	60.16
ROCK	-14.32	52.64

Table 4.3: Spatial clustering results

Looking at the results, the *k*-modes (Huang 1998) algorithm had the smallest log-likelihood of the three clustering schemes evaluated. The Squeezer (Zengyou 2002) algorithm produced a log-likelihood slightly better than that of the *k*-modes (Huang 1998) algorithm, however the additional cluster resulted in a much larger AIC due to the increased number of estimated parameters. The ROCK (Guha et al. 2000) algorithm performed much better than the other two algorithms in terms of both log-likelihood and AIC. This does not seem all too surprising since the ROCK (Guha et al. 2000) algorithm uses the notion of neighbors in its clustering process.

4.2.5 Spatial Weighting

Since one of the goals of precision agriculture is to define areas with potential for differentiated treatments, targeted (guided) samples are taken to provide detailed information about the agronomic properties of the land. Due to the high cost of obtaining these samples, only a limited number of them may be taken and should come from relatively homogenous spatially contiguous areas of the field. Also, small patches of similar observations may not be suitable for the application of lime, fertilizers or other agriculture inputs. Therefore, the clusters should be formed to produce the most spatially contiguous clustering of the data as possible. Thus, it would be more beneficial to include small patches of dissimilar values into the surrounding larger clusters of similar value to minimize the cost associated with site-specific management. Since the differences in the response values and the cluster means are squared in the likelihood calculation, the spatial location of the observations does not have a strong effect on the likelihood function. Therefore, weighting the purely spatial component of the likelihood $\frac{1}{|\mathbf{\Sigma}|^{1/2}}$ will increase the spatial information used in the clustering process to produce more spatially contiguous clusters for management purposes. Thus, the multivariate normal distribution will become

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{Nv/2}|\mathbf{\Sigma}|^{W/2}} e^{-1/2(\mathbf{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \tag{4.8}$$

where W is the weighting factor.

Figure 4.8 gives a set of data with three dichotomous variables which have been grouped into two clusters: cluster 1 in light blue and cluster 2 in dark blue. Notice that

observation 1 (circled in the top left corner) appears to belong in cluster 2 when looking at the value for variate 1. However, the values for variates 2 and 3 are most similar to those in cluster 1. Therefore, the weight (if any) required to keep observation 1 in cluster 1 (where it is spatially contiguous) should be rather small since two of the three variates are identical to the other observations in the first cluster.

1 0 0	0 0 0	0 0 0	1 1 1	1 1 1
0 0 0	0 0 0	0 0 0	1 1 1	1 1 1
0 0 0	0 0 0	1 1 1	1 1 1	1 1 1
0 0 0	0 0 0	1 1 1	1 1 1	1 1 1
1 1 1	1 1 1	1 1 1	1 1 1	1 1 1

Figure 4.8: One variate is different

There were not enough observations present in either cluster to adequately estimate the cluster mean and the spatial parameters. Therefore, the cluster means were estimated using the data in each cluster and the spatial parameters were again estimated by variate using ArcGIS (ESRI 2006). The range and sill estimates computed for this configuration of the data (Figure 4.8) were used to evaluate the cluster schemes in Figures 4.9 and 4.10 as well since the proportion of 1s and 0s is about the same in those scenarios.

Figure 4.9 gives the case when only one variate is an identical match to the other cluster 1 observations. Therefore, it would seem that some weight would be required to

keep observation 1 in cluster 1 where it is spatially contiguous with the other observations.

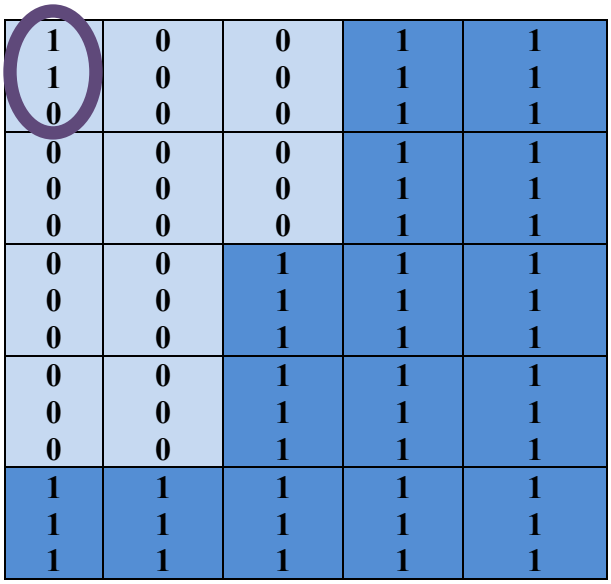

Figure 4.9: Two variates are different

Finally, the case when observation 1 has no matching attributes with the other cluster 1 observations is shown in Figure 4.10. This scenario would produce the largest weight needed since observation 1 is as dissimilar from the other cluster 1 observations as possible.

1 1 1	0 0 0	0 0 0	1 1 1	1 1 1
0 0 0	0 0 0	0 0 0	1 1 1	1 1 1
0 0 0	0 0 0	1 1 1	1 1 1	1 1 1
0 0 0	0 0 0	1 1 1	1 1 1	1 1 1
1 1 1	1 1 1	1 1 1	1 1 1	1 1 1

Figure 4.10: Three variates are different

As suspected, no additional weighting was required to keep observation 1 in cluster 1 with two matching attributes with the other observations in the cluster. A small weight was needed to keep observation 1 in cluster 1 when it only had one matching attribute with the other observations in the cluster. The largest weight required was needed when observation 1 had no matching attributes to the other cluster 1 observations. Hence, a larger weight is needed to ensure spatial contiguity when there is more dissimilarity between an observation and the other members of the cluster. The weights required are summarized in Table 4.4 below.

Number of Matching Attributes	Weight
2	0.15
1	4.63
0	9.12

Table 4.4: Weighting results for categorical attributes

4.2.6 Spatially Cluster Multinomial Categorical Variables

Not all categorical variables recorded are dichotomous, so the multinomial case must be addressed as well. In Section 4.2.2 the dichotomous categorical variables were re-coded to be either 0 or 1 depending on whether a certain attribute was absent or present. If a categorical variable had three attributes one might try re-coding them as 0, 1 and 2. However, this approach infers an order of importance in the attributes and implies that 0 and 2 are more different than 0 and 1 which is not usually the case. Therefore, another method to re-code the data must be employed. The data are re-coded as a set of dichotomous categorical variables. For example, suppose a soil type variable with three attributes (silt, clay and sand) has been recorded on three observations as seen in Table 4.5.

Observation	Attribute Recorded	Clay	Sand
1	Silt	0	0
2	Clay	1	0
3	Sand	0	1

Table 4.5: Re-coding of multinomial data

Since the soil type variable has three attributes, the observation values are re-coded into two dichotomous variables: clay and sand. Therefore, if both the sand and clay variables have values of 0 for an observation, the attribute recorded must have been silt as shown above. If a categorical variable has four attributes, three dichotomous variables are needed to summarize the original multinomial variable. Hence, $b-1$ dichotomous variables are needed to summarize one multinomial categorical variable with b attributes.

Once the data are re-coded, the multivariate normal distribution is used to cluster the observations in a hierarchical manner as discussed in Section 4.2.2. The multivariate normal distribution can be written as,

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{N\nu/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-1/2(\mathbf{x}-\mathbf{u})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{u})} \tag{4.9}$$

where ν is the number of dichotomous variates and $N = n_1 + n_2 + \ldots + n_c$ (the sum of the number of observations which fall into each cluster) where c is the number of clusters.

$$\mathbf{x}' = \left(x_{111} \quad \cdots \quad x_{11n_1} \quad x_{121} \quad \cdots \quad x_{12n_1} \quad \cdots \quad x_{211} \quad \cdots \quad x_{21n_2} \quad x_{221} \quad \cdots \quad x_{cvn_c} \right)$$

where x_{ijk} is the variate value of the k^{th} observation for the j^{th} variate in the i^{th} cluster

$i = 1, \ldots, c$ where c is the number of clusters

$j = 1, \ldots, \nu$ where ν is the total number of dichotomous categorical variables

$k = 1, \ldots, n_i$ where n_i is the total number of observations in the i^{th} cluster

$\boldsymbol{\mu}' = \left(\mu_{11} \quad \cdots \quad \mu_{11} \quad \mu_{12} \quad \cdots \quad \mu_{12} \quad \cdots \quad \mu_{21} \quad \cdots \quad \mu_{21} \quad \mu_{22} \quad \cdots \quad \mu_{cv} \right)$ where μ_{ij} is the mean for each cluster variate combination – there are n_i μ_{ij}'s in the i^{th} cluster of the j^{th} variate

The variance-covariance matrix in Equation (4.9) is given by $\boldsymbol{\Sigma} = \bigoplus_{i=1}^{c} \boldsymbol{\Sigma}_i^*$ which assumes each variable is uncorrelated with itself between clusters. $\boldsymbol{\Sigma}_i^*$ is the cross-covariance matrix between variates within a cluster and is computed using the spherical covariance function from Equation (4.1). In this case, it may not be assumed that all the

variates within a cluster are uncorrelated. Those variates which were formed based upon a single multinomial variable are correlated with one another. Therefore, $\boldsymbol{\Sigma}_i^*$ is a block diagonal matrix where each block represents a single multinomial variable:

$$\boldsymbol{\Sigma}_i^* = \begin{bmatrix} \boldsymbol{\Sigma}_{i11} & 0 & \cdots & 0 \\ 0 & \boldsymbol{\Sigma}_{i22} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \boldsymbol{\Sigma}_{itt} \end{bmatrix}. \tag{4.10}$$

In Equation (4.10) t is the total number of multinomial categorical variables recorded on the observations and

$$\boldsymbol{\Sigma}_{ijj} = \begin{bmatrix} \boldsymbol{\Sigma}_{ij_1j_1} & \boldsymbol{\Sigma}_{ij_1j_2} & \cdots & \boldsymbol{\Sigma}_{ij_1j_{b-1}} \\ \boldsymbol{\Sigma}_{ij_2j_1} & \boldsymbol{\Sigma}_{ij_2j_2} & \cdots & \boldsymbol{\Sigma}_{ij_2j_{b-1}} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_{ij_{b-1}j_1} & \boldsymbol{\Sigma}_{ij_{b-1}j_2} & \cdots & \boldsymbol{\Sigma}_{ij_{b-1}j_{b-1}} \end{bmatrix} \tag{4.11}$$

where $b-1$ is the number of dichotomous variates formed based on the multinomial categorical variable j with b attributes. $\boldsymbol{\Sigma}_{ij_sj_{s'}}$ $\left(s, s' = 1, \ldots, b-1\right)$ is the cross-covariance matrix between the dichotomous variates s and s' of the multinomial variable j. $\boldsymbol{\Sigma}_{ij_sj_{s'}}$ is computed using Oliver's (2003) method for finding the cross-covariance between two variates. First, the variance-covariance matrix $\boldsymbol{\Sigma}_{ij_sj_s}$ is computed and a Cholesky decomposition is performed, resulting in $\boldsymbol{\Sigma}_{ij_sj_s} = \mathbf{L}_{ij_s}\mathbf{L}_{ij_s}^{'}$. Similarly, a Cholesky decomposition is performed on the variance-covariance matrix $\boldsymbol{\Sigma}_{ij_{s'}j_{s'}}$, resulting in $\boldsymbol{\Sigma}_{ij_{s'}j_{s'}} = \mathbf{L}_{ij_{s'}}\mathbf{L}_{ij_{s'}}^{'}$. The decomposed variates $\left(\mathbf{L}_{ij_s}, \mathbf{L}_{ij_{s'}}\right)$ are used to compute

$\boldsymbol{\Sigma}_{ij_s j_{s'}} = \rho \mathbf{L}_{ij_s} \mathbf{L}'_{ij_{s'}}$ where ρ is the correlation between variates s and s' (Equation (4.12)) of the multinomial distribution:

$$\rho = -\sqrt{\frac{p_s p_{s'}}{(1-p_s)(1-p_{s'})}}. \tag{4.12}$$

In Equation (4.12) p_s and $p_{s'}$ are the proportion of 1s observed for the dichotomous variates s and s' of the multinomial variable j.

For each multinomial variable, the range, sill and mean (assuming no nugget effect) must be estimated for each of the $b-1$ dichotomous variates formed. However, the sill and mean are functions of p_s, which means that only p_s needs to be estimated for each dichotomous variable. Also, the correlation between each pair of dichotomous variates must be estimated, but since the correlation is comprised of p_s and $p_{s'}$ no additional parameters need to be estimated. Therefore, only the range and p_s are estimated for each dichotomous variable resulting in $2*(b-1)$ estimated parameters for each multinomial variable j. Since there are t total multinomial variables in a cluster, $\sum_{j=1}^{t}\left(2*\left(b_j-1\right)\right)$ parameters are estimated for each. Thus, it is recommended that at least $2*\sum_{j=1}^{t}\left(2*\left(b_j-1\right)\right)$ observations are present in each cluster to adequately estimate all the parameters using the data. If there are fewer than $2*\sum_{j=1}^{t}\left(2*\left(b_j-1\right)\right)$ observations present in each cluster, the range is estimated by variate using the entire data set and p_s is

estimated using the data in the cluster if there are at least five 1s and five 0s for each dichotomous variable. If not, then p_s is estimated by variate using the entire data set.

4.3 Clustering Categorical and Numeric Data

4.3.1 Background

There have been various approaches to clustering data which are comprised of both numeric and categorical (mixed-type) attributes. One of these algorithms is *k*-prototypes (Huang 1997, Huang 1998), a combination of the *k*-means and *k*-modes algorithms which are used to cluster purely numeric and purely categorical attributes, respectively. Recall that the main goal of any *k* type clustering is to group the data into *k* clusters which minimize a given cost function. The cost function to be minimized when using mixed-type data is found in Equation (4.13)

$$E = \sum_{l=1}^{k} \sum_{i=1}^{n} y_{il} d\left(X_i, Q_l\right) \tag{4.13}$$

where $Q_l = \left[q_{l1}, q_{l2}, \ldots, q_{lm}\right]$ is a representative vector for the cluster *l*, y_{il} is a member of a partition matrix and *d* is the similarity matrix (Huang 1997). Huang (1997) defines the similarity matrix as $d\left(X_i, Q_l\right) = \sum_{j=1}^{m_r} \left(x_{ij}^r - q_{lj}^r\right)^2 + \gamma_l \sum_{j=1}^{m_c} \delta\left(x_{ij}^c, q_{lj}^c\right)$ where m_r and m_c are the number of numeric and categorical attributes respectively, x_{ij}^r and q_{lj}^r are the numeric values, x_{ij}^c and q_{lj}^c are the categorical values, γ_l is a weight on the categorical values, and $\delta\left(p, q\right) = \begin{cases} 0 & p = q \\ 1 & p \neq q \end{cases}$ (Huang 1997, Huang 1998). Thus, the similarity function is a

combination of some distance function on the numeric values (Euclidean for example) and the number of mismatches between categorical attributes. Inserting the similarity function into Equation (4.13), Huang (1997 & 1998) defines the updated cost function as,

$$\begin{aligned} E &= \sum_{l=1}^{k}\left(\sum_{i=1}^{n} y_{il} \sum_{j=1}^{m_r} \left(x_{ij}^r - q_{lj}^r \right)^2 + \gamma_l \sum_{i=1}^{n} y_{il} \sum_{j=1}^{m_c} \delta\left(x_{ij}^c, q_{lj}^c \right) \right) \\ &= \sum_{l=1}^{k}\left(E_l^r + E_l^c \right) = \sum_{l=1}^{k} E_l^r + \sum_{l=1}^{k} E_l^c \\ &= E^r + E^c \end{aligned} \tag{4.14}$$

which is influenced partly by each type of variable.

As seen in Equation (4.14) a weight function γ_l has been introduced on the categorical component in the clustering process. Therefore, a weight must somehow be estimated for each cluster which can be problematic. There has not been any one solution for this problem at this point in time. Huang (1997 & 1998) suggested using the average standard deviation of either the numeric attributes in each cluster or of the overall dataset. If the average standard deviation for the numeric attributes in the entire data set is used, then the weights will be the same within each cluster. Huang (1997 & 1998) also suggests that the researcher's knowledge of the data may play an important role in determining just what the weight should be. Through simulations, Huang (1997 & 1998) determined that a small γ_l will result in more emphasis on the numeric attributes in the clustering process and a larger γ_l will result in more emphasis on the categorical attributes during the clustering process.

Another approach to clustering mixed-type attributes is proposed by Chiu et al. (2001). Each data point is read and a determination as to whether observations should be

merged with previously scanned points is made based upon some distance criterion. Once this initial scan of the data is completed, the clustered points which have formed dense regions in the data are stored in memory as a set of summary statistics. The next step in their algorithm treats the dense regions as individual points to be clustered in a hierarchical manner (Chiu et al. 2001). The log-likelihood used in this algorithm is

$$l = \sum_{j=1}^{J} \sum_{i \in I_j} \log p\left(x_i \mid \theta_j\right) = \sum_{j=1}^{J} l_{C_j} \tag{4.15}$$

where $p(\mathbf{x} \mid \theta)$ is the probability density function of $\mathbf{x}$ in the cluster C_j and θ_j are the model parameters (Chiu et al. 2001). The inner sum in Equation (4.15) represents the contribution of the j^{th} cluster to the log-likelihood. The model parameters for the clustering process are estimated using maximum likelihood.

The authors assume the numeric attributes are independent and normally distributed with mean μ_{jk} and variance σ_{jk}^2 while the categorical attributes are independent following a multinomial distribution with probability vector $\left(q_{jk1}, q_{jk2}, \ldots, q_{jkL_k}\right)$ where there are L_k categories (Chiu et al. 2001). Therefore, the log-likelihood may be written as the sum of the continuous attribute component and the categorical attribute component. Once the maximum likelihood estimates have been inserted, the log-likelihood then becomes $\hat{l} = \sum_{j=1}^{J} \left(\hat{l}_{AC_j} + \hat{l}_{BC_j}\right)$ (Chiu et al. 2001). Chiu et al. (2001) use a distance measure based on a decrease in the log-likelihood due to merging two clusters together as the clustering criteria.

The last approach for clustering mixed-type attributes to be discussed is an algorithm proposed by Simbahan and Dobermann (2006) and is comprised of four steps:

Step 1: Pre-processing of the data

Step 2: Initial allocation through spatially constrained cluster analysis

Step 3: Spatial aggregation and peripheral re-allocation of individuals

Step 4: Optional hierarchical or non-hierarchical classification to summarize individual patches in few classes for interpretive purposes

In the pre-processing of the data things such as standardization of the continuous variables, modeling of the semivariograms (both experimentally and theoretically) and construction of a nearest neighbor matrix may be completed (Simbahan & Dobermann 2006).

During the initial allocation stage of the algorithm, the observations are assigned to one of the *k* seed clusters based on a spatially weighted dissimilarity matrix. The Euclidean distance metric is usually used to assess the dissimilarity between the numeric observations, and does not work for categorical attributes. Thus, a weighted dissimilarity, $dc_{ik} = w_c^T \delta(c_i, \overline{c}_k)$ is computed between the categorical observations and the mode of the cluster $(\overline{c}_k)$ where $\delta(c_i, \overline{c}_k) = \begin{cases} 0 & c_i = \overline{c}_k \\ 1 & c_i \neq \overline{c}_k \end{cases}$ and w_c is a vector of weights which has been defined by the user. Therefore, the dissimilarity used combines both the numeric and categorical dissimilarities by $d_{ik} = dq_{ik} + dc_{ik}$ (Simbahan & Dobermann 2006).

However, d_{ik} can be modified by weighting the dissimilarity function to take into account the actual geographic separation between the observations. Thus, the dissimilarity between two observations at the i^{th} and l^{th} locations is $d_{il}^{*} = d_{il} * f\left(u_i - u_l\right)$ where d_{il} is the dissimilarity between the locations, f is the distance weighting function between the observations and u_i and u_l are the geographic locations (Simbahan & Dobermann 2006). Simbahan and Dobermann (2006) used the semivariogram as the weighting function for dq_{ik} (dissimilarity measure for the numeric variable) in their clustering algorithm. If there are a large number of numeric variables in the data, Simbahan and Dobermann (2006) suggest using principle component analysis to find the major principle components and calculate the semivariogram on those values rather than the actual numeric variables. Also, the multivariate variogram may be used to model the spatial correlation rather than using the semivariogram.

The spatially constrained cluster analysis in the initial allocation step minimizes the objective function $SS_w = \sum_{i=1}^{N}\sum_{k=1}^{N_k} y_{ik}\left(\left(x_i - \overline{x}_k\right)'\mathbf{G}\left(x_i - \overline{x}_k\right) + w_c^{'}\delta\left(c_i, \overline{c}_k\right)\right)$ where $\mathbf{G}$ is a diagonal matrix of the spatial weights for the numeric variables and N_k is the total number of clusters (Simbahan & Dobermann 2006). This clustering process continues until the desired number of clusters $\left(N_k\right)$ has been reached.

Once the set of clusters has been found, a random peripheral point is chosen and moved to its neighboring cluster. The objective function SS_w is recalculated to determine whether the point should remain in its original cluster or the current cluster. This

reallocation process continues until no peripheral point changes its cluster assignation. (Simbahan & Dobermann 2006)

This paper proposes a spatial clustering method to cluster both numeric and categorical attributes while taking into account the actual geographic or spatial location of the observations. The final clustering algorithm accounts for the spatial pattern of the data, the actual numeric responses and the categorical responses.

4.3.2 Spatially Cluster One Categorical and One Numeric Variable

Chapters 2 and 3 of this dissertation discussed how to cluster observations in which only numeric responses were recorded and Section 4.2.2 above discussed how to cluster observations with purely categorical responses. However, suppose it is of interest to cluster observations from a field based on soil type (clay or silt) and potassium; one response is categorical while the other one is numeric. Neither of the previous approaches alone would be appropriate to use in this case. Therefore, this section proposes a method to cluster observations with one dichotomous categorical and one numeric variable while taking into account the geographic or spatial location of the observations in the process.

In Section 4.2.2 the dichotomous categorical variables were re-coded in order to use the multivariate normal distribution in the clustering process. This again will be the case, which may cause a large discrepancy between the categorical and numeric response values. Therefore, each variate should first be standardized to ensure one variate does not weigh too heavily on the clustering process due solely to the nature of the responses

recorded. Once the responses have been standardized, the clustering process proceeds in a similar fashion. Each observation begins as a cluster resulting in n initial clusters. The likelihood for each possible pairing of clusters is calculated and the pair which yields the largest likelihood is merged to form a new cluster. Each possible pairing of the $n-1$ clusters is then evaluated and the pair which yields the largest likelihood is merged. This process continues until all the observations have been merged into a single cluster.

The geographic location of the observations is accounted for in the variance-covariance matrix of the likelihood using any specific covariance function. The categorical and numeric variables are assumed to be independent of one another so the variance-covariance matrix for each cluster is block diagonal. The variance-covariance matrix for each cluster is

$$\boldsymbol{\Sigma}_i^* = \begin{bmatrix} \boldsymbol{\Sigma}_{im} & 0 \\ 0 & \boldsymbol{\Sigma}_{ir} \end{bmatrix}, \tag{4.16}$$

where $\boldsymbol{\Sigma}_{im}$ represents the variability for the categorical attribute in the i^{th} cluster and $\boldsymbol{\Sigma}_{ir}$ represents the variability for the numeric attribute in the i^{th} cluster.

The multivariate normal distribution used for clustering is

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{Nv/2}\left|\boldsymbol{\Sigma}\right|^{1/2}} e^{-1/2(\mathbf{x}-\mathbf{u})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{u})}, \tag{4.17}$$

where the overall variance-covariance matrix is given by $\boldsymbol{\Sigma} = \bigoplus_{i=1}^{c} \boldsymbol{\Sigma}_i^*$ which assumes that the same variable is uncorrelated with itself between clusters. Also,

$$\mathbf{x}' = \left(x_{111}^m \quad \cdots \quad x_{11n_1}^m \quad x_{121}^r \quad \cdots \quad x_{12n_1}^r \quad \cdots \quad x_{211}^m \quad \cdots \quad x_{221}^r \quad \cdots \quad x_{22n_2}^r \quad \cdots \quad x_{cvn_c}^r \right)$$

where x_{ijk}^m is the categorical variate value of the k^{th} observation for the

j^{th} variate in the i^{th} cluster and $x^{r}_{i(j+1)k}$ is the numeric variate value of the k^{th} observation for the $(j+1)^{th}$ variate in the i^{th} cluster (if the distinction between categorical and numerical variates is not made then x^{m}_{ijk} and $x^{r}_{i(j+1)k}$ reduce to x_{ijk} as shown in Equation (4.4))

$i = 1, \ldots, c$ where c is the number of clusters

$j = 1, \ldots, v$ where v is the number of variates observed

$k = 1, \ldots, n_i$ where n_i is the total number of observations in the i^{th} cluster

$\boldsymbol{\mu}' = (\mu_{11} \cdots \mu_{11} \; \mu_{12} \cdots \mu_{12} \cdots \mu_{21} \cdots \mu_{21} \; \mu_{22} \cdots \mu_{cv})$ where μ_{ij} is the mean for each cluster variate combination – there are n_i μ_{ij}'s in the i^{th} cluster of the j^{th} variate

$N = n_1 + n_2 + \ldots + n_c$ where c is the number of clusters

Suppose four samples were taken from a field and the soil type (clay or silt) and the amount of potassium (K) in the soil (measured in mg/kg) were recorded. The data were then placed in a two cluster arrangement, based on expert opinion, with two observations in each cluster as shown in Table 4.6.

Cluster	Soil Type	Re-coded	K
1	Clay	1	40
1	Clay	1	43
2	Clay	1	80
2	Silt	0	72

Table 4.6: Data collected

It is of interest to evaluate the above clustering scheme using the likelihood, but the soil variable must first be re-coded: clay = 1 and silt = 0. The cluster 1 mean for the soil

variable is 1, whereas the mean for potassium is 41.5. Similarly, the Cluster 2 mean for the soil variable is 0.50 whereas the mean for the potassium variable is 76. There seems to be a rather large discrepancy in magnitude between the variables. Thus, each variable should be standardized before analysis.

The overall mean of the soil variable is 0.75 with a standard deviation of 0.50 and the overall mean for potassium is 58.75 with a standard deviation of 20.22. The standardized responses found in Table 4.7 are used in the analysis.

Cluster	Soil Type	K
1	0.50	-0.9273
1	0.50	-0.7789
2	0.50	1.0509
2	-1.50	0.6553

Table 4.7: Standardized response values

Now, the response vector used in the likelihood function is $\mathbf{x}' = \begin{pmatrix} 0.50 & 0.50 & -0.9273 & -0.7789 & 0.50 & -1.50 & 1.0509 & 0.6553 \end{pmatrix}$ and the mean for each cluster variate combination calculated using the standardized values produces $\boldsymbol{\mu}' = \begin{pmatrix} 0.50 & 0.50 & -0.8531 & -0.8531 & -0.50 & -0.50 & 0.8531 & 0.8531 \end{pmatrix}$.

Figures 4.11 and 4.12 below are plots of the cluster means for each variable. Notice in Figure 4.11 that the soil cluster means are small while the potassium cluster means are much larger. When the likelihood is calculated the difference in the observed responses and means for the potassium variable will be quite a bit larger and may dominate in the clustering process. Therefore, differences in the soil type (categorical) variable could be undetectable and the clustering process would be driven by the potassium (numeric) variable.

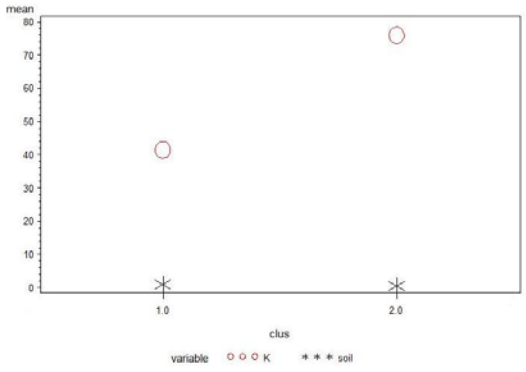

Figure 4.11: Cluster means for the non-standardized data

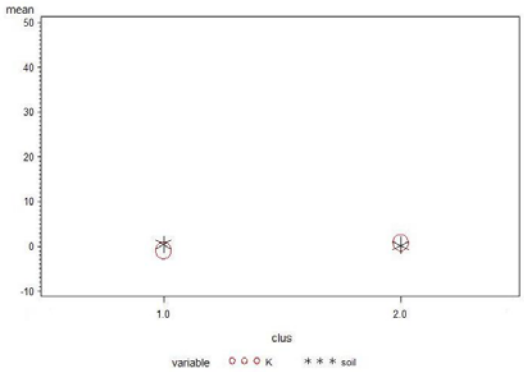

Figure 4.12: Cluster means for the standardized data

After the standardization process, the cluster means are much closer in magnitude as shown in Figure 4.12. Neither variable should dominate the clustering process based solely on the differences in the observed responses and the means. Therefore, a single variable has less chance of dominating the clustering process.

The variance-covariance matrix for each cluster $\boldsymbol{\Sigma}_i^* = \begin{bmatrix} \boldsymbol{\Sigma}_{im} & 0 \\ 0 & \boldsymbol{\Sigma}_{ir} \end{bmatrix}$ is a 4×4 matrix

where $\boldsymbol{\Sigma}_{im}$ is the 2×2 matrix

$$\boldsymbol{\Sigma}_{im} = \begin{bmatrix} \sigma_{i11}^2 & sph(d_{12}) \\ sph(d_{21}) & \sigma_{i11}^2 \end{bmatrix}, \tag{4.18}$$

and

$$\boldsymbol{\Sigma}_{ir} = \begin{bmatrix} \sigma_{i22}^2 & sph(d_{12}) \\ sph(d_{21}) & \sigma_{i22}^2 \end{bmatrix}. \tag{4.19}$$

In Equation (4.18) σ_{i11}^2 is the sill for the soil type attribute and $sph(d_{12})$ is the spherical covariance function between observations 1 and 2 in the cluster. Similarly, in Equation (4.19) σ_{i22}^2 is the sill for the potassium variable and $sph(d_{12})$ is the spherical covariance

function between observations 1 and 2 in the cluster. Both matrices, $\boldsymbol{\Sigma}_{im}$ and $\boldsymbol{\Sigma}_{ir}$, are symmetric since $d_{kk'}$ is the actual physical distance between observations at locations k and k', so $sph(d_{12}) = sph(d_{21})$ (Isaaks & Srivastava 1989).

The range, sill and mean (assuming no nugget effect) must be estimated for each variate (categorical and numeric) in each cluster. Therefore, each cluster would need at least 12 observations in order to carry out the estimation using the observations in the cluster. If this is not the case, only the mean for each cluster variate combination can be estimated using the data. The range and sill estimates are calculated using the entire data set. The categorical responses are combined to estimate their range and sill and then all the numeric responses are combined to estimate their range and sill values.

4.3.3 Spatially Cluster Multivariate Numeric and Multinomial Categorical Variables

In most situations, numerous multinomial categorical and multivariate numeric responses are recorded on the observations. Therefore, the spatial clustering algorithm must incorporate the clustering methods from Chapter 3 and Section 4.2.6 above. The multinomial categorical variables must first be re-coded into sets of dichotomous variables before the clustering process may begin. Once the multinomial variables have been re-coded, each variate (categorical and numeric) must be standardized as discussed in Section 4.3.2. Then the multivariate normal distribution may be written as,

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{Nv/2}\left|\boldsymbol{\Sigma}\right|^{1/2}} e^{-1/2(\mathbf{x}-\mathbf{u})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{u})} \tag{4.20}$$

$$\mathbf{x}' = \left(x^m_{111} \quad \cdots \quad x^m_{11n_1} \quad x^m_{121} \quad \cdots \quad x^m_{1l1} \quad x^r_{1(l+1)1} \quad \cdots \quad x^r_{1\nu n_1} \quad x^m_{211} \quad \cdots \quad x^r_{2\nu n_2} \quad \cdots \quad x^r_{c\nu n_c} \right)$$

where x^m_{ijk} is the categorical variate value of the k^{th} observation for the j^{th} variate in the i^{th} cluster and $j = 1,\ldots,l$ where l is the number of dichotomous categorical attributes

x^r_{ijk} is the numeric variate value of the k^{th} observation for the j^{th} variate in the i^{th} cluster and $j = (l+1),\ldots,\nu$ where ν is the total number of clustering variates

$i = 1,\ldots,c$ where c is the number of clusters

$k = 1,\ldots,n_i$ where n_i is the total number of observations in the i^{th} cluster

$\boldsymbol{\mu}' = \left(\mu_{11} \quad \cdots \quad \mu_{11} \quad \mu_{12} \quad \cdots \quad \mu_{12} \quad \cdots \quad \mu_{21} \quad \cdots \quad \mu_{21} \quad \mu_{22} \quad \cdots \quad \mu_{c\nu} \right)$ where μ_{ij} is the mean for each cluster variate combination – there are n_i μ_{ij}'s in the i^{th} cluster of the j^{th} variate

$N = n_1 + n_2 + \ldots + n_c$ where c is the number of clusters

The variance-covariance matrix in Equation (4.20) is given by $\boldsymbol{\Sigma} = \oplus_{i=1}^{c} \boldsymbol{\Sigma}^*_i$ which assumes that the same variable is uncorrelated with itself between clusters and will again be computed using any specific covariance function. Chiu et al. (2001) assume the categorical and numeric variables are independent which means $\boldsymbol{\Sigma}^*_i$ is block diagonal where

$$\boldsymbol{\Sigma}^*_i = \begin{bmatrix} \boldsymbol{\Sigma}_{im} & 0 \\ 0 & \boldsymbol{\Sigma}_{ir} \end{bmatrix} \tag{4.21}$$

In $\boldsymbol{\Sigma}_i^*$, $\boldsymbol{\Sigma}_{im}$ summarizes the variability for the categorical attributes in cluster i, and $\boldsymbol{\Sigma}_{ir}$ summarizes the variability for the numeric attributes in cluster i.

$\boldsymbol{\Sigma}_{im}$ is the matrix formulated in Equations (4.10), (4.11) and (4.12) discussed in Section 4.2.6 above, and

$$\boldsymbol{\Sigma}_{ir} = \begin{bmatrix} \boldsymbol{\Sigma}_{i(l+1)(l+1)} & \boldsymbol{\Sigma}_{i(l+1)(l+2)} & \cdots & \boldsymbol{\Sigma}_{i(l+1)\nu} \\ \boldsymbol{\Sigma}_{i(l+2)(l+1)} & \boldsymbol{\Sigma}_{i(l+2)(l+2)} & \cdots & \boldsymbol{\Sigma}_{i(l+2)\nu} \\ \vdots & & \ddots & \vdots \\ \boldsymbol{\Sigma}_{i\nu(l+1)} & \boldsymbol{\Sigma}_{i\nu(l+2)} & \cdots & \boldsymbol{\Sigma}_{i\nu\nu} \end{bmatrix} = \left[\boldsymbol{\Sigma}_{ijj'} \right]. \tag{4.22}$$

When $j = j'$ the cross-covariance matrix will be of the form

$$\boldsymbol{\Sigma}_{ijj} = \begin{bmatrix} \sigma_{ijj}^2 & sph\left(d_{12}\right) & \cdots & sph\left(d_{1n_i}\right) \\ & \sigma_{ijj}^2 & \cdots & sph\left(d_{2n_i}\right) \\ & & \ddots & \vdots \\ & & & \sigma_{ijj}^2 \end{bmatrix}. \tag{4.23}$$

When $j \neq j'$ the cross-covariance matrix will be comprised of the Cholesky decomposition from each variate, $\mathbf{L}_{ij}$ and $\mathbf{L}_{ij'}$, where $\boldsymbol{\Sigma}_{ijj} = \mathbf{L}_{ij}\mathbf{L}_{ij}^{'}$ and $\boldsymbol{\Sigma}_{ijj'} = \mathbf{L}_{ij'}\mathbf{L}_{ij'}^{'}$. Therefore,

$$\boldsymbol{\Sigma}_{ijj'} = \rho \mathbf{L}_{ij}\mathbf{L}_{ij'}^{'} \tag{4.24}$$

where ρ is the correlation between variates j and j'. Due to the complexity of the variables, a large number of observations must be present in each cluster to adequately estimate all the parameters using the data.

4.4 Conclusions

The first portion of this chapter incorporated dichotomous categorical variables into the spatial clustering algorithm. Since the dichotomous variables are non-numeric, the categorical responses were first re-coded into a set of numeric responses containing 0s and 1s. The actual geographic location of the categorical observations can be incorporated into the variance-covariance matrix of the multivariate normal distribution using any specific covariance function. The spherical covariance function is one of the most common and was chosen for this research. It is assumed that the categorical variables are uncorrelated, so the variance-covariance matrix is block diagonal. Weighting of the purely spatial component of the likelihood was also examined in order to ensure the spatial component was emphasized in the spatial clustering algorithm.

Not all categorical variables recorded are dichotomous so the spatial clustering algorithm was adapted for multinomial variables as well. The multinomial categorical variables must first be summarized by a set of dichotomous variables. The spherical covariance function was used to incorporate the spatial location into the variance-covariance matrix of the likelihood. However, the cross-covariance between the dichotomous variables for each multinomial variable must be taken into account. Oliver's (2003) method for finding the cross-covariance was implemented using the correlation from the multinomial distribution to calculate the cross-covariance.

The case when one dichotomous and one numeric variable have been recorded in each location was also incorporated into the spatial clustering algorithm. The clustering process assumes the two types of variables are independent of one another resulting in a

block diagonal covariance matrix with one component from the categorical variable and the other component from the numeric variable. Since the categorical variables are re-coded for the analysis, a discrepancy in the magnitude between the numeric and categorical responses could drive the clustering algorithm. To minimize these effects, the responses were standardized by variate prior to the clustering process.

Lastly, an extension of the spatial clustering algorithm was adapted to cluster both multinomial categorical variables as well as multivariate numeric variables. The multinomial categorical variables were first re-coded into a set of dichotomous categorical variables and then both the numeric and categorical variables were standardized before the clustering process began. The variance-covariance matrix is block diagonal with a categorical component and a numeric component for each cluster. Oliver's (2003) method for computing the cross-covariance was incorporated into both the numeric component and the categorical component of the variance-covariance matrix.

4.5 References

Akaike, H. 1974 A New Look at the Statistical Model Identification. *IEEE Transaction on Automatic Control*, AC 19, 716-723.

Barbará, Daniel, Couto, Julia & Li, Yi. 2002 COOLCAT: An entropy-based algorithm for categorical clustering. *Proceedings of the eleventh international conference on Information and knowledge management*. McLean, VA. 582-589.

Chiu, Tom, Fang, DongPing, Chen, John, Wang, Yao & Jeris, Christpoher. 2001 A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment. *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. San Francisco, CA. 263-268.

Cressie, N. 1991 *Spatial Statistics for Spatial Data*. New York: John Wiley & Sons, Inc.

ESRI. 2006 *ArcGIS Desktop Help 9.2*. Version 9.2. ESRI, Inc. Redlands, CA

Ganti, Venkatesh, Gehrke, Johannes, & Ramakrishman, Raghu. 1999 CACTUS-Clustering Categorical Data Using Summaries. *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. San Diego, CA. 73-83.

Gower, J. C. & Ross, G. J. S. 1969 Minimum Spanning Trees and Single Linkage Cluster Analysis. *Applied Statistics*. Vol. 18, No. 1, 54-64.

Guha, Sudipto, Rastogi, Rajeev & Shim, Kyuseok. 2000 ROCK: A Robust Clustering Algorithm for Categorical Attributes. *Information Systems*. Vol. 25, No. 5, 345-366.

Huang, Zhexue. 1997 Clustering Large Data Sets with Mixed Numeric and Categorical Values. *Proceedings of the First Pacific Asia Knowledge Discovery and Data Mining Conference*. Singapore: World Scientific, 21-34.

Huang, Zhexue. 1998 Extensions to the *k*-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery*. Vol. 2, No. 3, 283-304.

Isaaks, E. H. & Srivastava, R. M. 1989 *An Introduction to Applied Geostatistics*. New York: Oxford University Press, Inc.

Oliver, D. S. 2003 Guassian Cosimulation: Modelling of the Cross-Covariance. *Mathematical Geology*. Vol. 35, No. 6, 681-698.

Schabenberger, O. & Gotway, C. A. 2005 *Spatial Methods for Spatial Data Analysis*. New York: Chapman & Hall/CRC Press.

Simbahan, Gregorio & Dobermann, Achim. 2006 An algorithm for spatially constrained classification of categorical and continuous soil properties. *Geoderma*. Vol. 136, Issues 3-4, 504-523.

Zengyou, HE, Xiaofei, XU & Shengchun, DENG. 2002 Squeezer: An Efficient Algorithm for Clustering Categorical Data. *Journal of Computer Science and Technology*. Vol. 17, No. 5, 611-624.

Chapter 5

Conclusions

This dissertation includes three papers that discuss incorporating the actual geographic location of an observation into the clustering process, known as spatial clustering. The spatial clustering algorithm proposed is a hierarchical clustering method which maximizes the multivariate normal distribution at every step in the clustering process. The geographic location of the observations was explicitly taken into account in the variance-covariance matrix of the multivariate normal distribution. In this research, the variance-covariance matrix was computed using the spherical covariance function. Therefore, pairs of observations which yielded a larger spherical covariance function value were close to one another in geographic location, and pairs of observations which produced a smaller spherical covariance function value were farther apart in geographic location.

Since the spatial clustering algorithm produces numerous possible clustering of the data, Akaike's Information Criterion (Akaike 1974) was one of the methods used to determine the appropriate number of clusters that fit the data. This criterion also uses the estimated likelihood, whose variance-covariance matrix was computed using the spherical covariance function, while at the same time penalizing for the number of estimated parameters which is directly proportional to the number of clusters in which the data are placed.

Chapter 2 looked at the case when only one numeric response was recorded on the observations. Therefore, the spatial clustering algorithm clustered the points based on the

recorded response and the geographic location of the observations. The likelihood may be used not only to cluster the observations in a hierarchical manner, but also to evaluate clustering arrangements created based upon expert opinion as well. Two examples were provided which utilized the likelihood as an evaluation tool of different clustering schemes proposed. Therefore, the likelihood and AIC were used to determine which proposed clustering scheme was the best fit for the data. The first example was a simulation study, while the second example analyzed pH readings from a 23-ha field in Kansas. Experts used their knowledge of precision agriculture to form the clusters of observations. The algorithm then evaluated these "expert" clusters.

Since the goal of the spatial clustering algorithm is to incorporate the geographic location of the observations, weighting of the purely spatial component of the multivariate normal distribution was also investigated in this chapter. The weighting allowed the spatial component to play a larger role in the clustering process. It was found that as the range increased, the effect of the spatial component in the likelihood decreased making the spatial locations less influential in the clustering algorithm. As the sill increased, the effect of the spatial component in the likelihood increased making the spatial locations more influential in the clustering algorithm.

Chapter 3 extended the spatial clustering algorithm to account for more than one numeric response, i.e. the multivariate case. The main challenge of this chapter was to model the cross-covariance matrix between response variables while taking into account the spatial structure between them. This was done using Oliver's (2003) approach for computing the cross-covariance matrix using the variance-covariance matrix for each

numeric response, as well as its Cholesky decomposition. The variance-covariance matrix for each numeric response was computed using the spherical covariance function so the actual geographic location of the observations was accounted for in the clustering process. A simulation study using the likelihood to evaluate proposed clustering schemes was conducted.

Categorical responses were incorporated into the spatial clustering algorithm in Chapter 4. First, the case when only dichotomous categorical responses are recorded on the observations was examined. Since the responses recorded are categorical (i.e. non-numeric), they must first be re-coded into numeric responses. Therefore, the response vector used in the multivariate normal distribution was comprised of 0s and 1s. It is assumed that the categorical variables were uncorrelated, so the variance-covariance matrix is a block diagonal matrix. The variance-covariance matrix was again computed using the spherical covariance function.

The simulated example presented a set of data which were clustered using the *k*-modes (Huang 1998), Squeezer (Zengyou 2002) and ROCK (Guha et al. 2000) clustering algorithms. The likelihood was used to evaluate which of the three clustering methods was best when the geographic location of the observations was explicitly taken into account. Weighting of the spatial component was also investigated in this chapter. It was found that the more dissimilar an observation is with the other members of the cluster, the larger the weight is needed to make certain spatial contiguity is dominant.

When multinomial categorical responses were recorded on the observations the variables had to first be re-coded into a set of dichotomous categorical variables. Since

each multinomial observation is summarized by a set of dichotomous variables, the dichotomous variates were no longer uncorrelated. Thus, the correlation between dichotomous variates for each multinomial variable was taken into account in the variance-covariance matrix. Again, Oliver's (2003) method was used to find the cross-covariance using the correlation coefficient for the multinomial distribution.

The case when one dichotomous categorical variable and one numeric variable were recorded on each observation was investigated. It was assumed that the two types of variables were uncorrelated with each another so the variance-covariance matrix was block diagonal with one component from the categorical variable and the other from the numeric variable. To make certain the nature of the responses did not overwhelm the clustering process, all responses were standardized by variable before clustering. Once the data have been standardized, the spatial clustering process proceeds as discussed.

Lastly, an extension of the spatial clustering algorithm was adapted to cluster both multinomial categorical variables as well as multivariate numeric variables. The multinomial categorical variables were first re-coded into a set of dichotomous categorical variables and then both the numeric and categorical variables were standardized before clustering. The variance-covariance matrix was block diagonal with a categorical component and a numeric component for each cluster. Oliver's (2003) method for computing the cross-covariance was incorporated into both the numeric component and the categorical component of the variance-covariance matrix as discussed in Chapters 3 and 4 respectively.

Since the log-likelihood was used only as an evaluation tool in this dissertation, implementation of the hierarchical clustering algorithm requires further investigation. The ultimate goal is to have a working software program to read in the data and cluster the observations hierarchically in an automated process.

Bibliography

Adamchuk, V. I. 2006 *Site-Specific Management Guidelines*. Potash & Phosphate Institute, in cooperation with the Foundation of Agronomic Research.

Adamchuk, V.I., D.B. Marx, A.T. Kerby, A.K. Samal, L.K. Soh, R.B. Ferguson, and C.S. Wortmann. 2007. Guided soil sampling for enhanced analysis of georefernced sensor-based data. In: Proceedings of the Ninth International Conference on Geocomputation 2007 Conference, Maynooth, Ireland, 3-5 September 2007, U. Demsar, ed. Maynooth, Ireland: NCG - National University of Ireland (E-proceedings, 4 pages).

Akaike, H. 1974 A New Look at the Statistical Model Identification. *IEEE Transaction on Automatic Control*, AC 19, 716-723.

Barbará, Daniel, Couto, Julia & Li, Yi. 2002 COOLCAT: An entropy-based algorithm for categorical clustering. *Proceedings of the eleventh international conference on Information and knowledge management*. McLean, VA. 582-589.

Bourgault, Gilles, Marcotte, Denis, & Legendre Pierre 1992 The Multivariate (Co)Variogram as a Spatial Weighting Function in Classification Methods. *Mathematical Geology*. Vol. 24, No. 5, 463-478.

Chiu, Tom, Fang, DongPing, Chen, John, Wang, Yao & Jeris, Christpoher. 2001 A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment. *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. San Francisco, CA. 263-268.

Cressie, N. 1991 *Spatial Statistics for Spatial Data*. New York: John Wiley & Sons, Inc.

Cuzick, J. & Edwards, R. 1990 Spatial Clustering for Inhomogeneous Populations. *Journal of the Royal Statistical Society*, Series B (Methodological), Vol. 52, No. 1, 73-104.

ESRI. 2006 *ArcGIS Desktop Help 9.2*. Version 9.2. ESRI, Inc. Redlands, CA

Everitt, B. 1974 *Cluster Analysis*. Toronto: Heinemann Educational Books Ltd.

Frogbrook, Z. L. & Oliver, M. A. 2007 Identifying management zones in agricultural fields using spatially constrained classification of soil and ancillary data. *Soil Use and Management*. 23, 40 – 51.

Ganti, Venkatesh, Gehrke, Johannes, & Ramakrishman, Raghu. 1999 CACTUS-Clustering Categorical Data Using Summaries. *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. San Diego, CA. 73-83.

Gower, J. C. 1971 A General Coefficient of Similarity and Some of its Properties. *Biometric*. Vol. 27, No. 4, 857 – 871.

Gower, J. C. & Ross, G. J. S. 1969 Minimum Spanning Trees and Single Linkage Cluster Analysis. *Applied Statistics*. Vol. 18, No. 1, 54-64.

Guha, Sudipto, Rastogi, Rajeev & Shim, Kyuseok. 2000 ROCK: A Robust Clustering Algorithm for Categorical Attributes. *Information Systems*. Vol. 25, No. 5, 345-366.

Hartigan, J. A. 1975 *Clustering Algorithms*. New York: John Wiley & Sons, Inc.

Huang, Zhexue. 1997 Clustering Large Data Sets with Mixed Numeric and Categorical Values. *Proceedings of the First Pacific Asia Knowledge Discovery and Data Mining Conference*. Singapore: World Scientific, 21-34.

Huang, Zhexue. 1998 Extensions to the *k*-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery*. Vol. 2, No. 3, 283-304.

Huang, Zhexue & Ng, K. 1999 A Fuzzy *k*-Modes Algorithm for Clustering Categorical Data. *IEEE Transactions on Fuzzy Systems*. Vol. 7, No. 4, 446-452.

Isaaks, E. H. & Srivastava, R. M. 1989 *An Introduction to Applied Geostatistics*. New York: Oxford University Press, Inc.

Johnson, D. E. 1998 *Applied Multivariate Methods for Data Analysis*. Pacific Grove: Brooks/Cole Publishing Company.

Johnson, R. A. & Wichern, D. W. 2002 *Applied Multivariate Statistical Analysis*. Upper Saddle River: Prentice-Hall, Inc.

Kaufman, L. & Rousseeuw, P. J. 1990 *Finding Groups in Data An Introduction to Cluster Analysis*. New York: John Wiley & Sons, Inc.

Kerby, A., D. Marx, S. Kachman, A. Samal, & V. Adamchuk. Spatial Clustering Using the Likelihood Function. In: Proceedings of the 8th Annual Hawaii International Conference on Statistics, Mathematics and Related Mathematics, Honolulu, Hawaii, 13-15 January 2009. Honolulu, Hawaii.

Kerby, A., D. Marx, A. Samal, and V. Adamchuck. 2008. Spatial clustering using the likelihood function. In: Proceedings of the Kansas State University Conference on Applied Statistics in Agriculture, Manhattan, Kansas, 27-29 April 2008. Manhattan, Kansas: KSU.

Kerby, A., D. Marx, A. Samal, and V. Adamchuk. 2007. Spatial clustering using the likelihood function. In: Proceedings of Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007), Omaha, Nebraska, 28-31 October 2007, 637-642, K. Anthony, H. Tung, and Q. Zhu, eds. Washington, DC: IEEE Computer Society.

Lee, Ickjai. 2005 Geospatial Clustering in Data-Rich Environments: Features and Issues. *Proceedings of the 9th International Conference on Knowledge-Based Intelligent Engineering Systems*. Australia, 336-342

Ng, R. T. & Han, J. 1994 *Efficient and Effective Clustering Methods for Spatial Data Mining*. Proceedings of the 20th VLDB Conference, Sanitago, Chile.

Oliver, D. S. 2003 Guassian Cosimulation: Modelling of the Cross-Covariance. *Mathematical Geology*. Vol. 35, No. 6, 681-698.

Oliver, M. A. & Webster, R. 1989 A Geostatistical Basis for Spatial Weighting in Multivariate Classification. *Mathematical Geology,* 21, 1, 15-35.

Press, W. Flannery, B. Teukolsky, S. & Vetterling, W. 1988 *Numerical Recipes in C The Art of Scientific Computing*. New York: Cambridge University Press.

R Development Core Team. 2007 *R: a language and environment for statistical computing*. Vienna, Austria. R Foundation for Statistical Computing.

SAS Institute. 2008 *SAS Online Doc.* Version 9.2. SAS Institute, Inc. Cary, NC.

Schabenberger, O. & Gotway, C. A. 2005 *Spatial Methods for Spatial Data Analysis*. New York: Chapman & Hall/CRC Press.

Schwarz, G. 1978 Estimating Dimensions of a Model. *Annals of Statistics,* 6, 461-464.

Simbahan, Gregorio & Dobermann, Achim. 2006 An algorithm for spatially constrained classification of categorical and continuous soil properties. *Geoderma*. Vol. 136, Issues 3-4, 504-523.

Zengyou, HE, Xiaofei, XU & Shengchun, DENG. 2002 Squeezer: An Efficient Algorithm for Clustering Categorical Data. *Journal of Computer Science and Technology*. Vol. 17, No. 5, 611-624.

APPENDIX

R program used to find the correlation which maximized the log-likelihood function.

```
#This is needed for the ginv function to find the generalized inverse of the variance-covariance matrix
library(MASS)

dat = read.table( )

#Definitions to be used in calculations
n = nrow(dat)/2
resp.vec = matrix(dat[ ,3], nrow = n*2, ncol = 1)
mean.vec = matrix(dat[ ,4], nrow = n*2, ncol = 1)
rng1 = 5
rng2 = 5
sill1 = 1
sill2 = 5

#Creating the distance matrix for variate 1
dist.mat1 = matrix(0, nrow = n, ncol = n)
for(i in 1:n){
  for(j in 1:n){
     dist.mat1[i,j] = sqrt(((dat[i,1] - dat[j,1])^2) + ((dat[i,2] - dat[j,2])^2))
     if(dist.mat1[i,j] > rng1) {dist.mat1[i,j] = rng1}
     dist.mat1[j,i] = dist.mat1[i,j]
  }
}

#Calculating the cubed distances used in the spherical covariance function
dist.mat1.3 = matrix(0, nrow = n, ncol = n)
for(i in 1:n){
  for(j in 1:n){
    dist.mat1.3[i,j] = dist.mat1[i,j]^3
    dist.mat1.3[j,i] = dist.mat1.3[i,j]
  }
}

#Calculating the variance-covariance matrix for variate 1
c1 = dist.mat1*-1.50*(sill1/rng1)
c2 = dist.mat1.3*0.50*(sill1/rng1^3)
c3 = matrix(sill1,n,n)
c4 = c1 + c2 + c3
L1 = chol(c4)
```

```
#Creating the distance matrix for variate 2
dist.mat2 = matrix(0, nrow = n, ncol = n)
for(i in 1:n){
  for(j in 1:n){
     dist.mat2[i,j] = sqrt(((dat[i,1] - dat[j,1])^2) + ((dat[i,2] - dat[j,2])^2))
     if(dist.mat2[i,j] > rng2) {dist.mat2[i,j] = rng2}
     dist.mat2[j,i] = dist.mat2[i,j]
  }
}

#Calculating the cubed distances for the spherical covariance function
dist.mat2.3 = matrix(0, nrow = n, ncol = n)
for(i in 1:n){
  for(j in 1:n){
    dist.mat2.3[i,j] = dist.mat2[i,j]^3
    dist.mat2.3[j,i] = dist.mat2.3[i,j]
  }
}

#Calculationg the variance-covariance matrix for variate 2
d1 = dist.mat2*-1.50*(sill2/rng2)
d2 = dist.mat2.3*0.50*(sill2/rng2^3)
d3 = matrix(sill2,n,n)
d4 = d1 + d2 + d3
L2 = chol(d4)

sig1 = c4
sig2 = d4

differ = resp.vec - mean.vec

#Function which will be optimzed in regrads to the correlation
log.like = function(rho, L1, L2, sig1, sig2, differ){
    sig12 = rho*t(L1)%*%L2
    sig21 = rho*t(L2)%*%L1
    sig.row1 = cbind(sig1,sig12)
    sig.row2 = cbind(sig21,sig2)
    sig = rbind(sig.row1,sig.row2)
    det.sig = det(sig)
    sig.inv = ginv(sig)
    -n*log(2*pi)-0.50*log(det.sig)-0.5*t(differ)%*%sig.inv%*%(differ)
    }

#Finding the maximum correlation
```

```
max.rho = optimize(log.like, lower = 0, upper = 1, maximum = TRUE, L1 = L1, L2 = L2,
sig1 = c4, sig2 = d4, differ = differ)
max.rho$maximum
```