

2007

Spatial Clustering Using the Likelihood Function

April Kerby

Winona State University, akerby@winona.edu

David Marx

University of Nebraska - Lincoln, david.marx@unl.edu

Ashok Samal

University of Nebraska - Lincoln, asamal1@unl.edu

Viacheslav Adamchuk

University of Nebraska - Lincoln, vadamchuk2@unl.edu

Follow this and additional works at: <http://digitalcommons.unl.edu/cseconfwork>



Part of the [Computer Sciences Commons](#)

Kerby, April; Marx, David; Samal, Ashok; and Adamchuk, Viacheslav, "Spatial Clustering Using the Likelihood Function" (2007). *CSE Conference and Workshop Papers*. 23.

<http://digitalcommons.unl.edu/cseconfwork/23>

This Article is brought to you for free and open access by the Computer Science and Engineering, Department of at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in CSE Conference and Workshop Papers by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

Spatial Clustering Using the Likelihood Function

April Kerby

*Department of Statistics
University of Nebraska-Lincoln
Lincoln, NE 68588-0115
akerby1@bigred.unl.edu*

David Marx

*Department of Statistics
University of Nebraska-Lincoln
Lincoln, NE 68588-0115
dmarx1@unl.edu*

Ashok Samal

*Department of Computer Science and
Engineering
University of Nebraska-Lincoln
Lincoln, NE 68588-0115
samal@cse.unl.edu*

Viacheslav Adamchuk

*Department of Biological Systems
Engineering
University of Nebraska-Lincoln
Lincoln, NE 68588-0115
vadamchuk2@unl.edu*

Abstract

Clustering has been widely used as a tool to group multivariate observations that have similar characteristics. However, there have been few attempts at formulating a method to group similar multivariate observations while taking into account their spatial location [12, 13, 14]. This paper proposes a method to spatially cluster similar observations based on their likelihoods. 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 can be computed using any specific spatial covariance structure. Therefore, observations within a cluster which are spatially close to one another will have a larger likelihood than those observations which are not close to one another. This results in spatially close observations being placed into the same cluster.

1. Introduction

Cluster analysis has been used as a tool to place similar observations in groups or clusters. Clusters are formed based on measures of similarity or dissimilarity. Observations are placed in clusters to maximize the similarity among observations within a cluster while at the same time maximizing the

dissimilarity to observations in other clusters [1, 2, 7, 8, 9].

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

$$d_{rs} = \sqrt{(x_r - x_s)'(x_r - x_s)} \quad (1)$$

standardized Euclidean distance,

$$d_{rs} = \sqrt{(z_r - z_s)'(z_r - z_s)} \quad (2)$$

and Mahalanobis distance

$$d_{rs} = \sqrt{(x_r - x_s)' \Sigma^{-1} (x_r - x_s)} \quad (3)$$

In Equations (1) and (3) above, x_r and x_s are multivariate observations. In Equation (2) z_r and z_s are the standardized observation values. Equation (3) uses Σ , the variance-covariance matrix between pairs of observations [1]. These distances can be used in a variety of hierarchical or nonhierarchical clustering methods. The hierarchical clustering methods place observations together in a nested sequence of clusterings. Nearest Neighbor and Hierarchical Tree Dendograms are popular forms of hierarchical clustering methods [1, 2].

These clustering methods do not allow one to account for spatial structure. 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 precision agriculture technology which has become an important aspect of agriculture production in recent years. Precision agriculture uses multiple data layers within

spatially variable observations to fine-tune cropping decisions. Since conventional coarse 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 is to delineate field areas with potential for differentiated treatments (management zones). The limited number of guided samples should be collected from homogenous areas of the field and away from the boundaries or locations where sensor data changes significantly over short distances. The soil samples should also uniformly cover the entire range of measurements, indicating spots of high, medium or low readings [3]. Therefore, a proper clustering method should be developed 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 is proposed to explicitly incorporate the spatial structure. This is accomplished by using likelihoods to form the clusters. The spatial structure is present as part of the variance-covariance matrix. That is, if two points are located far apart, their likelihood will be smaller than if the points were closer together.

2. Clustering using the likelihood function

The procedure proposed here maximizes the likelihood for the multivariate normal distribution at every step (hierarchical clustering). Initially, each observation will be considered to form its own cluster, resulting in n clusters. The likelihood is computed for each possible pairing of two “clusters”. The pair 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 pair which gives the largest likelihood is selected as the new merged cluster. This continues until there is only one cluster. The optimal number of clusters may be determined by plotting the likelihood against the number of clusters and looking for a sharp increase. This would indicate the appropriate number of clusters much like a dendrogram does.

To account for the spatial structure in the likelihood, the variance-covariance matrix is computed using any specific covariance function; exponential, Gaussian, or spherical are the most common. The spherical covariance function is

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

where d is the distance between two points and a is the range [4, 5, 6]. The Gaussian covariance function is

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

and the exponential covariance function is $C(d) = \sigma^2 e^{-\frac{3d}{a}}$. The Gaussian and exponential covariance functions have a similar range, a , but they are not strictly identical, as it refers to the rate at which the covariance function approaches the sill. Figure 1 compares the three covariance functions [4, 5, 6].

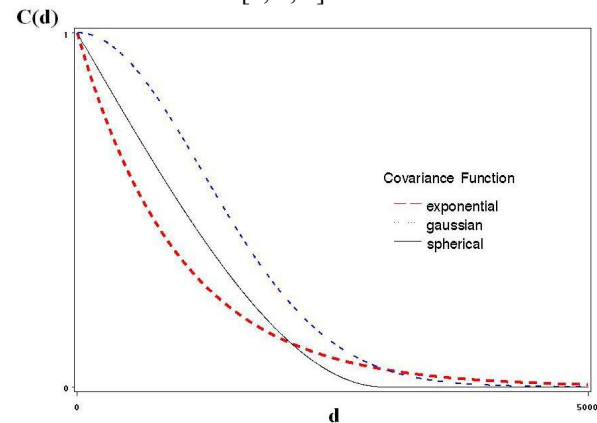


Figure 1. Comparison of covariance functions

The nugget effect is defined as the vertical jump from 0 at the origin to the variogram value at extremely small distances [4]. An example using only the spherical covariance function and assuming no nugget effect will be provided in this paper.

The likelihood of the multivariate normal distribution can be written as

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

where p is the number of variates, $\mathbf{x}' = (x_{11} \dots x_{1n_1} \ x_{21} \dots x_{cn_c})$ and \mathbf{x}_{jk} is the $(i, k)^{\text{th}}$ observation; $i = 1, \dots, c$ where c is the number of clusters and $k = 1, \dots, n_i$ where n_i is the total number of observations in the i^{th} cluster [2]. The mean vector is $\boldsymbol{\mu}' = (\mu_1 \dots \mu_1 \ \mu_2 \dots \mu_c)$ where μ_i is the mean from each cluster with $n_i \mu$'s for each cluster; $i = 1, \dots, c$. The variance-covariance

matrix, Σ , is given by $\Sigma = \mathbf{I}_c \otimes \Sigma_i$ where

$$\Sigma_i = \sigma_i^2 \begin{bmatrix} 1 & sph(d_{12}) & \cdots & sph(d_{1n_i}) \\ & 1 & \cdots & sph(d_{2n_i}) \\ & & \ddots & \vdots \\ & & & 1 \end{bmatrix} \quad (4)$$

$$= \sigma_i^2 [sph(d_{jk})]$$

In the variance-covariance matrix Σ_i is symmetric,

d_{jk} = spatial distance from x_j to x_k

and $sph(d_{jk}) = \sigma_i^2 \left(1 - 1.5 \left(\frac{d_{jk}}{a} \right) + 0.5 \left(\frac{d_{jk}}{a} \right)^3 \right)$ where

σ_i^2 = sill (for cluster i) = variance of the independent observations and a = range. Also, note that since d_{jk} is the spatial distance between x_j and x_k , $sph(d_{jk}) = sph(d_{kj})$ because $d_{jk} = d_{kj}$ [4].

Extending the likelihood to v variates, each observation would be x_{ijk} and

$\mathbf{x}' = (x_{111} \cdots x_{11n_1} \cdots x_{211} \cdots x_{cvn_c})$
 $= (x_{ijk}); i = 1, \dots, c$ where c is the number of clusters,
 $j = 1, \dots, v$ where v is the number of variates and $k = 1, \dots, n_i$ where n_i is the number of observations in the i^{th} cluster. The mean vector is

$\boldsymbol{\mu}' = (\mu_{11} \cdots \mu_{11} \cdots \mu_{21} \cdots \mu_{cv}) = (\mu_{ij});$
with n_i μ 's for each variate of each cluster. The

variance-covariance matrix becomes

$$\Sigma = \mathbf{I}_c \otimes \Sigma_i^* \text{ where}$$

$$\Sigma_i^* = \begin{bmatrix} \Sigma_{i11} & \Sigma_{i12} & \cdots & \Sigma_{i1v} \\ \Sigma_{i21} & \Sigma_{i22} & \cdots & \Sigma_{i2v} \\ \vdots & & \ddots & \vdots \\ \Sigma_{iv1} & \Sigma_{iv2} & \cdots & \Sigma_{ivv} \end{bmatrix} = [\Sigma_{ijj'}].$$

$\Sigma_{ijj'}$ (when $j = j'$) = Σ_i from before, but will differ for each i . $\Sigma_{ijj'}$ (when $j \neq j'$) will be of the same form as Equation (4). However, there will be a sill value from the first variate, σ_j^2 and one from the second variate, $\sigma_{j'}^2$. Therefore, in order to ensure $\Sigma_{ijj'}$ is positive definite the sill of the cross-covariance matrix can be no larger than $\sqrt{\sigma_j^2 \sigma_{j'}^2}$. Similarly, there will be two different range values for each variate as well,

say a_j and $a_{j'}$. The range used in computing $\Sigma_{ijj'}$ can be no larger than $\sqrt{a_j a_{j'}}$.

Also, it will be assumed that observations in different clusters are independent even though they may be next to each other spatially. If this assumption is not made, all the variance-covariance matrices would not change as clusters changed and the spatial structure would not add anything to the likelihood.

3. Optimal number of clusters

For determining the optimal number of clusters an improvement over plotting the likelihood against the number of clusters would be to use Akaike's Information Criteria (AIC) [10]. This criterion also uses the likelihood computed using a covariance function, while penalizing for the number of parameters being estimated. It is given by:

$$AIC = -2 \log \{ L(\hat{\boldsymbol{\mu}}, \hat{\Sigma} | \mathbf{x}) \} + 2k$$

where k is the number of parameters and $L(\hat{\boldsymbol{\mu}}, \hat{\Sigma} | \mathbf{x})$ is the estimated likelihood given the data. For each cluster there will be three parameters to estimate; sill, range, and mean (assuming no nugget effect). Therefore, a penalty will be imposed for having more clusters, i.e. more parameters to estimate. Thus, smaller AIC values are better. The AIC will be 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 just looking at the likelihood. Thus, both the likelihood and AIC values will be given in the examples, but the decisions will be made based solely on the AIC values. Although the goal of this paper is to cluster using multivariate data, our example will illustrate the univariate case which can be extended to the multivariate case as shown.

4. Example 1

The data for this example has been simulated to have no nugget effect, a sill value of 1, and a range of 20. A 10x10 grid was generated and the center 6x6 grid of the data was used. The smallest number of clusters results when all the data falls 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. Figures 2, 3, 4, and 5 show which clusters the points fall in when there are one, two, three, and four clusters respectively.

20.78	19.84	18.88	34.56	32.62	33.01
20.85	16.77	33.98	33.96	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	32.6	19.43	17.82
35.43	34.00	33.88	17.63	18.4	17.96

Figure 2. Data values as one cluster

A	A	A	B	B	B
A	A	B	B	B	B
A	B	B	B	B	B
B	B	B	B	B	A
B	B	B	B	A	A
B	B	B	A	A	A

Figure 3. Data in two clusters

A	A	A	B	B	B
A	A	B	B	B	B
A	B	B	B	B	B
B	B	B	B	B	C
B	B	B	B	C	C
B	B	B	C	C	C

Figure 4. Data in three clusters

A	A	A	B	B	B
A	A	B	B	B	B
A	B	B	B	B	B
D	B	B	B	B	C
D	B	B	B	C	C
D	B	B	C	C	C

Figure 5. Data in four clusters

Table 1 summarizes the AIC and likelihood values for a number of different cluster sizes.

Number of Clusters	Likelihood	AIC
1	5.34×10^{-49}	228.3
2	9.76×10^{-43}	205.47
3	9.91×10^{-43}	211.43
4	5.77×10^{-43}	218.52
33	3.28×10^{-47}	352.07
34	2.49×10^{-47}	354.62
35	1.90×10^{-47}	357.16
36	1.44×10^{-47}	359.71

Based on the results, the number of clusters with the highest likelihood value is three. However, the number of clusters with the lowest AIC is two. In this

case, even though three clusters had the highest likelihood value, the penalty for adding another cluster is enough to result in two clusters being the best fit for the data. Figures 6 and 7 show how the AIC and log-likelihood values change as a function of the number of clusters.

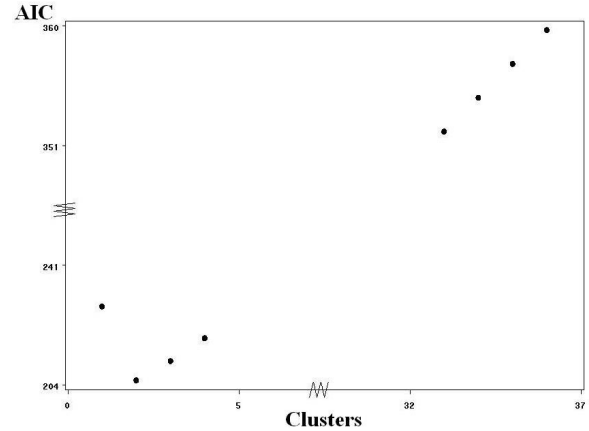


Figure 6. Plot of AIC values

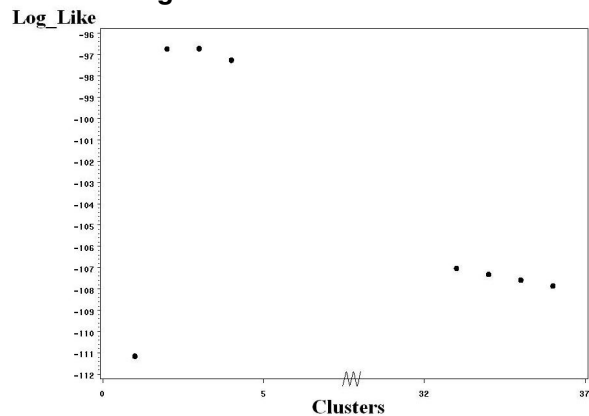


Figure 7. Plot of log-likelihood values

5. Example 2

The following example used a 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) [3]. The data layer used in this research was univariate (soil pH only). No nugget effect was assumed when estimating the parameters of the variogram. Therefore, only three parameters were estimated for each cluster; sill, range, and mean.

If there is no idea of what the clustering arrangement of the data should be, hierarchical clustering methods would be used. However, in this case experts not only used knowledge of the response variable, but other qualitative information as well. The clusters were assigned on the perceptions of what four individuals thought to be appropriate

management zones of the data in regards to pH and spatial location. The data are shown below in Figure 8.

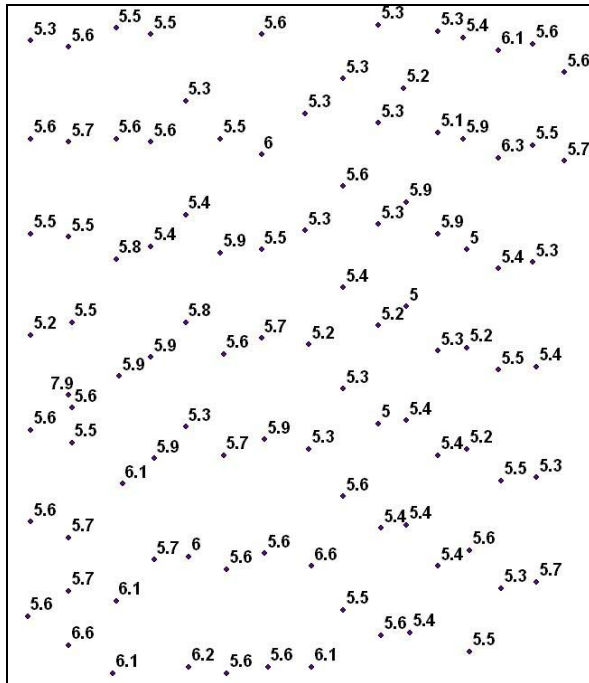


Figure 8. Data values

The data were broken into either three or four clusters with four illustrations of each. The three cluster examples were compared and the best was chosen based upon the likelihood as well as the AIC. Then the four cluster examples were compared and the best was chosen based on the likelihood and AIC. Finally, all eight variations were compared to see which example performed the best, that is which had the largest likelihood and the smallest AIC. The main goal was to see which example of the four would be better for each cluster size and then to determine whether three or four clusters would be more appropriate. Figures 9, 10, 11, and 12 show the illustrated examples for three clusters.

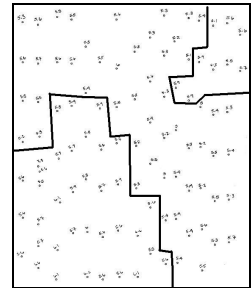


Figure 9. Variation 1

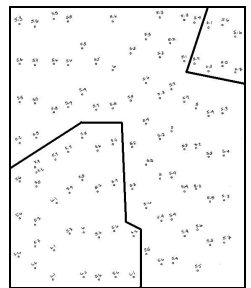


Figure 10. Variation 2

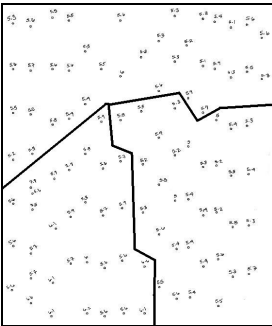


Figure 11. Variation 3

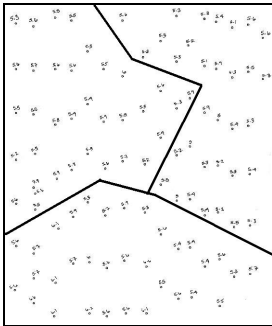


Figure 12. Variation 4

Table 2 summarizes the results of the three cluster analysis.

Table 2. Three cluster results

Variation	Likelihood	AIC
1	2.02×10^{-14}	81.06
2	1.01×10^{-18}	100.87
3	4.08×10^{-34}	171.76
4	3.56×10^{-17}	93.75

Figures 13, 14, 15, and 16 show the illustrated examples for four clusters.

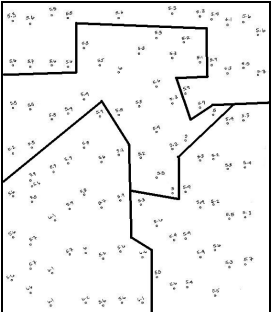


Figure 13. Variation 1

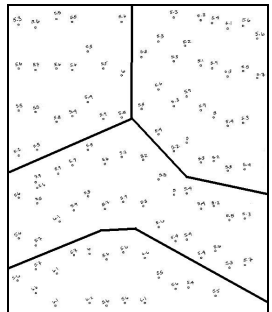


Figure 14. Variation 2

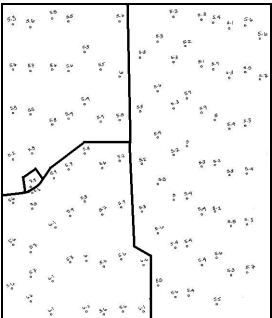


Figure 15. Variation 3

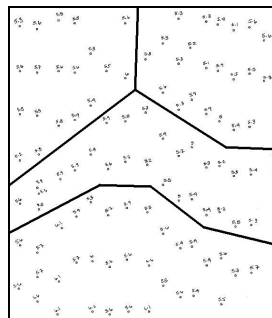


Figure 16. Variation 4

Table 3 summarizes the results of the four cluster analysis.

Table 3. Four cluster results

Variation	Likelihood	AIC
1	1.91×10^{-9}	64.15
2	7.51×10^{-14}	84.44
3	3.40×10^{-4}	37.98
4	1.59×10^{-16}	94.76

6. Summary and future work

Looking at Example 1, it can be seen that two clusters performed the best. The likelihood was 9.76×10^{-43} and the AIC was 205.47. Although the likelihood of 9.91×10^{-43} for three clusters was larger, due to the penalty of adding a cluster the AIC value of 211.43 was also larger. Thus, choosing two clusters is optimal.

When looking at the results from Example 2 and comparing the variations of three clusters, variation 1 had the largest likelihood, 2.02×10^{-14} and the smallest AIC, 81.06. When grouping the observations into four clusters, variation 3 performed the best. The likelihood was 3.40×10^{-4} and the AIC was 37.98. When determining whether three or four clusters would be more appropriate for the data, it appeared that four clusters was better. The likelihood computed with four clusters (3.40×10^{-4}) was larger than the likelihood for three clusters (2.02×10^{-14}). Also, the AIC was smaller, 37.98 compared to 81.06. Overall, variation 3 using four clusters best suited the data.

This paper only looks at the AIC as a possible way to assign a penalty for having a large number of clusters. Other information criteria will be explored, including Schwartz's Bayesian Information Criterion (SBC) which provides a larger penalty for more clusters [11].

We have shown 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. However, only the univariate case was considered in this paper. Therefore, the next step is to extend this work to the multivariate case. When looking at the multivariate case the spatial relationship between clusters of different variates must be taken into consideration. Once this is incorporated into the likelihood, the same approach as described in this paper may be taken.

After incorporating more than one variate into the likelihood, the ultimate goal will be to automate this process. The hopes are that a user can input the data and the program will systematically find the best possible clustering for the data.

7. References

- [1] D.E. Johnson, *Applied Multivariate Methods for Data Analysis*, Brooks/Cole Publishing Company, Pacific Grove, CA, 1998
- [2] R.A. Johnson and D.W. Wichern, *Applied Multivariate Statistical Analysis*, Prentice-Hall, Inc., Upper Saddle River, NJ, 2002
- [3] V.I. Adamchuk, D.B. Marx, A.T. Kerby, A.K. Samal, L.K. Soh, R.B. Ferguson, and C.S. Wortmann, *Guided Soil Sampling for Enhanced Analysis of Georeferenced Sensor-Based Data*, Geocomputation Conference, 2007
- [4] E.H. Isaaks and R.M. Srivastava, *An Introduction to Applied Geostatistics*, Oxford University Press, Inc., New York, NY, 1989
- [5] O. Schabenberger and C.A. Gotway, *Spatial Methods: for Spatial Data Analysis*, Chapman & Hall/CRC Press, New York, NY, 2005
- [6] N. Cressie, *Statistics for Spatial Data*, John Wiley & Sons, Inc., New York, NY, 1991
- [7] J.A. Hartigan, *Clustering Algorithms*, John Wiley & Sons, Inc., New York, NY, 1975
- [8] B. Everitt, *Cluster Analysis*, Heinemann Educational Books Ltd., London, 1974
- [9] L. Kaufman and P.J. Rousseeuw, *Finding Groups in Data An Introduction to Cluster Analysis*, John Wiley & Sons, Inc., New York, NY, 1990
- [10] H. Akaike, "A New Look at the Statistical Model Identification", *IEEE Transaction on Automatic Control*, AC 19, 1974, 716-723
- [11] G. Schwarz, "Estimating the Dimensions of a Model", *Annals of Statistics*, 6, 1978, 461-464
- [12] R. T. Ng and J. Han, *Efficient and Effective Clustering Methods for Spatial Data Mining*, Proceedings of the 20th VLDB Conference, Santiago, Chile, 1994
- [13] J. Cuzick and R. Edwards, "Spatial Clustering for Inhomogeneous Populations", *Journal of the Royal Statistical Society, Series B (Methodological)*, Vol. 52, No. 1, 1990, 73-104
- [14] G. C. Simbahan and A. Dobermann, "An algorithm for spatially constrained classification of categorical and continuous soil properties", *Geoderma*, Vol. 136, Issues 3-4, 2006, 504-523