

# Optimization of Cluster Analysis Using Autocorrelation

Shing Lin  
Department of geography  
Southwest Texas State University  
[slin@swt.edu](mailto:slin@swt.edu)

## Abstract

Most current methods for measuring the number of clusters use optimum index values by comparing within-group and between-group variations at a global scale. An indicator based on local spatial variation to decide the existence of local scale cluster is needed. This study employed a new k-means algorithm for cluster analyses and geostatistical variograms for spatial autocorrelation. The criteria to optimize the numbers of clusters using global/local scale autocorrelation were discussed. Average annual temperature and precipitation from 1971 to 2000 at 350 weather stations throughout the State of Texas were used in two analyses. The first one investigated the relationship between cluster analysis and autocorrelation for a single variable - average spring temperature. Another analyzed multi-variant cluster and autocorrelation for average annual temperature and precipitation. The results suggest that global and local variograms can be used to indicate if further partitioning is needed within each cluster and in turn help to optimize the number of clusters. Future research is needed for the criteria to decide if the poorly fitted autocorrelation function data is due to the results of combination of smaller scales variograms or plain noises.

## 1. Introduction

Identifying the optimal number of clusters of hotspots is one of the most fundamental questions when using non-hierarchical (or partition) algorithm cluster analysis. There are three main approaches to optimize the number of clusters: distance-based such as k-means method, model-based such as probabilities distribution method, and density-based methods (Guo, 2002). However, two major drawbacks exist in all the methods for the purpose of cluster optimization. First, only the global-scaled clusters were verified. Various spatial scales of cluster analysis were not considered (Grubestic, 2002) and this can affect the optimization of clusters. Secondly, although the Gaussian distribution of model-based methods were proposed (Duta, Hart *et al.* 2000), the relationship between global and local perspectives for cluster detection and between global and local scales of spatial autocorrelation were not explored. If essential relationships exist between various scales of cluster detection and spatial correlation, the information of spatial autocorrelation can be used as a guideline to identify the number of clusters.

## 2. Background and Related Research

The works of Grubestic (2002) and Vesanto and Alhoniemi (2000) briefly describe two main approaches of cluster analysis: hierarchical algorithms and partition algorithms. Hierarchical

algorithms can be divided into agglomerate (bottom-up) and divisive algorithms (top-down) to build a hierarchical clustering tree and are used when an *a priori* number of clusters is known or easily guessed. Partition algorithms divide a data set into an arbitrary number of clusters by minimizing the error functions and are used when an *a priori* number of clusters is unknown such as crime hotspot cluster analysis. Murry and Estaville-Castro (1998) reviewed three partitioning algorithms. The first one is Observation Interaction Clustering Problem (OCIP) which employs the algorithm of minimizing total weighted difference in the assignment of observations to clusters. The second algorithm is Center Point Clustering Problem (CPCP). Instead of measuring between observation differences within clusters, a central point is assigned for cluster members; K-means is one of the examples. The last one was Median Cluster Problem (MCP) which defines cluster membership based on assigning observations to a representative observation.

There are three important application issues in cluster identification associated with these partitioning algorithms described by Murry and Estaville-Castro (1998). The first is that dynamic programming cannot solve the non-linear functions given the need for extensive computation time when the number of clusters is large. The use of other methods such as neural networks or heuristics may resolve this difficulty. For instance, Vesanto and Alhoneimi (2000) used two level neural network approach combining Self-Organization Map (SOM) and k-means to greatly reduce computation time.

The second problem with k-means algorithms is the membership function (or error function), which includes Euclidean distance ( $d_{ij}$ ) and weighting factor. Euclidean distance is a function of geographical space and observation association. Some k-means algorithms use square distance matrices ( $d_{ij}^2$ ) rather than distance matrices. Murry and Estaville-Castro (1998) defended the usage of distance matrices based on the superior performance of distance matrices when the data has spatial attributes and square distance matrices have greater impacts on outliers. However, Murry and Estaville-Castro used the attribute values as the weights of the membership function in their CPCP algorithm. Attribute values can be used to as the contributions to the estimated value of centroid, yet they cannot not reflect the properties of geostatistical anisotropy. Therefore, to use the attribute values as the weight of membership function is questionable.

The third problem of applying partition algorithms is defining the criteria to select the most appropriate number of clusters. Grubestic (2002) described discrepancies that can range up to 30% in the numbers of clusters when comparing the results from commercial products such as SAS, SPSS and SPlus. According to Grubestic, the most widely used indices to decide the number of clusters is Cubic Clustering Criteria (Sarle, 1983) or the Calinski and Harabasz index (1974). These methods deal with global rules based on complete data sets to seek an optimal index value to compare both within-group and between-group variations. It should be noted that some k-means algorithms as well as Cubic Clustering Criteria are mainly applied to spherical clusters (Sarle, 1983). In reality, geographical clusters are often related to regional rather than global rules and have highly elongated or irregular shapes.

Spatial autocorrelation, which refers to the degree of association of a variable in relation to its location or other attributes, can be used to investigate regional effects. The most widely used autocorrelations, such as Moran's I and Geary's G index, seek optimum index values by comparing within-group and between-group variation in a global scale; local scales of analysis

that may affect the correlation are not considered. Recognizing this limitation of global spatial correlation, some scholar such as Ratcliffe and McCullagh (1999) emphasized the local patterns influence by using Local Indicators of Spatial Association (LISA) statistics. They adopted the index of  $G_i$  and  $G_i^*$  proposed by Ord and Getis (1995). The problem with the Ord and Getis indicator is selecting a suitable distance  $d$  for the index  $G_i(d)$ . Ratcliffe *et al* (1999) attempted to derive the suitable distance from the intersection (crossing) of the curves of crimes per cell and the number of crimes. This approach has two main drawbacks: First, they did not explain why the cross point is an optimal indicator of distance. Second, the cross point can be arbitrary chosen depending on the units chosen.

### 3. Methodology

This study is an attempt to resolve two issues: First, to revise the existing cluster analysis algorithm so better cluster divisions can be calculated; second, to use distance-based spatial autocorrelation to verify the existence of clusters and in turn help to optimize the number of clusters. For cluster analysis, I revised Murry's algorithm into the following equation:

$$\text{Min } Z = \sum \sum |Z_i - Z_k| * d_{ik} \quad (1)$$

Where

$Z$  is the membership function

$Z_i$  is the attribute value of the measurement

$Z_k$  is the attribute value of the cluster centroid

$d_{ik}$  is the spatial Euclidean distance

The reason for adopting the attribute value differences, instead of the attribute values of the measuring points as Murry and Estaville-Castro (1998) pointed out in the original algorithm, was based on the assumption that the less value difference between the centroid and the measuring point, the greater the geostatistical similarity between the two. The same criteria can be applied to Euclidean distance which assumes the less distance between two measurement points the more similar they are. The algorithm of the program is described in the following:

1. Initialize *a priori* number of centroid points.
2. Use the membership function described above to assign the measuring point to one single cluster.
3. Average the value of all the points assigned to the same cluster to locate a new centroid for the new cluster.
4. Calculate the distance difference between old and new centroid.
5. Iterate the steps 3 and 4 until the distance differences converge to certain small value.

For autocorrelation, I applied geostatistics variograms to determine the global and regional spatial autocorrelation. Range of the variograms was used as an indicator since it is the spatial distance within which all the data are related to each other. However, variograms should be used cautiously. Atkinson and Tate (2000) pointed out that variograms are additive and some models

imply a combined distribution of spatial variation at different scales, as opposed to a single predominant scale of variation. They also distinguished the dual usages of 'scale'. The first is the scale of measurements or study areas. The second is the distance between observations and their support, i.e., the area or volume of a sample. They emphasized the importance of scale-dependent heterogeneity as a fundamental constraint on the comparison of multi-scaled phenomenon, such as those found in remote sensing and ecology. Webster and Oliver (1990) mirrored similar ideas and introduced nested variation and theory of nested sampling by using the 'reconnaissance variogram' to determine the general spatial scale of variation by employing several orders of magnitude for distance measurements for samples. The underlying theory is that different distances between sampling points could represent the stages in the hierarchy for a spatial property.

Two cluster and spatial analyses were conducted. The first analysis used a single variant - average spring temperature; the second used multi-variant - average annual temperature and precipitation. The data used in the study are the temperature and precipitation from 1971 to 2000 reported by the National Oceanic and Aeronautical Agency at 350 weather stations throughout the State of Texas. The revised k-means algorithm described previously was applied to the first analysis since the membership function involves both attribute value differences and distance. Multi-variant K-means clusters used Euclidian distance of temperature and precipitation as the membership function. Because the scales and the units of temperature and the precipitation are different, all the values of temperature and precipitation were standardized as input values in the multi-variant k-means algorithm. Both algorithms were implemented with Java programs I wrote. Spatial autocorrelation was conducted in ESRI ArcGIS geostatistical extension. To achieve a stable variogram, at least 50-100 points are needed to avoid a noisy variogram (Burrough and McDonnell, 1997). Therefore, the range of total number of clusters  $p$  was selected from 2 to 6 because when the total number of clusters is greater than 6, some of the clusters are more likely to contain less than 50 points.

## **4. Analyses and Results**

### **4.1 Single variant**

Texas spring average temperatures were used to conduct the cluster analysis and spatial correlation. Cluster analysis of real world temperature rendered distinct patterns that require geographical interpretation. When the number of clusters ( $p$ ) is low, such as 2 or 3, the cluster shows ellipsoidal or irregular shape (Figure 1). When  $p$  is 4 and up, the clusters become more spherical. Unlike hierarchical clustering algorithms, there are no distinct hierarchical relationships between clusters of different number of  $p$ . Also, unlike most clusters showing distinct boundaries, these results show interlaced boundaries for all numbers of  $p$ . For autocorrelation, initially ordinary kriging method was used to estimate the range of variogram. However, the results did not fit into any experimental geostatistical function. After examining the rough temperature map generated from the ordinary kriging, I refined the geostatistical variogram and kriging by removing the global trend of the temperature by using the method universal kriging provided in ARCGIS. The result showed a well-fitted spherical variogram with a major global range of about 3.5 h, which is about one-third of the width and

length of Texas (Figure 2), based on removing the first order trend. The refined average temperature map generated from universal kriging based on this spherical variogram showed not only smooth contours but also no outliers in any part of the map. On thing noteworthy is that at distance of zero, the value of semi-variance should be zero. But the model shows the value (or nugget) of 1.4. This nugget effect may indicate noises during measurement.

The length of range of this global variogram, which is about one-third of Texas, can be used as an indicator to decide if further partition into smaller cluster is needed within each cluster. Table 1 shows the results of this verification. When the number of cluster p is 5 or 6, the ratio of the each cluster needs to be broken down further is low (0.2) compare to when the p number is 2 and 3. Caution should be made that the range used here is a generalized global scale measurement. To be more precisely to determine if the partition within each cluster is needed, local variograms within each cluster should be generated and compared within each cluster.

### K-MEANS CLUSTERS RUN WITH DIFFERENT NUMBERS OF CLUSTER

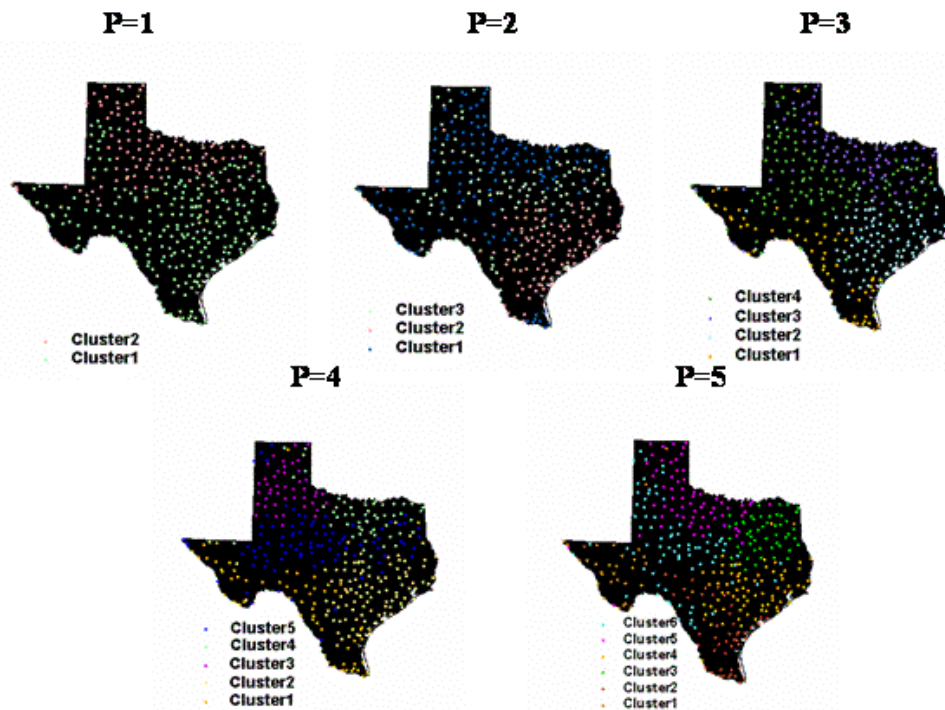


Figure 1 Results of k-means clusters at number of cluster p from 2 to 6.

## VARIOGRAM FROM UNIVERSAL KRIGING

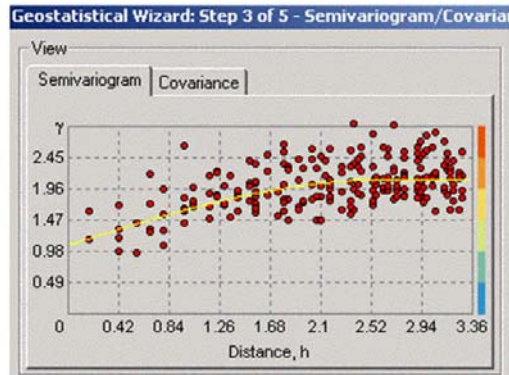


Figure. 2 Variogram created with universal kriging

Total Number of clusters P	Cluster number	Cluster partitioning are needed within cluster based on the value of range	Ratio of cluster needs to be broken up
2	2.1	Yes	1.0
	2.2	Yes	
3	3.1	Yes	0.66
	3.2	Yes	
	3.3	Maybe	
4	4.1	Maybe	0.5
	4.2	Yes	
	4.3	Maybe	
	4.4	Yes	
5	5.1	Maybe	0.2
	5.2	Yes	
	5.3	No	
	5.4	Maybe	
	5.5	No	
6	6.1	No	0.2
	6.2	No	
	6.3	Maybe	
	6.4	No	
	6.5	No	
	6.6	No	

Table 1 shows the ratio of each cluster number p needs to be broken up

## 4.2 Multi-variant

Annual average temperatures and precipitation were used to conduct the multi-variant cluster analysis and spatial correlation. Unlike the single variant k-means algorithm, which considers the relations of attribute values and spatial distance, the multi-variant k-means algorithm calculates the minimum distance at the domain of temperature and precipitation instead including the variable of spatial distance. After all the clusters were calculated in the statistical domain (Figure 3), each point within cluster was then projected onto the spatial domain (Figure 4).

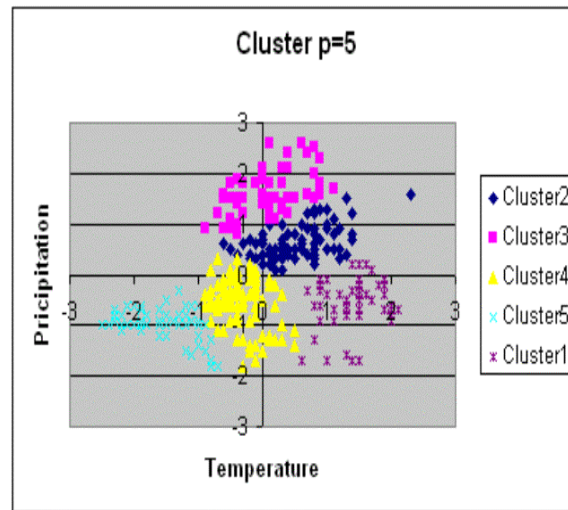


Figure 3 multi-variant k-means cluster  $p = 5$  renders on the statistical domain of temperature and precipitation. The values of both temperature and precipitation are normalized.

Unlike the shapes of temperature clusters, the shapes of multi-variant spatial clusters showed more ellipsoidal than elongated except when the number of cluster  $p$  is 2. The ellipsoid-shaped clusters are not only shown on the geographical domain but also show on the statistical domain. This phenomenon can be explained by the global trend of temperature and precipitation. Based on the maps generated from universal kriging, they indicate that there is an increasing trend from northwest to southeast for temperature, and increasing trend from west to east for precipitation. Unlike the temperature clusters, multi-variant clusters showed some hierarchical relationships between small and large numbers of clusters  $p$ . For instance, cluster 2 of  $p=1$  can be divided into cluster 4 and 5 of  $p=5$ ; cluster 1 of  $p=1$  can be divided into cluster 1, 2, 3 of  $p=5$ . The reason behind this phenomenon may be related to the subdivision of Texas physiography.

**MULTI-VARIANTS K-MEANS CLUSTERS RUN WITH DIFFERENT  
NUMBERS OF CLUSTER**

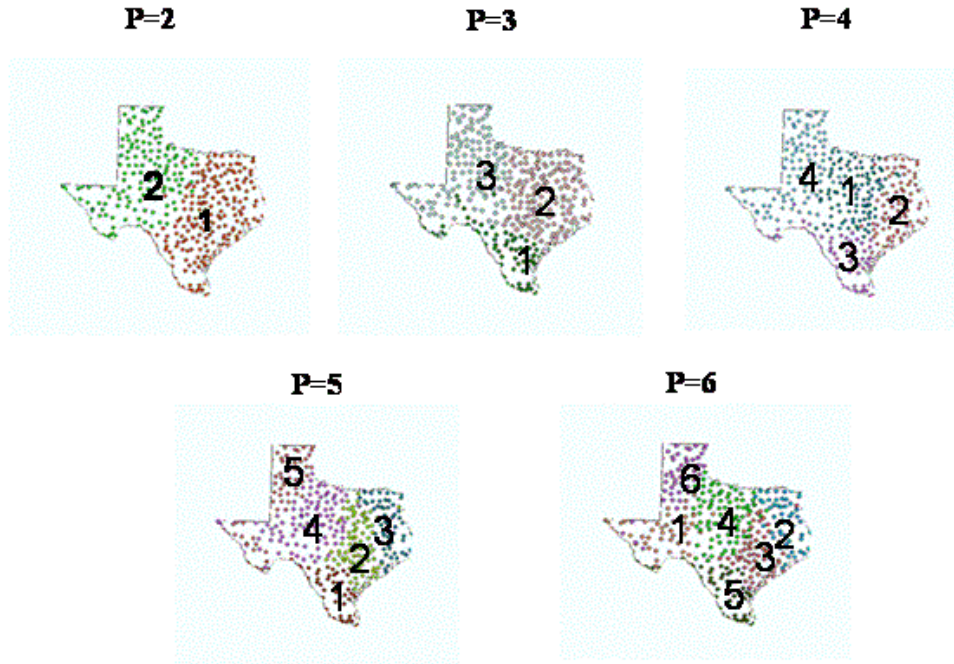


Figure 4 Results of multi-variant k-means clusters at number of cluster p from 2 to 6.

Both global and local temperature and precipitation variograms were used to compare the homogeneity and scale of autocorrelation. Spherical function is used for curve fitting in variogram to decide if there is a dominant pattern of range exists. If the dominant pattern exists and the range is smaller than the diameter of cluster, it indicates that further partition within this cluster is needed. When the spherical function pattern is not distinct, there are two possibilities. If the minimum and maximum semi-variance difference at each lag is large, it is possible that the autocorrelation is determined by larger scale of range. Further partition of this cluster is not necessary. If the minimum and maximum semi-variance difference are small throughout some part of the range, it is possible that the spatial autocorrelation is consists of a combination of various ranges of smaller scales and further partition is needed. The distinction between these two possibilities, however, is arbitrary and further studies are needed.

Figure 5 and 6 illustrates global and local temperature and precipitation variograms when cluster number p is 6. For local temperature variograms, the spherical function is distinct for cluster 1, 2 and 4. The range in cluster 1, 2 and 4 is 1.7, 0.97 and 1.07 respectively, all of which are smaller than the global variogram range 2.6. For local precipitation variograms, there is only one distinct spherical function at cluster 1 and its range is 2.5. There is no distinct spherical function for global precipitation variogram. However, there is a distinct autocorrelation at the range of 3.0,

which implies it may consist of several ranges of autocorrelation. To decide which clusters need to be partitioned requires the consideration of both temperature and precipitation autocorrelation. It is not clear to see the values just with one cluster number of  $p$ . But when all numbers of clusters  $p$  were compared, it is easy to see which optimum number of cluster  $p$  can be determined.

Table 2 summarizes the results of the variograms of temperature and precipitation for each cluster. The ranges of temperature concentrate at four means - 1.1, 1.8, 2.5, and 3.0. The ranges of precipitation are not as concentrated at certain values as temperature even though it does show some distinct spherical functions. The ranges are around 1.45, 2.7 and 3.3. The last column of that table saves the total ratio of distinct spherical function for temperature and precipitation. It can be used as an indicator to check if further partition is needed. Higher ratios indicate the cluster needs to be partitioned into smaller clusters. The result shows that cluster 5 and 6 have lower total ratios than the rest of the cluster numbers. Therefore, Cluster number  $p$  5 and 6 are the most optimum numbers of cluster.

### GLOBAL AND LOCAL TEMPERATURE VARIOGRAMS

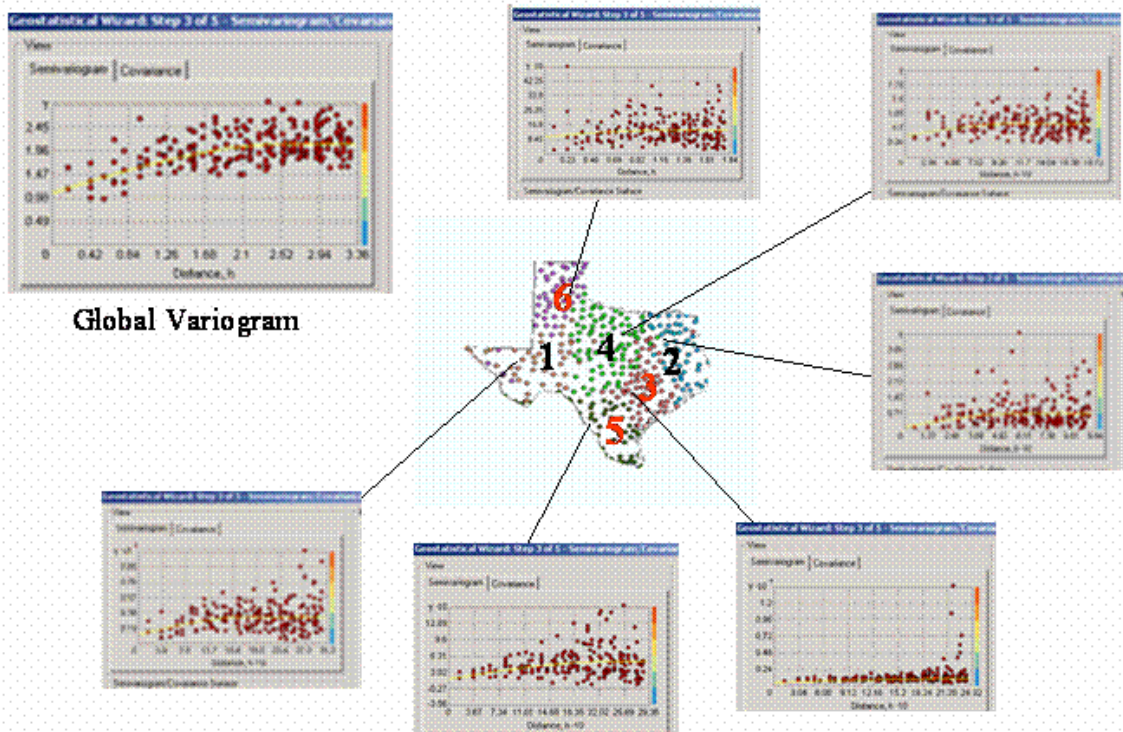


Figure 5 Global temperature variogram shows distinct spherical function and larger ranges than local variogram. The black label of cluster 2 and 4 indicates a distinct range and further division of the cluster is needed whereas the red label of cluster of 1, 3, 5 and 6 indicates no distinct spherical function.

## GLOBAL AND LOCAL PRECIPITATION VARIOGRAMS

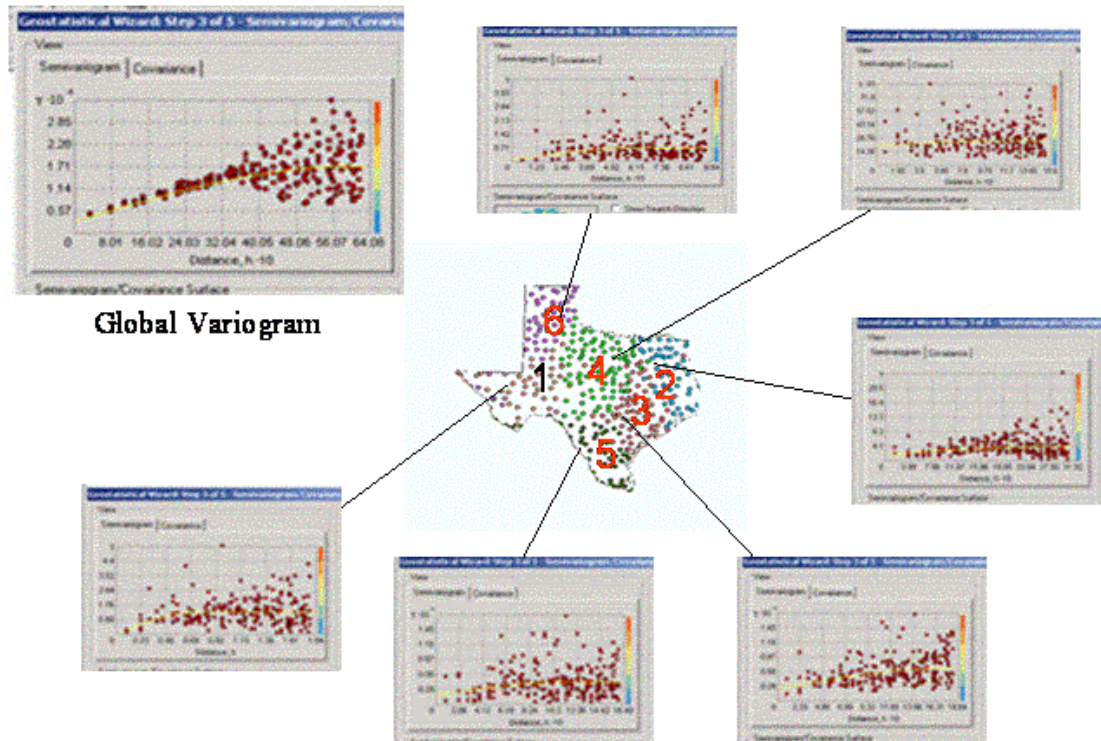


Figure 6 Neither global nor local precipitation variogram (when  $p = 6$ ) shows distinct spherical function and ranges. The black label of cluster 1 indicates a distinct range and further division of the cluster is needed whereas the red label of cluster from 2 to 6 indicates no distinct spherical function.

Temperature and precipitation clusters should be the collections of measuring points with similar geostatistical properties. In geography, this can be the regimes or areas with similar geographical attributes. The global temperature distribution is the result of latitude whereas the global precipitation distribution is the result of longitude. Regional influences can include elevation, geological features, and the distance to water bodies. Since the universal kriging is used in spatial autocorrelation, the global trend of both temperature and precipitation had been removed during data processing, I compared the regional physiography with the clusters.

Total Number of Clusters p	Cluster Name	Temperature Ranges	Ratio of Distinct Temperature Spherical Function	Precipitation Ranges	Ratio of Distinct Precipitation Spherical Function	Total Ratios
1	1.1	2.6		3.4*		
2	2.1	2.3	0.5	3.1*	0.5	1
	2.2	NA		NA		
3	3.1	NA	0.33	1.27	0.67	1
	3.2	NA		3.26		
	3.3	3		NA		
4	4.1	1.9	0.5	1.6	0.75	1.25
	4.2	NA		NA		
	4.3	3		2.4		
	4.4	NA		1.1		
5	5.1	NA	0.2	1.1*	0.4	0.6
	5.2	NA		2		
	5.3	NA		NA		
	5.4	1.78		NA		
	5.5	NA		NA		
6	6.1	1.7	0.5	2.5	0.167	0.667
	6.2	0.97		NA		
	6.3	NA		NA		
	6.4	1.1		NA		
	6.5	NA		NA		
	6.6	NA		NA		

Table 2 shows the ratio of distinct spherical function for temperature and precipitation. Higher ratios indicate the cluster needs to be divided into smaller clusters. The results show that a total number of clusters of 5 has the lower combining ratios of both temperature and precipitation. Therefore, it is the most optimum cluster number. NA means there is no distinct pattern of spherical or other good fitting function. \* indicates there is no distinct spherical function, but there is distinct autocorrelation.

The physiographic map of Texas by Bureau of Economic Geology (1997) was compared with each of the multi-variant k-means clusters. The shapes, areas, and locations of clusters had the best match with the physiography of Texas when  $p$  was 6 (Fig. 7). Cluster 1 is coincident with the Basin and Range (West Texas, dark brown area), part of the High Plains (pink area) and part of Edwards Plateau (green area). The combination of clusters 2, 3, and 5 is roughly coincident with the Coastal Plain (orange and yellow areas). Due to the effects of longitude and latitude, the further partition of Coastal Plains into three sub-regions of 2, 3 and 5 is reasonable. Cluster 4 is coincident with the combination of the Central Plains (blue area) and part of Edwards Plateau (green area). Cluster 5 includes almost all the High Plains (pink area). Cluster 4 and 1 are the areas need to be partitioned further according to results of local temperature variograms. Therefore, if more stations are provided within clusters 4 and 1, further partition may show better match with the physiographic map. Caution should be used because physiography is an arbitrary perception depending on geomorphic interpretation. But in general, it provides good reference for the interpretation of local temperature and precipitation clusters.

**COMPARISON BETWEEN PHYSIOGRAPHIC  
MAP AND MULTI-VARIANTS CLUSTER P = 6**

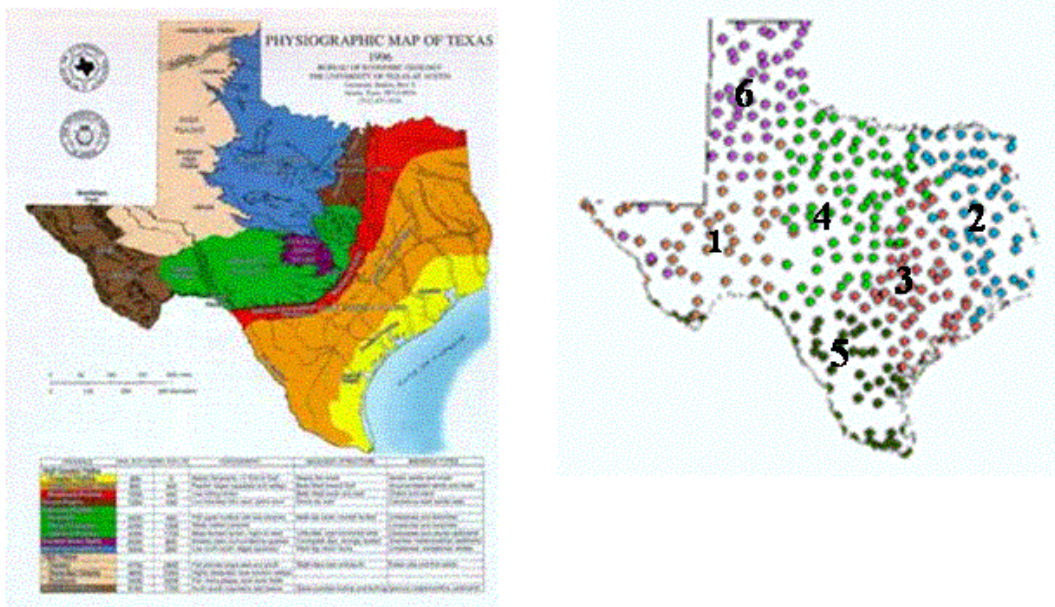


Figure 7 Comparison of physiographic maps with k-means clusters  $p = 6$ .

## 6. Future Studies

Current criteria to partition cluster is based on distinct patterns of spherical function or the maximum-minimum semi-variance differences. However, the categorization of the maximum-minimum semi-variance differences needs more studied. To ensure the effectiveness of k-means algorithms and to establish the criteria for the range of variograms to determine the optimum number of clusters, a diversity of geographical phenomenon, both natural and social such as environmental hazards or crime analysis, need to be considered. Another improvement is to use data sets which are more related to regional scale instead of a combination of both global and local scale. Good candidates are the crime or real estate data sets. The visualization can also be improved by using fuzzy k-means interface or three- dimensional interfaces. Several scales of geostatistical properties such as anisotropy need to be taken into consideration to compare the shapes and types of clusters.

## 7. Conclusion

This study is intended to develop a new k-means algorithm for cluster analysis and to use autocorrelation for optimizing the number of clusters. Two analyses were conducted: one with single-variant average spring temperatures between 1971 and 2000 in Texas, another with multi-variant average annual temperatures and participation between 1971 and 2000 in Texas. Single-variant analysis showed global spherical autocorrelation model and ellipsoidal, spherical and irregular shape of clusters. The range of the global variogram can verify the geostatistical validities of clusters when the number of clusters,  $p$ , is 4, 5 and 6. Multi-variant analysis showed obvious global and local temperature spherical autocorrelation model. Global precipitation showed a well-correlated trend without autocorrelation function to fit the model. This may be due to having no dominant range of precipitation at global scale. Instead, the global variogram is the combination of several local ranges variograms. The ranges of global and local variograms can verify the geostatistical validities of clusters when the number of clusters,  $p$ , is 5 and 6. One physiographic map of Texas was used to evaluate various multi-variant cluster diagrams. The results indicated that when  $p$  is 6, the cluster diagrams correlated well with physiography. Future research is needed to decide if the data, which were not well fitted into any autocorrelation function, is due to either the combination of smaller scales of variogram functions or noises. Future studies may include: using data without global trends, better k-means algorithms, and methods of visualization.

## Reference

- Atkinson, P., N. J. Tate. "Spatial scale problems and Geostatistical solutions: A Review", *Professional Geographer*, Vol. 52, No. 4, pp. 607-623, 2000.
- Blatt, M., S. Wiseman, and E. Domany. "Superparamagnetic clustering of data", *Phys. Rev. Lett.*, Vol. 76, No. 18, pp. 3251-3254, 1996.
- Burrough P. A., R. A. McDonnell. *Principles of Geographical Information Systems*. Oxford University Press, 1998.
- Grubestic, T. H, "Detecting hot spots using cluster analysis and GIS", 2002  
<http://www.ojp.usdoj.gov/nij/maps/Conferences/01conf/Grubestic.doc>.
- Milligan, G. W. and M. C. Cooper, "An examination of procedures for determining the number of clusters in a data set.", *Psychometrika*, Vol. 50 No. 2, pp. 159-179, 1985.
- Murry, A, "Spatial analysis using clustering methods: Evaluating central point and median approaches", *Journal of Geographical Systems*, Vol.1, pp. 367-383, 1999.
- \_\_\_\_\_, V. Estaville-Castro, "Cluster discovery techniques for exploratory spatial data analysis", *Int. J. Geographical Information Science*, Vol. 12, No. 5, pp. 431-443, 1998
- Ratcliffe, J. H., M. J. McCullagh, "Hotbeds of crime and search for spatial accuracy", *Journal of Geographical Systems*, Vol. 1, pp. 385-398, 1999.
- Sarle, W.S., "Cubic Clustering Criterion", SAS Technical report A-108. Cary, NC: SAS Institute Inc., 1983.
- Vesanto, J, E. Alhoniemi, "Clustering of the Self-Organizing Map", *IEEE transactions on Neural Networks*, Vol. 11, No. 3, pp. 586-600, 2000.