

Multiscale Study of the Spatial Variability in the Cluster Analysis of Rainfall

Jeng-Jong Pan¹, Sheng-Tun Li², Shuzheng Cong¹

¹NOAA, NWS, Office of Hydrology,
Silver Spring, Maryland 20910, USA
E-mail: jjpan@nhds1.ssmc.noaa.gov

²Department of Information Management Technology,
National Institute of Technology at Kaohsiung, Taiwan

The postal address of the author:

Dr. J. J. Pan
NOAA/NWS/OH/HRL, SSMC2
1325 East West Highway
Silver Spring, MD 20910, USA

Telephone: (301) 713-0640 X 149
FAX: (301) 713-0963
email: jjpan@nhds1.ssmc.noaa.gov

Multiscale Study of the Spatial Variability in the Cluster Analysis of Rainfall

ABSTRACT

There are three issues that should be considered in the cluster analysis of rainfall stations: (1) input data including concerns on the data length, scale, and transform, (2) analysis methods based on dissimilarity measure, correlation matrix, or cognitive approaches, and (3) groupings which generate non-overlapping regions or transition zones between regions. We reviewed these three issues and paid additional attention to the scales used in the input data. Regions determined from different scales of input data could be varied. To reduce the regionalization uncertainties by using a single scale time series, it will be useful to generate two-dimensional, scale-based data which covers a range of scales as input in cluster analysis. A multiscale approach using the continuous wavelet transform (CWT) is proposed to investigate the nonstationary characteristics of rainfall for studying the spatial variability in the cluster analysis of rainfall. The scalogram generated from CWT has a higher dimension to analyze the scale-dependent variances. The rainfall over Iowa was used for demonstration. The regions determined by using 3- to 30-day scales can reduce the local small features by using one small scale (e.g., 3-day) input or improve the over-smoothed regions by using one large (e.g., 30-day) scale input.

Key Words: Cluster analysis; wavelet transform; principal components analysis; rainfall; Iowa, USA.

1. INTRODUCTION

The purpose of cluster analysis of rainfall stations, also called regionalization, is to decompose a large complex area into several smaller homogeneous regions for research and applications in climatology and hydrology. Wide regionalization studies using cluster analysis and multivariate statistics have been published in the last 20 years. In general, to study regionalization, three issues -- input data, analysis methods, and groupings -- should be discussed for the interpretation of the subregion features. For example, in the input data issue, the spatial variations of rainfall could be a function of temporal scale indicating that the results from the cluster analysis of rainfall could be varied by using hourly, daily, weekly, monthly, seasonal, or annual rainfall data. The selection of an appropriate temporal scale is really dependent on the purpose of application. There could be multiple options involved in each of issues.

First, we briefly review these three issues required for cluster analysis and then focus on the temporal scale problem in regionalization. The continuous wavelet transform (CWT) is applied in the analysis of nonstationary characteristics of rainfall and in the multiscale study of its spatial variability. Rather than using a single scale input data (e.g., 3-day average), the homogeneous regions can be determined based on a range of scales of rainfall (e.g., 3- to 30-day scales). This provides an option to investigate short- and long-term spatial variabilities of rainfall. In section 2, we discuss concerns in input data, analysis methods, and grouping for regionalization; in section 3, the CWT is introduced; in section 4, a demonstration of the application of CWT to cluster analysis is shown; and section 5 contains the conclusions from this study.

2. CONCERNS FOR REGIONALIZATION

Regionalization study is to delimit homogeneous regions based on the spatial variability of one or more physical variables (e.g., rainfall, temperature, etc.). The results from the regionalization study could be varied according to three issues: (1) input data, (2) analysis methods, and (3) groupings. There are several options in each issue which make it very difficult to compare performances among the different options. In this section we provided an outline (Figure 1) which includes concerns that should be considered in a regionalization study.

2.1 Input Data

Quality control of input data is critical for outlier detection before processing. Outliers, usually extremely high values in the positive tail of a distribution, can produce distortions in the correlation coefficients calculation which are sensitive to most correlation matrix-based multivariate statistical analyses. If no outlier detection is performed for quality control, then a robust analysis method, such as the robust principal components analysis, will be useful to separate outliers from clusters. In addition, rain gages could be moved during data collection. Inconsistency checks of rainfall using double-mass analysis (DMA) is necessary to make an appropriate correction, particularly if monthly or annual data are used for long-term rainfall

analysis. The trend analysis is also required to check temporal inhomogeneity and remove nonclimatic induced variations.

Because rainfall stations are grouped based on the similarity of input data characteristics, the regions determined from cluster analysis could vary with data length, scale, and form. Form indicates that the input data is of a raw or transformed type. All three of these factors could change the clustering results.

2.1.1. Length of data

Assume daily rainfall data for over 20 years are available. We may use the entire 20 years of data or select a portion of the data (e.g., 5 years) as input. If the regions obtained from cluster analysis are different when we use different length of input data, it implies that regions could be changed with time due to climate change. A basic assumption is that the regions determined from a time period are homogeneous in that time period only. Sequentially analyzing a short time period of data can be used to study changes in the region, while a longer time period data could be necessary for long-term estimation of rainfall or water resources management.

2.1.2. Scale of data

Regions determined from cluster analysis are sensitive to the scale of input data. The selection of an appropriate scale is really dependent on the application purpose. For example, Richman and Lamb (1985) used 3- and 7-day summer rainfall to determine homogeneous regions for a short-time weather forecast, climate change study, and crop-yield modeling. Gadgil and Iyengar (1980) used the mean 5-day rainfall to study the relationship between rainfall distribution and the monsoon. Gong and Richman (1995) used the 7-day rainfall to study the growing season. Fovell and Fovell (1993) used monthly temperature means and rainfall accumulations to specify climate divisions. Jackson and Weinand (1994) discussed the annual and seasonal variables in the study of tropical rainfall. Johnson and Hanson (1992) studied the spatial variations for daily summer, daily winter, monthly summer, and monthly winter.

A preliminary comparison might be helpful to decide the scale of the input data. Van Regenmortel (1995) first evaluated the percentage of cumulative variance explained by the principal components analysis for daily, 5-day, 10-day, and monthly rainfall sum. He then selected the 10-day average rainfall to study the soil-moisture status and drought assessment. In general, daily or weekly rainfall showed high frequency and local features, while the monthly or annual rainfall characterized by its low frequency and large spatial scale. An option of using a range of temporal scales as input might be useful in a regionalization study. The CWT, described in more detail in section 3, provides the capability for generating a scalogram which reflects the rainfall intensity distribution over a range of scales.

2.1.3. Transform of data

An appropriate transform of data can enhance required features and reduce ‘noise’ effects. For example, daily data are not normally distributed, and square-root or logarithmic transformation can reduce the skewness of the distribution so that the impact of extreme values on the computation of correlation coefficients will be reduced (Van Regenmortel, 1995). In the climatic pattern analysis of 3- and 7-day summer rainfalls, Richman and Lamb (1985) evaluated the input data as raw, square-root, and \log_{10} transformed; they selected the square-root transform after checking the differences of means and standard deviations before and after the transforms. However, if the purpose of the regionalization is to identify the location of short duration storm events, the raw data could be more appropriate as input. In general, these nonlinear transforms work as low-pass filtering and could reduce the number of homogeneous regions.

In addition to the above simple numerical transforms, more complicated transforms, such as the Fourier transform, CWT, etc., are also available by investigating the similarities in the frequency or scale domain. In spectral analysis, data are transformed from time domain into frequency domain using Fourier transform, and stations are grouped based on the frequency distribution which reflects similar periodical features. Spectral analysis is useful for stationary time series analysis. To study long-term seasonal or annual cycles in rainfall, regions can be defined based on the distribution of first several harmonics amplitudes (Kirkyla and Hameed, 1989). Using monthly rainfall, Krepper et al. (1989) applied spectral analysis to study interannual, annual, and intraannual variability and delineated the transition zone between wet and dry regimes. In addition, maximum entropy spectral analysis is a robust way to deal with short-length time series (Leite and Peixoto, 1995). CWT, the method used in this paper, generates a two-dimensional, time-scale distribution for analyzing nonstationary characteristics of rainfall and provides more detailed information embedded in a one-dimensional time series for cluster analysis.

It is possible to determine regions by comparing probability distribution functions based on the regional analysis in the L-Moments methodology (Guttman, 1993). Easterling (1989) studied the regionalization of thunderstorm rainfall by use of the incomplete gamma distribution. Other statistical parameters, such as entropy, can also be considered as ancillary information in regionalization study.

2.2 Analysis Methods

Methods used in regionalization study can be divided into three categories: methods based on dissimilarity measure, cognition, or correlation matrix.

2.1.1. Methods based on dissimilarity measure

Consider x_i is an M -point vector of the time series at station i . The dissimilarity between two stations, x_i and x_j , can be measured as a function of:

- | | |
|---------------------------------------|---|
| (1) Euclidean distance: | $d(x_i, x_j) = [(x_i - x_j)^T (x_i - x_j)]^{1/2}$, |
| (2) Manhattan distance: | $d(x_i, x_j) = x_{i1} - x_{j1} + x_{i2} - x_{j2} + \dots + x_{iM} - x_{jM} $, |
| (3) Mahalanobis distance: | $d(x_i, x_j) = [(x_i - x_j)^T C^{-1} (x_i - x_j)]^{1/2}$, |
| (4) the theta angle between stations: | $\theta(x_i, x_j) = \cos^{-1}[(x_i^T x_j) / (x_i x_j)]$, or |
| (5) inverse correlation coefficient: | $r(x_i, x_j)^{-1} = (y_i^T y_j) / (y_i y_j) ^{-1}$, |

where $y_i = x_i - \bar{x}$, \bar{x} is the mean of x_i , T means the transpose, and C is the covariance matrix of $(x_i - x_j)$.

Euclidean distance, the theta angle, and the inverse correlation coefficient have been used in the comparison of several cluster analysis methods by Gong and Richman (1995), while the Manhattan distance and Mahalanobis distance are more robust for noisy data.

In general, there are two types of techniques in this category, hierarchical and nonhierarchical clustering approaches. In a hierarchical clustering (e.g., linkage method, Ward's method, etc.) stations can be grouped via either top-down (division) or bottom-up (merger) by partition patterns from a dissimilarity matrix. Nonhierarchical clustering methods (e.g., K -means, vector quantization (VQ), etc.), specify a set of centroids of K groups. Based on the distance between one station and each centroid, the station is assigned to the nearest group. After the assignment of each station, the new centroids of clusters are recomputed, and the assignment of each station is repeated. The iterative procedure will continue until there is no change to the members in each group. A detailed review of these two cluster techniques can be found in Gong and Richman (1995). In this category, each station must belong to one and only one group, and a hard boundary exists between regions. It is impossible to identify the transition zones if they exist between regions.

2.2.2. Methods based on cognition

Cognitive methods, such as fuzzy logic, neural networks, etc., have been widely studied for cluster analysis. Fuzzy clustering uses the membership coefficients to describe the degree of one station belonging to a group. The range of membership coefficients is zero to 1, and the sum of coefficients at one station must be equal to 1. This approach indicates that a station could belong to several groups with different degrees of membership. Fuzzy clustering is a generalization of grouping whereas in hard clustering at each station, only one of the membership coefficients is 1 and the remainder are zeroes. For example, assuming there are K clusters, the fuzzy K -means algorithm is implemented in an iterative computations as follow (Bezdek, 1981):

- (1) Select K groups and compute their centroids c_j , $j = 1, 2, \dots, K$.
- (2) Compute the membership coefficients, $u(x_i, c_j)$ with a given real number $p (> 1)$,

$$u(x_i, c_j) = \frac{\left[\frac{1}{d(x_i, c_j)} \right]^{1/(p-1)}}{\sum_{k=1}^K \left[\frac{1}{d(x_i, c_k)} \right]^{1/(p-1)}}$$

- (3) Update the new centroids c_j' ,

$$c_j' = \frac{\sum_{i=1}^N u(x_i, c_j)^p x_i}{\sum_{i=1}^N u(x_i, c_j)^p}$$

and recompute new $u(x_i, c_j)$ according to (2).

- (4) The process will terminate if the $\max |u(x_i, c_j) - u'(x_i, c_j)|$ over all membership coefficients is less than a predefined small threshold, then stop; otherwise go to step (3).

As in most cluster analysis methods, the fuzzy K-means clustering requires information on the number of clusters and the locations of associated centroids. To avoid these problems, some adaptive and optimal fuzzy clustering algorithms have been published (e.g., Gath and Geva, 1989). The closeness of one station to each group can be described as a function of the membership coefficient providing the means to determine the fuzzy boundaries between regions.

There are various neural networks which provide unsupervised learning mechanism for cluster analysis. For example, the Kohonen's self-organizing-feature-maps (SOFM) method, which is similar to the K-means method, has the capability to extract features from large data sets without supervision (Kohonen, 1982). However, compared with the K-means method, a priori assumption of the number of clusters is eliminated in the SOFM method. In the SOFM method, the lattice structure of neurons in the output layer can show the topological features among neurons. By introducing fuzzy membership for output neurons, Sim and Huntsberger (1991) proposed fuzzy SOFM so that, during the training, one or more winners in the output layer for any input vectors are allowed. For an input neuron, those output neurons which are closer, in terms of distance measure, have a membership value of 1, and others have a value in the range of 1 and zero. The membership value then adjusts the connection weights to reflect the similarity in clustering. Therefore, the fuzzy SOFM method provides the ability of overlapping clusters in the input space.

2.2.3. Methods based on correlation matrix

There are three popular multivariate statistical methods widely used in this category: empirical orthogonal functions (EOF), principal components analysis (PCA), and common factor analysis (CFA). In the EOF method, the dispersion matrix is solved using singular value decomposition (SVD). The dispersion matrix can be either the covariance matrix or correlation matrix. The correlation matrix was adopted by most regionalization studies since standardized data were used to minimize the impact from stations with higher variability. However, the covariance might be more appropriate for locating the individual climate regions with large variance in cyclone climatology (Overland and Preisendorfer, 1982). The eigenvectors derived from the correlation matrix represent the orthogonal basis functions. The rainfall time series at each station is a linear combinations of these basis functions. The corresponding eigenvalues represent the amounts of the total variance that are explained by each eigenvector.

In PCA, a PC loading factor is equal to the multiplication of the associated EOF coefficient and the square root of the corresponding eigenvalue. Loading factors represent the correlation between stations and components. The difference between CFA and PCA/EOF is that CFs are extracted only from the variance between two or more stations and not the total variance for all stations as in the PCA/EOFs (Barring, 1988). According to the assignment of the dispersion matrix, there are six basic operational modes in PCA depending on which parameters are selected as variables, individuals, and fixed entities (Richman, 1986). S (Spatial)-mode (e.g., Dyer, 1975; Richman and Lamb, 1985; Barring, 1988; Johnson and Hanson, 1992; Van Regenmortel, 1995) and T (Temporal)-mode (e.g., Gadgil and Iyengar, 1980) are the two major modes used for regionalization. The S-mode, adopted by most applications, assigns the time series from one station as one column in the matrix, while T-mode is done in the opposite way. Therefore, the S-mode clusters stations with similar temporal patterns, but T-mode separates observations into subsegments with similar spatial characteristics. The T-mode analysis requires additional steps to determine the number of regions.

The rotation of principal components, using varimax method (e.g., Richman, 1986) or oblique method (e.g., Jackson and Weinand, 1995), attempts to maximize the variance of the component loadings between each component for producing a few large loading factors and reducing other factors making it easy to discriminate stations. In a systematic methodological review, Gong and Richman (1995) performed an intercomparison of various cluster analysis methods and indicated that the rotated PCA could be more accurate than other methods.

In PCA, the number of reserved components will be determined before the rotation procedure is performed. A simple procedure is based on the distribution of N sorted descending eigenvalues,

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \dots \geq \lambda_N$$

Only the first K components, where $\lambda_k \geq$ threshold (say, 1.0), are used in the rotation procedure. The variance explained by each component is defined as

$$f_i = \frac{\lambda_i}{\sum_{k=1}^N \lambda_k}$$

The total accumulative variance of the first K components F_K , where $F_K = f_1 + f_2 + \dots + f_K$, provides ancillary information in the determination of the number of groups. The higher F_K values (e.g., 0.8) will reserve more original information. It is possible that the number of groups will be reduced by checking the associated loading factors after the rotation.

'Rule N' is another approach in the determination of number of groups (Preisendorfer and Barnett, 1977). In 'Rule N', a Monte Carlo simulation is used to generate a set of uncorrelated Gaussian variables, then eigenvalues are compared to the distribution of these variables for selecting the number of components. However, there are probably no 'perfect' answers for this problem (Ferre, 1995), particularly when rainfall is not changing rapidly between regions. In addition to the above objective determination, the purpose of applications and the interpretation of regions are two major factors in the decision of the number of clusters.

2.3. Groupings

In general, regions after cluster analysis can be displayed as either (a) hard and nonoverlapping boundaries or (b) transition zones in the overlapping among regions. Methods based on dissimilarity distance assign one station to one and only one region in the former case, while methods based on cognition and correlation matrix are good for both cases. In a rotated PCA, a station is assigned to the component which has the highest loadings factor, or uses the component loading as an indicator of the correlation between each station and component. The loading isopleths with a constant (e.g., 0.65) may be selected to specify the boundary. There could be an overlapping or open space between regions. A station in an overlapped space means it is highly associated with the overlapped regions, while a station in an open space means it has the transition characteristics of the neighboring regions.

3. CONTINUOUS WAVELET TRANSFORM (CWT) IN MULTISCALE STUDY

To study the nonstationary characteristics of rainfall, CWT provides the capability to investigate temporal variation with a different scale. The CWT is defined as the convolution of a time series $x(t)$ with a wavelet function $\psi(t)$ shifted in time by a translation parameter b and a dilation parameter a (Morlet et al., 1982):

$$S(b,a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) \psi^* \left(\frac{t-b}{a} \right) dt$$

where $*$ is the complex conjugate, and $a (> 0)$ and b are real numbers and can be varied continuously. The calculation of $S(b,a)$ is more efficient using the corresponding Fourier transform:

$$S(b,a) = \sqrt{a} \int_{-\infty}^{\infty} X(\omega) \Psi^*(a\omega) e^{ib\omega} d\omega$$

where $X(\omega)$ and $\Psi(\omega)$ are the Fourier transform of $x(t)$ and $\psi(t)$, respectively. The scalogram is defined as $|S(b,a)|^2$. The wavelet function $\psi(t)$ has to satisfy the admissibility condition (i.e., zero mean), and localization support (i.e., fast decay from its center). The approximated Morlet wavelet with a constant c ($c=5.3$ used in this paper) is adopted here.

$$\psi(t) = e^{ict} e^{-\frac{t^2}{2}}$$

Apparently, the Morlet wavelet is a modulated Gaussian function with a zero mean and unit standard deviation. The magnitude of the Morlet wavelet is a Gaussian function which means the amplitude of data is smoothed via a low-pass filter. One advantage of using this low-pass filter is to reduce the Gibb's phenomena in its operation. For example, if the sum of 10 days' data from a daily data set is generated by multiplying a rectangular window with 10-day length to the original data, then the straight truncation at both edges in a rectangular window could cause the Gibb's phenomena while the Gaussian-type Morlet wavelet can reduce such kind problem.

The localization feature of $\psi(t)$ makes that $S(b,a)$ are computed only by data in the cone of influence (COI). As shown in figure 2, only data between b_1 and b_2 can influence the value of $S(b_0, a_0)$. Due to no information beyond edges of input data, $S(b,a)$ has uncertainties in the shaded areas. Using the Morlet wavelet, the radius of the COI at point b is $2a$. The scale is linearly proportional to the wavelength (or period). The period is 1.2 times the scale if $c=5.3$ in Morlet wavelet. Mayer, et al. (1993) discussed the impact of edge effects in implementation and proposed a method to reduce the uncertainties.

Wavelet variance (WV) is defined as the integration of the scalogram over time for given scales. Therefore, the WV is a function of scale which represents the marginal density function of energy and shows the relative intensities of a time series at different scales. It is similar to the power spectrum generated from the Fourier transform. The difference is that the scale is used in the WV while the frequency is used in the Fourier transform; the scale and the frequency have a reciprocal relationship.

4. DEMONSTRATION

The rainfall stations in the state of Iowa, United States, are used to demonstrate the results of using different time scales. The TD-3200 Daily Summary Observations from National Climatic

Data Center (NCDC) provided quality controlled daily rainfall data. We arbitrarily selected 1992 data, checked the associated quality flags, and rejected any stations which had suspected, missing, accumulated, or invalid data. Only 70 stations were reserved after the careful quality check.

4.1. Data analysis

Figures 3 and 4 show the CWT of daily and monthly rainfall, respectively, at a rainfall station in Iowa. Data for only 1 year (1992) are used in figure 3, while 20 years' (1973-1992) monthly data are used in figure 4. The associated scalograms show the dynamic variations as a function of the temporal scale. Due to uncertainties at both edges of the scalogram, the WVs shown here have a little distortion when the scale is large. The nonstationary characteristics in daily rainfall are typically different from the semi-stationary characteristics of monthly rainfall. As shown in figure 4, the long-range trend is identified with the scale 10, which is equivalent to the 12-month period. The scalogram generated from CWT is applicable for cluster analysis if a range of scales is interested in applications.

4.2. Regionalization

To compare the regions determined from different scales of input data, we tried four scales: 3-day, 15-day, 30-day, and 3- to 30-day scales. Raw data were used in each process since no outliers had been detected. The rotated PCA was applied in this regionalization study. Figure 5 shows the loading factors of the first four principal components using the scalogram of a 3- to 30-day scale as input. Only correlation coefficients greater than 0.5 are displayed with the stations, and the single contour line represents the correlation coefficient 0.65. The isolated regions are easily identified from each loading factor.

Figure 6 displays the mosaiced regions derived from figure 5 where the central small region is corresponding to the 5th loading factor. There are several stations (e.g., stations 5, 34, 57, 41, etc.) that are not firmly linked to one cluster. They are located in transition zones between two or more regions. These stations can be assigned to one or more clusters when the threshold used in the contour line decreases, and this could generate some overlapping areas among regions. The rotated PCA objectively provides loading factors, but it is a little subjective in selecting the thresholds in a grouping.

Figures 7-9 show the regions using the 3-day, 15-day, and 30-day rainfall data, respectively. Apparently, more local small regions appeared in the smaller scale (e.g., 3-day) data while the larger regions are generated from the larger scale (e.g., 30-day) data. Particularly, there are more transition zones or uncertainties between regions when using the smaller scale data. Comparing these figures with figure 6, the multiscale input data can integrate the information from a range of scales and compromise the uncertainties using a single scale in input data.

5. CONCLUSIONS

In this paper, we introduced a multiscale study of the spatial variability using CWT. The CWT provides an option to consider the range of scales in the input data which can reduce the local small regions using one small scale input, or improve the over-smoothed regions by using one large scale input in regionalization studies. We also reviewed concerns of input data, analysis method, and grouping issues in the cluster analysis of rainfall stations. The selection of options in each issue is dependent upon the purpose of applications. It will be helpful for future regionalization studies if a committee with members from the climatology and hydrology communities can generate a guideline which provides details of the regionalization procedure and comments or suggestions on these issues.

6. REFERENCES

- Barring, L. 1987. "Spatial patterns of daily rainfall in central Kenya: Application of principal component analysis, common factor analysis and spatial correlation," *J. Climatol.*, **7**, 267-289.
- Barring, L. 1988. "Regionalization of daily rainfall in Kenya by means of common factor analysis," *J. Climatol.*, **8**, 371-389.
- Bezdek, J. C. 1981. "Pattern Recognition with Fuzzy Objective Function Algorithms," Plenum, New York, p. 256.
- Dyer, T. G. J. 1975. "The assignment of rainfall stations into homogeneous groups: An application of principal component analysis," *Quart. J. R. Met. Soc.*, **101**, 1005-1013.
- Easterling, D. R. 1989. "Regionalization of thunderstorm rainfall in the contiguous United States," *Int. J. Climatol.*, **9**, 567-579.
- Ferre, L. 1995. "Selection of components in principal component analysis: A comparison of methods," *Comput. Statist. Data Anal.*, **19**, 669-682.
- Fovell, R. G. and Fovell, M. C. 1993. "Climate zones of the conterminous United States defined using cluster analysis," *J. Climatol.*, **6**, 2103-2135.
- Gadgil, S. and Iyengar, R. N. 1980. "Cluster analysis of rainfall stations of the Indian peninsula," *Quart. J. R. Met. Soc.*, **106**, 873-886.
- Gath, I. and Geva, A. B. 1989. "Unsupervised optimal fuzzy clustering," *IEEE Trans. Pattern Anal. Machine Intell.*, **11**, 773-781.

- Gong, X. and Richman, M. B. 1995. "On the application of cluster analysis to growing season precipitation data in north America east of the Rockies," *J. Climate*, **8**, 897-931.
- Guttman, N. B., Hosking, J. R. M., and Wallis, J. R. 1993. "Regional precipitation quantile values for the continental Unites States computed from L-moments," *J. Climate*, **6**, 2326-2340.
- Jackson, I. J. and Weinand, H. 1994. "Towards a classification of tropical rainfall stations," *Int. J. Climatol.*, **14**, 263-286.
- Jackson, I. J. and Weinand, H. 1995. "Classification of tropical rainfall stations: A comparison of clustering techniques," *Int. J. Climatol.*, **15**, 985-994.
- Johnson, G. L. and Hanson, C. L. 1992. "Time scale difference in the spatial variability of precipitation on a Mountainous watershed: A rotated principal components approach," *The 5th Intl. Meeting Stat. Climatol.*, Toronto, Canada, 539-542.
- Kaufman L. and Rousseeuw, P. J. 1990. "Finding groups in data - An introduction to cluster analysis," John Wiley & Sons, Inc., New York. p. 338.
- Kirkyla, K. I. and Hameed, S. 1989. "Harmonic analysis of the seasonal cycle in precipitation over the United States: A comparison between observations and a general circulation model," *J. Climatol.*, **2**, 1463-1475.
- Kohonen, T. 1982. "Self-organized formation of topologically correct feature map," *Biological Cybernetics*, **43**, 59-69.
- Krepper, C. M., Scian, B. V., and Pierini, J. O. 1989. "Time and space variability of rainfall in central-east Argentina," *J. Climate.*, **2**, 39-47.
- Leite S. M. and Peixoto, J. P. 1995. "Maximum entropy spectral analysis of the Duero basin," *Int. J. Climatol.*, **15**, 463-472.
- Meyers, S. D., Kelly, B. G., and O'Brien, J. J. 1993. "An introduction to wavelet analysis in oceanography and meteorology: With application to the dispersion of Yanai Waves," *Mon. Wea. Rev.*, **121**, 2858-2878.
- Morlet, J., Arens, G., Fourgeau, I., and Giard, D. 1982. "Wave propagation and sampling theory," *Geophysics*, **47**, 203-236.
- Overland, J. E. and Preisendorfer, R. W. 1982. "A significance test for principal components applied to a cyclone climatology," *Mon. Wea. Rev.*, **110**, 1-4.

Richman, M. B. and Lamb, P. J. 1985. "Climate pattern analysis of three- and seven-day summer rainfall in the central United States: Some methodological considerations and a regionalization," *J. Climate. Appl. Meteor.*, **24**, 1325-1343.

Richman, M. B. 1986. "Rotation of principal components," *J. Climatol.*, **6**, 293-335.

Sim, D. S. and Huntsberger, T. 1991. "Self-organizing neural networks for unsupervised color image recognition," *Tenth Annual Int. Phoenix Conf. on Computers & Communications*, 39-45.

Todorov, V. K., Neykov, N. M., and Neytchev, P. N. 1992. "Robust procedures for multivariate data analysis," *5th Int. Meet. on Stat. Climatol.*, Toronto, Canada, 583-586.

Van Regenmortel, G. 1995. "Regionalization of Botswana rainfall during the 1980s using principal component analysis," *Int. J. Climatol.*, **15**, 313-323.

Figure Captions

- Figure 1. Three issues in the cluster analysis of rainfall stations.
- Figure 2. Cone of influence and uncertainties in the scalogram of continuous wavelet transform.
- Figure 3. The continuous wavelet transform (CWT) of daily precipitation. (a) data, (b) CWT, and (c) wavelet variance (WV).
- Figure 4. The continuous wavelet transform (CWT) of monthly precipitation. (a) data, (b) CWT, and (c) wavelet variance (WV).
- Figure 5. The first four principal component station loadings using varimax rotation. The input data have scales 3- to 30-day generated from CWT. Only loading factors which are greater than 0.50 are shown. The contour line is equal to 0.65.
- Figure 6. Regions determined from 3- to 30-day scales. Numbers refer to the stations used in this study.
- Figure 7. Regions determined from 3-day scale. Numbers refer to the stations used in this study.
- Figure 8. Regions determined from 15-day scale. Numbers refer to the stations used in this study.
- Figure 9. Regions determined from 30-day scale. Numbers refer to the stations used in this study.

Cluster Analysis of Rainfall Stations

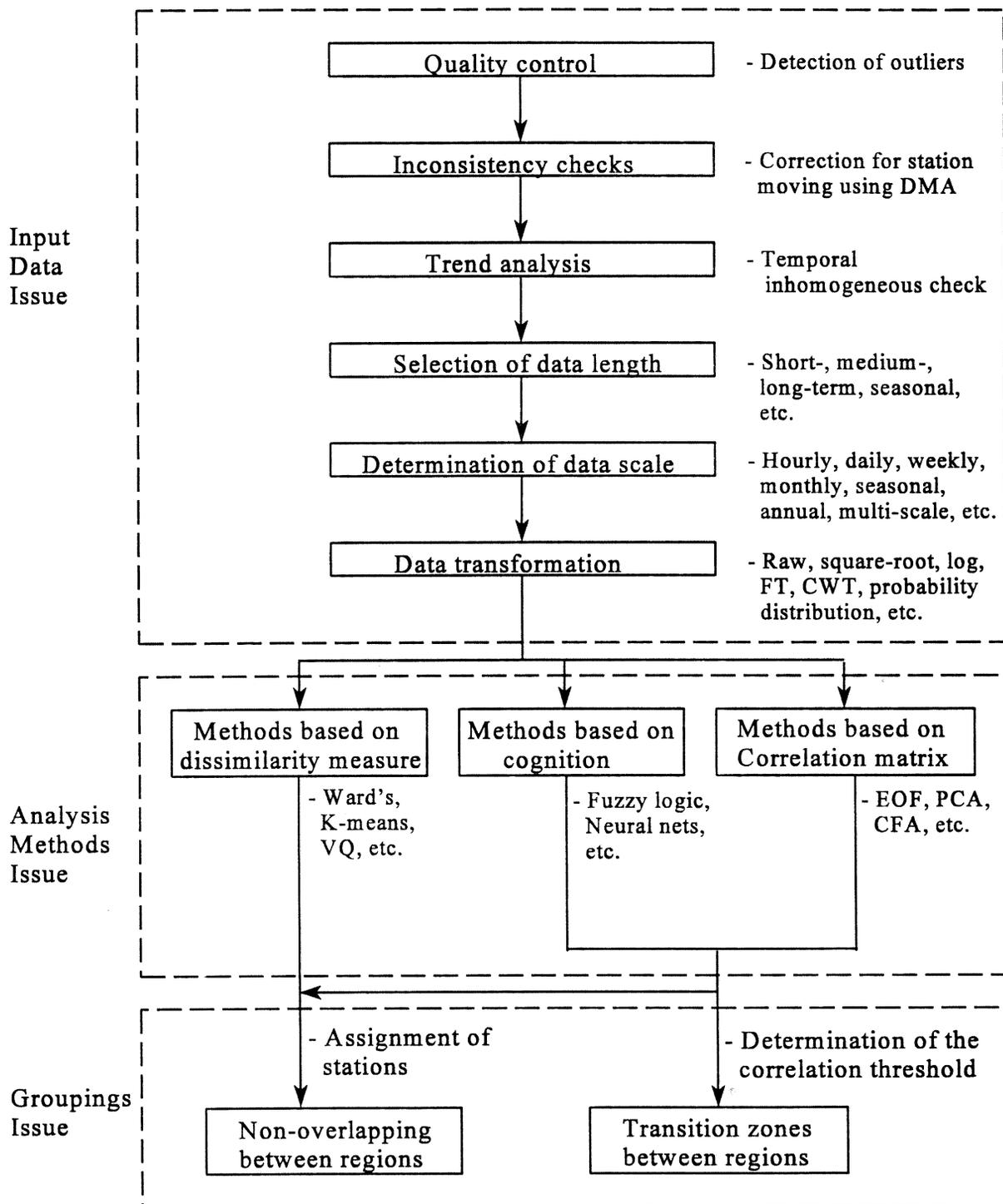
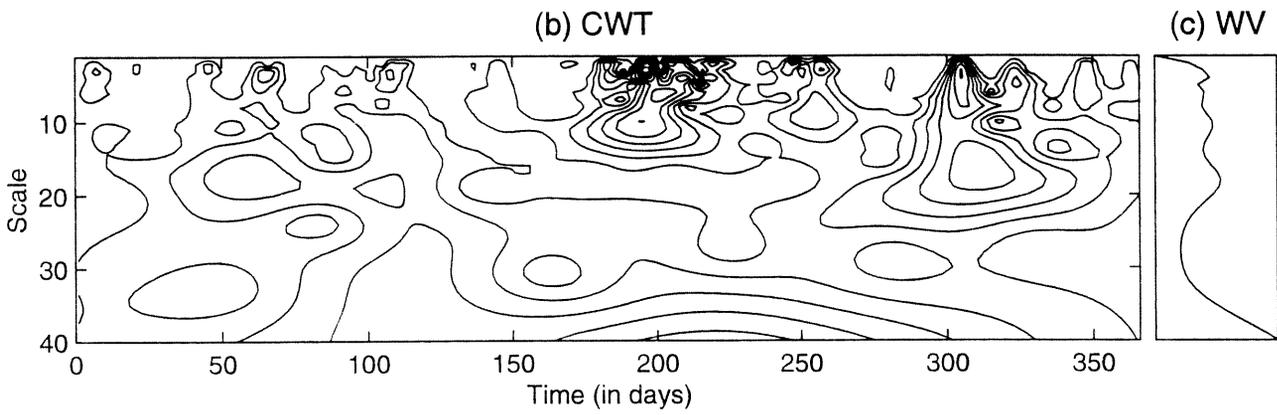
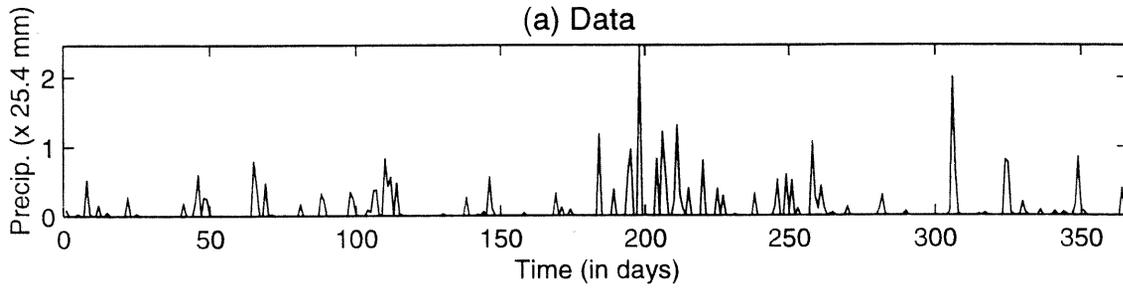
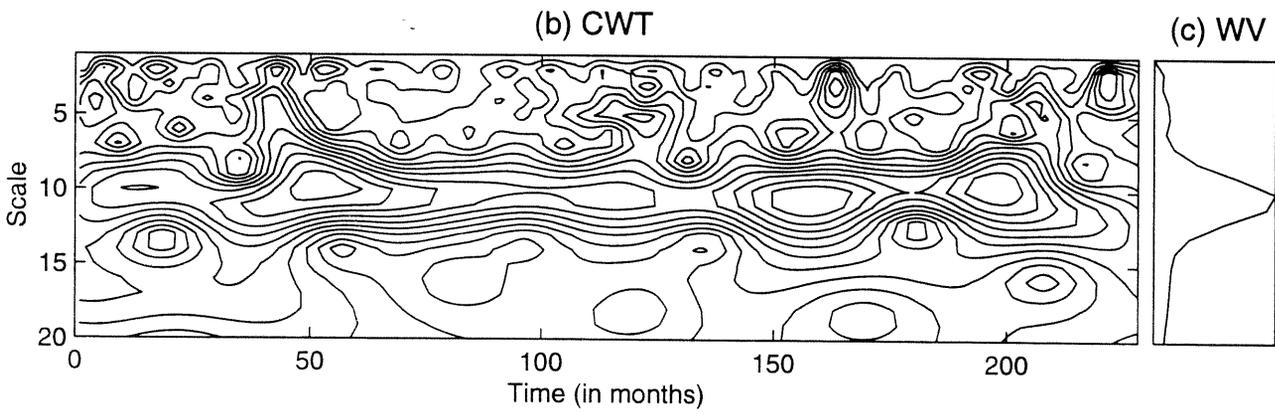
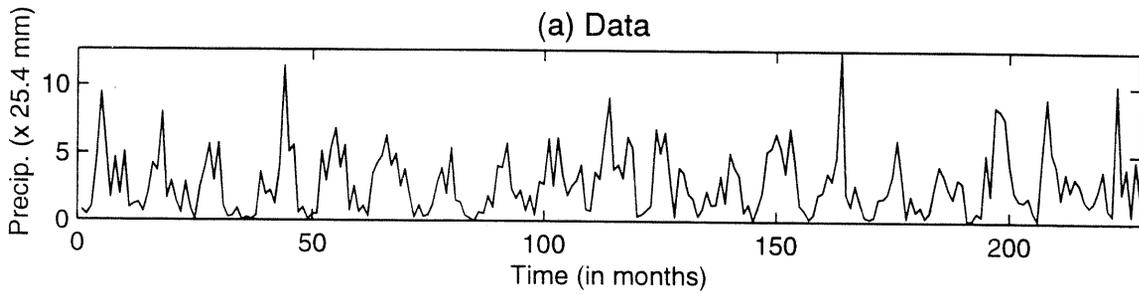


Figure 1. (J.J. Pan, S.T. Li, and S. Cong)

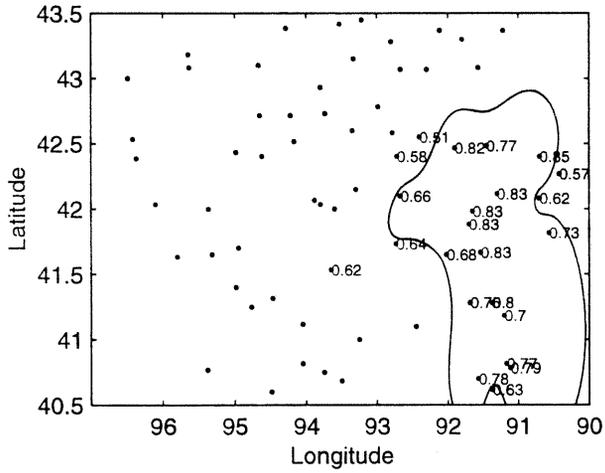
Daily Precip. 1/92 – 12/92 (Iowa Station ID=130200)



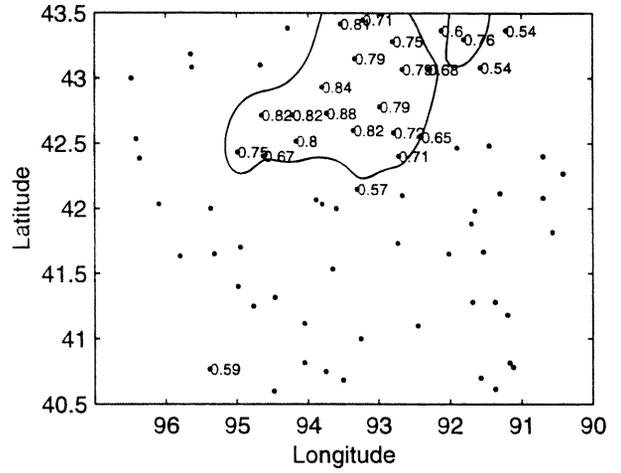
Monthly Precip. 1/73 - 12/92 (Iowa Station ID=130200)



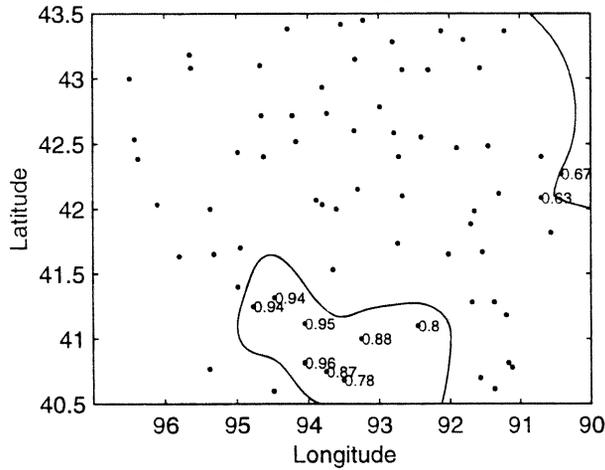
The 1st Loading Factor



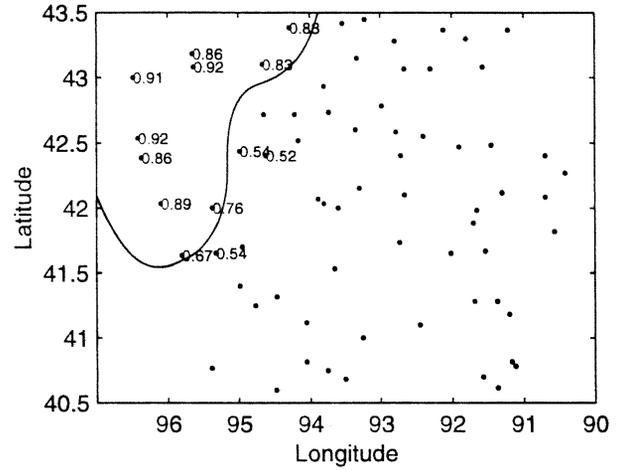
The 2nd Loading Factor



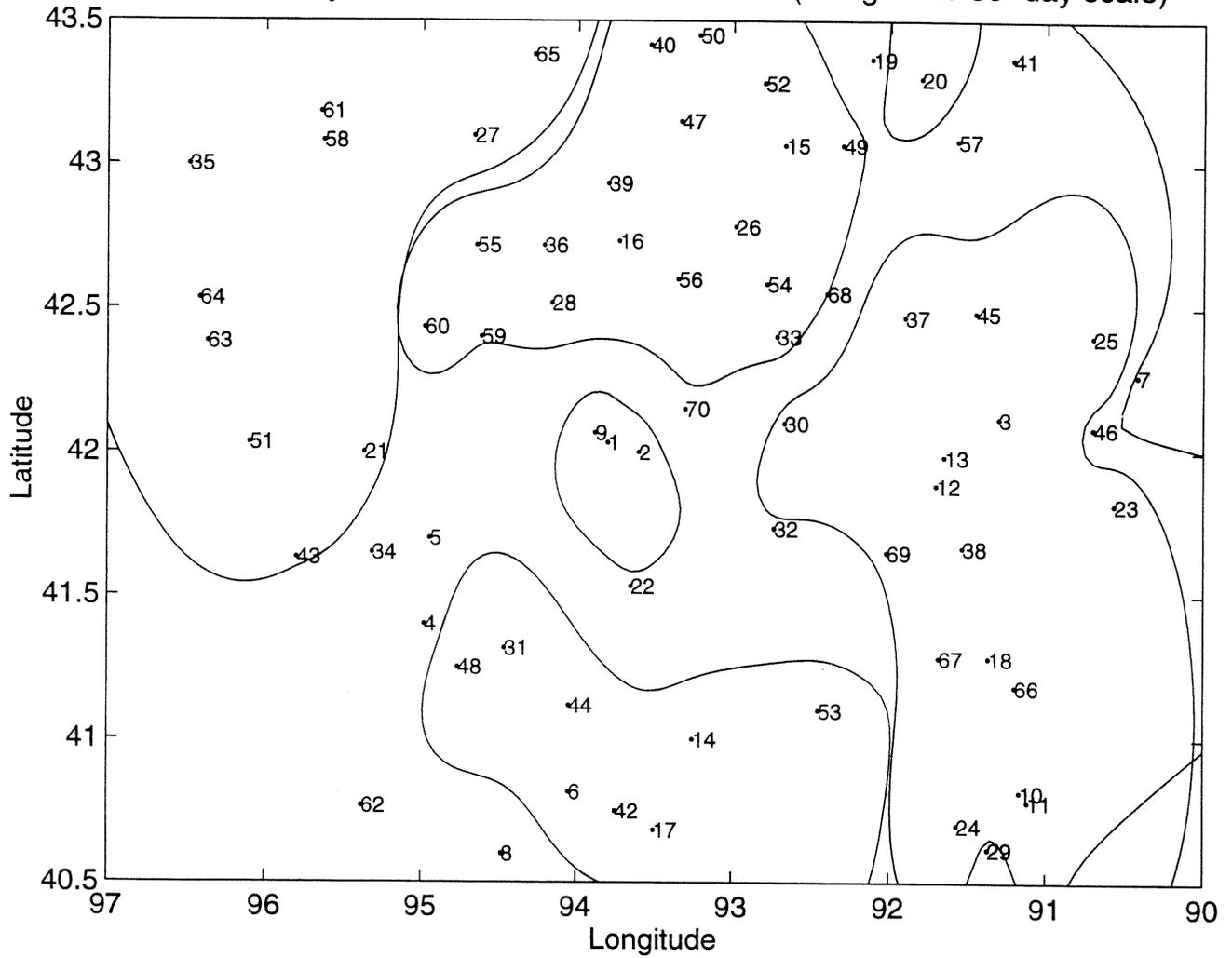
The 3rd Loading Factor



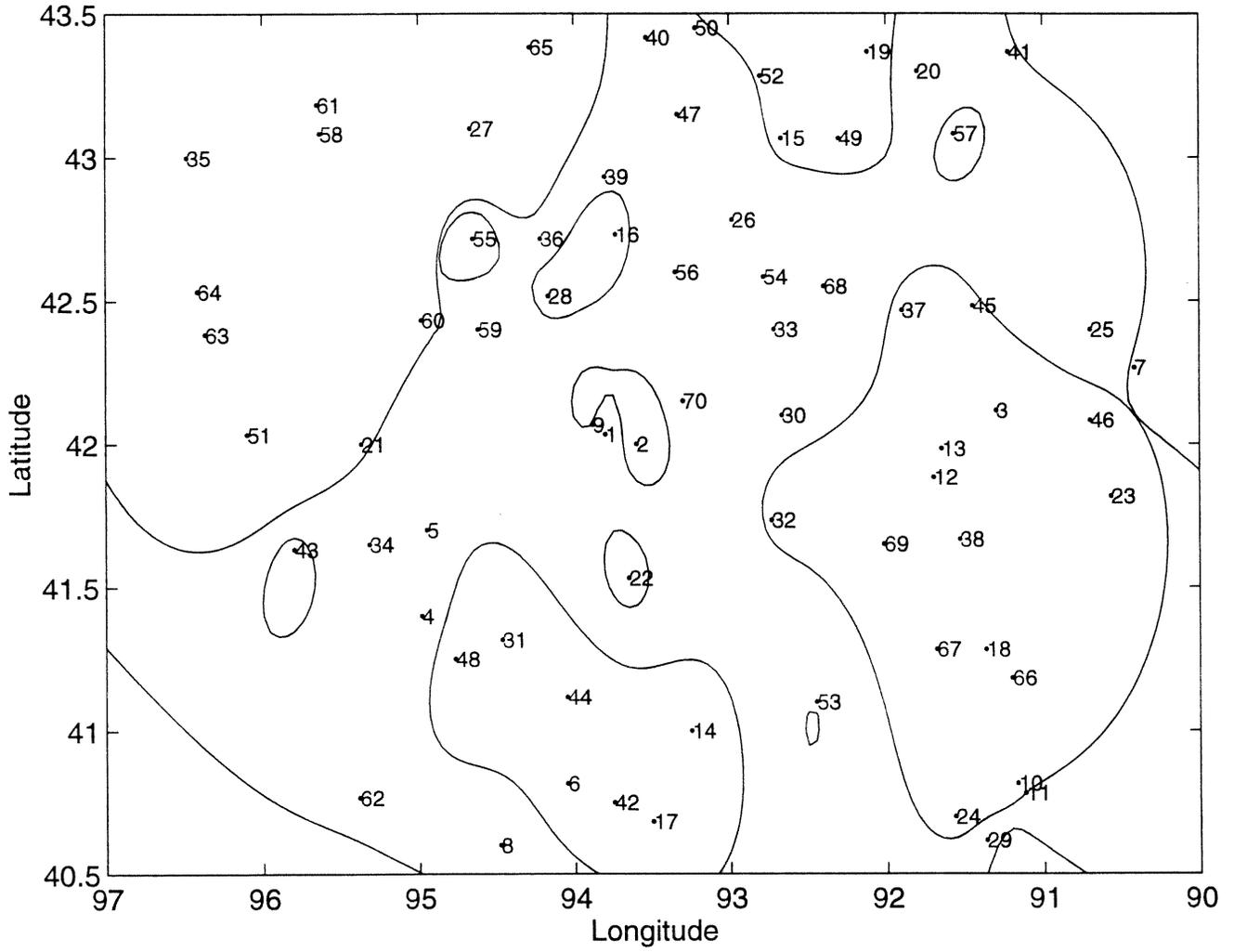
The 4th Loading Factor



Cluster Analysis of Rainfall Stations in IOWA (using 3- to 30-day scale)



Cluster Analysis of Rainfall Stations in IOWA (using 3-day scale)



Cluster Analysis of Rainfall Stations in IOWA (using 30-day scale)

