Kansas State University Libraries

New Prairie Press

Conference on Applied Statistics in Agriculture

2008 - 20th Annual Conference Proceedings

SPATIAL CLUSTERING USING THE LIKELIHOOD FUNCTION

April Kerby

David Marx

Ashok Samal

Viacheslav Adamchuk

See next page for additional authors

Follow this and additional works at: <https://newprairiepress.org/agstatconference>

Part of the [Agriculture Commons](#), and the [Applied Statistics Commons](#)

This work is licensed under a [Creative Commons Attribution-Noncommercial-No Derivative Works 4.0 License](#).

Recommended Citation

Kerby, April; Marx, David; Samal, Ashok; and Adamchuk, Viacheslav (2008). "SPATIAL CLUSTERING USING THE LIKELIHOOD FUNCTION," *Conference on Applied Statistics in Agriculture*. <https://doi.org/10.4148/2475-7772.1100>

This is brought to you for free and open access by the Conferences at New Prairie Press. It has been accepted for inclusion in Conference on Applied Statistics in Agriculture by an authorized administrator of New Prairie Press. For more information, please contact cads@k-state.edu.

Author Information

April Kerby, David Marx, Ashok Samal, and Viacheslav Adamchuk

SPATIAL CLUSTERING USING THE LIKELIHOOD FUNCTION

April Kerby¹, David Marx¹, Ashok Samal² and Viacheslav Adamchuk³

¹Department of Statistics
University of Nebraska-Lincoln
Lincoln, NE 68583-0963

²Department of Computer Science and Engineering
University of Nebraska-Lincoln
Lincoln, NE 68588-0115

³Department of Biological Systems Engineering
University of Nebraska-Lincoln
Lincoln, NE 68583-0726

Abstract

Clustering has been widely used as a tool to group multivariate observations that have similar characteristics. However, most attempts at formulating a method to group similar multivariate observations while taking into account their spatial location are relatively ad hoc and do not account for the underlying spatial structure of the variables measured [12, 13, 14]. This paper proposes a method to spatially cluster similar observations based on the likelihood function. 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 each other. This results in observations which are similar and spatially close to one another being placed into the same cluster.

Key Words: spatial clustering, geostatistics, multivariate likelihood, spherical covariance

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; the three most prominent are Euclidean distance,

$$d_{rs} = \sqrt{(\mathbf{x}_r - \mathbf{x}_s)'(\mathbf{x}_r - \mathbf{x}_s)} \quad (1)$$

standardized Euclidean distance,

$$d_{rs} = \sqrt{(\mathbf{z}_r - \mathbf{z}_s)'(\mathbf{z}_r - \mathbf{z}_s)} \quad (2)$$

and Mahalanobis distance

$$d_{rs} = \sqrt{(\mathbf{x}_r - \mathbf{x}_s)'\Sigma^{-1}(\mathbf{x}_r - \mathbf{x}_s)} \quad (3)$$

In Equations (1) and (3) above, $\mathbf{x}_r$ and $\mathbf{x}_s$ are multivariate observations. In Equation (2) $\mathbf{z}_r$ and $\mathbf{z}_s$ 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. 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 [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 to reveal a more detailed description of the soil characteristics. One of the major challenges is to delineate field areas with potential for differentiated treatments (management zones). The limited number of samples that are collected should come from homogenous areas of the field and away from the boundaries or locations where sensor data changes significantly over short distances. Also, the soil samples should uniformly cover the entire range of measurements, indicating spots of high, medium or low readings [3]. This way, certain agronomic properties could be related to a linear or nonlinear combination of multiple sensor data layers. However, the area of applicability of such relationships may be reduced 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 and their spatial distribution.

In this paper a clustering method is proposed to explicitly incorporate the spatial structure by using the likelihood function to form the clusters. The spatial structure is present as part of the variance-covariance matrix of the likelihood function. 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 dendogram 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 given by,

$$C(d) = \begin{cases} \sigma^2 \left\{ 1 - \frac{3}{2} \left(\frac{d}{a} \right) + \frac{1}{2} \left(\frac{d}{a} \right)^3 \right\} & \text{if } d \leq a \\ 0 & \text{if } d > a \end{cases} \quad (4)$$

where d is the distance between two points and a is the range of the variogram [4, 5, 6]. The

range is the point at which an increase in distance no longer produces an increase in the average squared distance between pairs of values [4]. The Gaussian covariance function is

$$C(d) = \sigma^2 e^{-\frac{3d^2}{a^2}} \quad (5)$$

Finally, the exponential covariance function is

$$C(d) = \sigma^2 e^{-\frac{3d}{a}} \quad (6)$$

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 these covariance functions [4, 5, 6]. The nugget effect is also considered when doing spatial analyses. 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 [4]. 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

$$f(x) = \frac{1}{(2\pi)^{Nv/2} |\Sigma|^{1/2}} e^{-1/2(x-\mu)' \Sigma^{-1}(x-\mu)} \quad (7)$$

where v is the number of clustering variates and $N = n_1 + n_2 + \dots + n_c$, 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. $v = 1$)

- $\mathbf{x}' = (x_{11} \quad x_{12} \quad \dots \quad x_{1n_1} \quad x_{21} \quad \dots \quad x_{cn_c})$ where x_{ik} is the variate value of the $(i, k)^{th}$ observation
- $i = 1, \dots, c$ where c is the number of clusters
- $k = 1, \dots, n_i$ where n_i is the total number of observations in the i^{th} cluster
- $\mu' = (\mu_1 \quad \dots \quad \mu_1 \quad \mu_2 \quad \dots \quad \mu_c)$ where μ_i is the mean of each cluster – there are $n_i \mu$'s in each cluster

The variance-covariance matrix in equation (7) is given by $\Sigma = \bigoplus_{i=1}^c \Sigma_i$ where Σ_i is computed using the spherical covariance function from equation (4).

$$\Sigma_i = \begin{bmatrix} \sigma_i^2 & sph(d_{12}) & \dots & sph(d_{1n_i}) \\ & \sigma_i^2 & \dots & sph(d_{2n_i}) \\ & & \ddots & \vdots \\ & & & \sigma_i^2 \end{bmatrix} \quad (8)$$

This is a symmetric matrix because d_{ik} is the actual physical distance between observation units so $sph(d_{12}) = sph(d_{21})$ [4].

Extending the likelihood in equation (7) to the multivariate case (i.e. $v > 1$),

- $\mathbf{x}' = (x_{111} \quad \cdots \quad x_{11n_1} \quad x_{121} \quad \cdots \quad x_{12n_1} \quad \cdots \quad x_{211} \quad \cdots \quad x_{21n_2} \quad x_{221} \quad \cdots \quad x_{cvm_c})$
- where x_{ijk} is the variate value of the $(i, j, k)^{th}$ observation
- $i=1, \dots, c$ where c is the number of clusters
- $j=1, \dots, v$ where v is the number of variates observed
- $k=1, \dots, n_i$ where n_i is the total number of observations in the i^{th} cluster
- $\boldsymbol{\mu}' = (\mu_{11} \quad \cdots \quad \mu_{11} \quad \mu_{12} \quad \cdots \quad \mu_{12} \quad \cdots \quad \mu_{21} \quad \cdots \quad \mu_{21} \quad \mu_{22} \quad \cdots \quad \mu_{cv})$ where μ_{ij} is the mean for each cluster variate combination - there are $n_i \mu_{ij}$'s in each cluster

The variance-covariance matrix in equation (7) then becomes $\boldsymbol{\Sigma} = \bigoplus_{i=1}^c \boldsymbol{\Sigma}_i^*$ where $\boldsymbol{\Sigma}_i^*$ is the cross-covariance matrix between variates computed using the spherical covariance function.

$$\boldsymbol{\Sigma}_i^* = \begin{bmatrix} \boldsymbol{\Sigma}_{i11} & \boldsymbol{\Sigma}_{i12} & \cdots & \boldsymbol{\Sigma}_{i1v} \\ \boldsymbol{\Sigma}_{i21} & \boldsymbol{\Sigma}_{i22} & \cdots & \boldsymbol{\Sigma}_{i2v} \\ \vdots & & \ddots & \vdots \\ \boldsymbol{\Sigma}_{iv1} & \boldsymbol{\Sigma}_{iv2} & \cdots & \boldsymbol{\Sigma}_{ivv} \end{bmatrix} = [\boldsymbol{\Sigma}_{ijj'}] \quad (9)$$

When $j = j'$ the cross-covariance matrix $\boldsymbol{\Sigma}_{ijj'}$ will be the same as equation (8). When $j \neq j'$ the cross-covariance matrix will be found as referenced in Oliver 2003 [15]. There will be two sill values for each pair of variates j and j' ; one for the first variate denoted σ_j^2 and one for the second variate denoted $\sigma_{j'}^2$. Therefore, in order to ensure $\boldsymbol{\Sigma}_{ijj'}$ 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 a different range value for each variate, denoted a_j and $a_{j'}$, which 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, 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,

$$\text{AIC} = -2 \log \left\{ L(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}} | \mathbf{x}) \right\} + 2k \quad (10)$$

where k is the number of parameters and $L(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}} | \mathbf{x})$ 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 with no nugget effect, a sill of 1, and a range of 20. A 10×10 grid was generated and the center 6×6 grid of the data were used. The smallest number of clusters results when 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. Figures 2, 3, 4, and 5 show the clustering when there are one, two, three, and four clusters respectively.

Table 1 summarizes the AIC and likelihood values for a number of different cluster sizes. When the number of clusters is greater than the ability to adequately estimate the spatial parameters and mean, these estimates will be derived using the entire data set. When the number of clusters reaches $\frac{n}{3}$ or fewer, the parameter values can usually be estimated using the current data configuration. Based on these 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 optimal number of clusters to use. Figures 6 and 7 show how the AIC and log-likelihood values change as a function of the number of clusters.

5 Example 2

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) [3]. The data layer used in this research was univariate (soil pH only) as shown in Figure 8. No nugget effect was assumed when estimating the parameters of the covariance function. 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 process described above can 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 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, and table 2 summarizes the results of the three cluster analysis. Figures 13, 14, 15, and 16 show the illustrated examples for four clusters, and table 3 summarizes the results of the four cluster analysis.

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 the four cluster scheme 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 normally distributed data in 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.

Acknowledgements

The authors would like to thank the contribution of the University of Nebraska Agricultural Research Division, supported in part by funds provided through Hatch Act. Example data were provided by Veris Technologies, Inc. Additional support was granted through the Channing B. and Katherine W. Baker Fund #3424 of the University of Nebraska-Lincoln.

References

- [1] D.E. Johnson, *Applied Multivariate Methods for Data Analysts*, 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., Toronto, 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
- [15] D.S. Oliver, "Gaussian Cosimulation: Modelling of the Cross-Covariance," *Mathematical Geology*, Vol. 35, No. 6, August 2003.

Figures

Figure 1: Comparison of covariance functions

Figure 2: Example 1 - Data values as one cluster

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 3: Example 1 - 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	A
B	B	B	B	A	A
B	B	B	A	A	A

Figure 4: Example 1 - Data in three 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 5: Example 1 - Data in four 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 6: Example 1 - Plot of AIC values

Figure 7: Example 1 - Plot of log-likelihood values

Figure 8: Example 2 - Data values

Figure 10: Example 2 – Three Clusters - Variation 2

Figure 11: Example 2 – Three Clusters - Variation 3

Figure 15: Example 2 - Four Clusters - Variation 3

Table 1: Example 1 - Clustering results

Number Clusters	of	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

Table 2: Example 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

Table 3: Example 2 - 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