1998 International Computer Symposium
Workshop on Artificial Intelligence
December 17-19, 1998, N.C.K.U., Tainan, Taiwan, R.O.C.

A MULTISCALE STUDY OF SPATIOTEMPORAL DATA MINING – CASE STUDY: PRECIPITATION PATTERN ANALYSIS[1]

Sheng-Tun Li

Department of Information Management
National Kaohsiung First University of Technology and Science
Yenchao, Kaohsiung, TAIWAN, R.O.C.
Email: stli@ccms.nkfu.edu.tw

ABSTRACT

The scale used in the input data is one of the key issues in conducting cluster analysis in spatiotemporal data mining. Results determined from different scales of the input data could be varied. To reduce the clustering uncertainties by using a fixed time scale, it will be useful to generate a two-dimensional scale-based data which covers a range of scales as input in cluster analysis. A multiscale rotated principal component analysis approach using the continuous wavelet transform is proposed to investigate the non-stationary characteristics of time series for the study of the spatial variability in the cluster analysis of rainfall. Experimental results of precipitation pattern analysis show that the proposed approach is capable of removing the local small features by using one small scale or improving the over-smoothed regions by using one large scale input.

1. INTRODUCTION

Spatiotemporal data mining is the process of discovering meaningful patterns, trends, and correlations from large spatiotemporal data which are 2-D or 3-D (spatial and temporal) in nature using statistical and mathematical techniques. Comparing to other analytical tools, which require users to assume specific data interrelationships and then use the tools to verify these assumptions, the data mining approach employ similarity-measure and pattern-matching techniques (e.g., rule-based analysis, neural networks, fuzzy logic, K-nearest neighbor, genetic algorithms, advanced visualization, principal components analysis, etc.) to determine the key relationships in the datasets.

Cluster analysis, an approach to identify distinguishing characteristics from large datasets and then, based on their similarities, group them into a relatively small set of groups, is often the first step in data mining analysis [1]. In this study, we will focus on the development of data mining techniques used for cluster analysis of spatiotemporal data and the problem of precipitation pattern analysis will be used as a case study. In particular, we will explore the possibility of applying the continuous wavelet transform (CWT) [4] and principal component analysis (PCA) to analyze the non-stationary characteristics of time series data and examine the scale-dependent variances embedded in it via the scalogram generated from CWT. Rather than using a single scale (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 term and long term spatial variability of rainfall. The rest of this paper

[1] This work is partially supported by National Science Council, R.O.C. under Grant NSC87-2213-E-327-001.

is organized as follows. Section 2 addresses the scale issue of performing data mining on spatiotemporal data. Section 3 establishes the theoretical foundation of mutliscale study. Section 4 briefly reviews general clustering methodologies and provides a detailed discussion on rotated principal component (RPC) clustering used in the study. Experiment results for clustering the rainfall stations in Iowa, USA using different time scales are given in Section 5. Section 6 concludes the paper.

2. THE SCALE ISSUE OF SPATIOTEMPORAL DATA MINING

In precipitation pattern analysis, clustering spatiotemporal data, so-called regionalization, is based on the spatial variability of one or more physical variables (e.g., rainfall, temperature, etc.), to decompose a large complex area into several smaller homogeneous regions for various research and applications in climatology and hydrology. Recently studies in extracting spatiotemporal patterns from geoscience data sets using cluster analysis and multivariate statistics are growing rapidly [3]. In spatiotemporal data mining, in addition to concerns on quality of input data and methods used in cluster analysis, the selection of an appropriate scale plays an important role in the interpretation of the features in clusters, that means groups determined from cluster analysis might be sensitive to the scale of input data. For example, the spatial variations of precipitation patterns could be a function of temporal scales, therefore, the results from the cluster analysis of precipitation could be varied by using hourly, daily, weekly, monthly, seasonal, or annual precipitation data. The selection of an appropriate scale is really dependent on the application purpose.

A preliminary comparison could be helpful to decide the scale of the input data. For example, Van Regenmortel first evaluated the percentage cumulative variance explained by principal components analysis for daily, 5-day, 10-day, and monthly rainfall sum, then selected 10-day average rainfall to study the soil-moisture status and drought assessment [7]. In general, daily or weekly rainfall shows its 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 regionalization study. The CWT, described more detailedly in the next section, generates a two-dimensional time-scale distribution, so called scalogram, for analyzing non-stationary characteristics of rainfall. The scalogram reflects the rainfall intensity distribution over a range of scales and provides a more detailed information embedded in a one-dimensional time series for cluster analysis.

3. CONTINUOUS WAVELET TRANSFORM IN MULTISCALE STUDY

To study the non-stationary characteristics of rainfall, CWT provides the capability to investigate the temporal variation with 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 [6]:

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

where * is the complex conjugate, 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 zero mean and unit standard deviation. The magnitude of the Morlet wavelet is a Gaussian function which makes the amplitude of data are smoothed via a low-pass filter before operation. One of advantages 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 in 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 1, 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 Morelet wavelet, the radius of the COI at a point of b is $2a$. The relationship between wavelet scale and Fourier wavelength is dependent on the characteristics of the wavelet. Using the Morlet wavelet with c = 5.3, the wavelength is equivalent to the multiplication of 1.19 and the scale. Mayer discussed the impacts of edge effects in implementation and proposed a method to reduce their uncertainties [5].

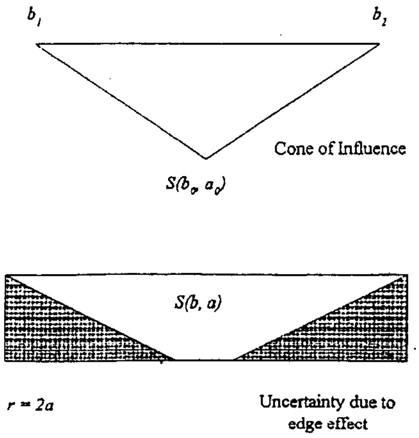

Figure 1

Wavelet variances (WV) is defined as the integration of the scalogram over time for given scales. Therefore, 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 Fourier transform. The difference is that scale is used in the WV while the frequency is used in the Fourier transform, and the scale and frequency has a reciprocal relationship.

4. CLUSTER ANALYSIS OF SPATIOTEMPORAL PATTERNS

Cluster analysis has been applied in diverse disciplines such as sciences, engineering, psychology, and behavioral sciences. The task of clustering attempts to discover groups existing in data under investigation based on measuring the degree of similarity (or dissimilarity) among them to exhibit internal cohesion and external isolation properties (high intra-cluster similarity and low inter-cluster similarity) [2]. In the present study, we perform cluster analysis on spatiotemporal data observed at rainfall stations and delimitate them into homogeneous regions.

In general, there are three clustering methodologies: hierarchical, nonhierarchical, and rotated principal component [3]. In hierarchical clustering (e.g., linkage method, Ward's method, etc.) stations can be grouped via either top-down (division) or bottom-up (merger) by partitioning patterns from a dissimilarity matrix. Nonhierarchical clustering methods (e.g., K-means, vector quantization, etc.) specify a set of centroids of K groups initially, 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.

The rotated principal component (RPC)

methodology using varimax method or oblique method tries to maximize the variance of the component loadings between each component for producing a few large loading factors and reducing else factors which makes it easy to discriminate stations. The RPC clustering differs from hierarchical and nonhierarchical clustering in overlapping solutions in which some stations could belong to more than one cluster. In a systematic methodological review, Gong and Richman performed an intercomparison of various cluster analysis methods and indicated that the rotated principal component analysis (RPCA) could be more accurate than other methods [3]. In RPCA, the number of reserved components will be determined before performing rotation. A simple way is based on the distribution of N sorted descending eigenvalues,

$$\lambda_1 \geq \lambda_2 \cdots \geq \lambda_N$$

Only the first K components, where $\lambda_k \geq$ some 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 $F_k = f_1 + f_2 + \ldots + f_K$, provides an ancillary information in the determination of number of groups. The higher F_K (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. 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 RPCA, a station is assigned to the component which has the highest loading 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 have an overlapping or open space between regions. A station in the overlapping means that it is highly associated with these overlapped regions while as the stations in an open space means that they have transition characteristics of their neighbor regions.

5. EXPERIMENTS

The rainfall stations in Iowa, United States, are used to demonstrate the results of using different time scales. We arbitrary select the data in 1992, check the associated quality flags, and reject any stations which have suspected, missing, accumulated, or invalid data. Only 70 stations are reserved after careful quality check.

5.1 Data Analysis

Figures 2 and 3 show the CWT of daily and monthly rainfall, respectively, at a rainfall station in Iowa. Only one year (1992) daily data are used in Figure 2 while 20 years' (1973-1992) monthly data are used in Figure 3.

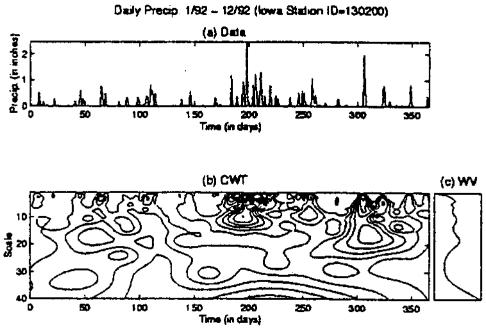

Figure 2

The associated scalograms can 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 scale is large. The non-stationary characteristics in daily rainfall are typical different from the semi-stationary characteristics of monthly rainfall. As shown in Figure 3, 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 regionalization if a range of scales is interested in applications.

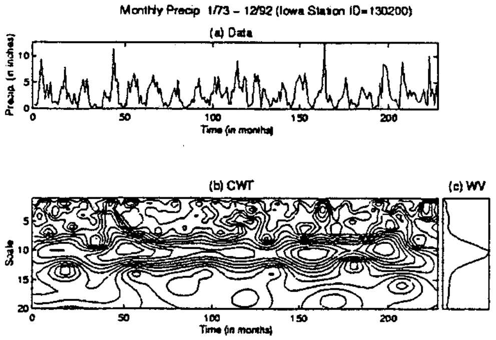

Figure 3

5.2. 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. The raw data are used in each process since no outliers have been detected. The time series from one station is assigned as one column in the correlation matrix thus the rotated PCA is applied to cluster stations with similar temporal patterns in this regionalization study. Figure 4 shows the loading factors of the first four principal components using the scalogram of 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 5 displays the mosaiced regions derived from Figure 4 where the central small region is corresponding to the 5th loading factor. There are several stations (e.g., stations 5, 34, 57, 41, etc.) are not firmly linked to one cluster. They are located in transition zones between two or more regions.

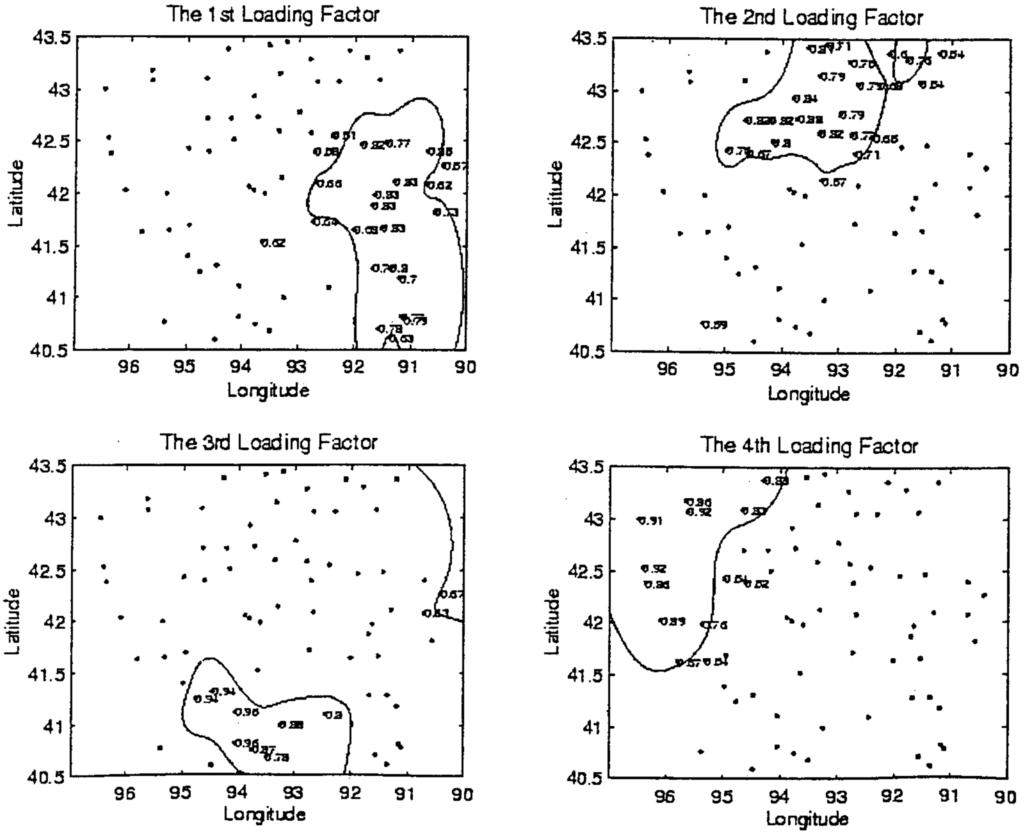

Figure 4

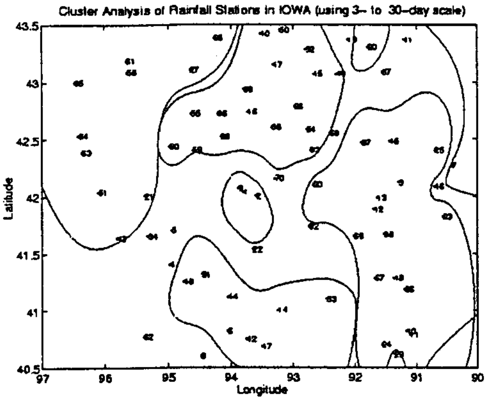
Cluster Analysis of Rainfall Stations in IOWA (using 3- to 30-day scale)
43.5
43
42.5
42
41.5
41
40.5
Latitude
97
96
95
94
93
92
91
90
Longitude

Figure 5

These stations can be assigned to one or more clusters when the threshold used in the contour line decreases. That could generate some overlapping areas among regions. Rotated PCA objectively provides the loading factors, but it is a little subjective to make a decision for the selection of threshold in grouping.

Figures 6 to 8 show the regions using the 3-day, 15-day, and 30-day rainfall data, respectively.

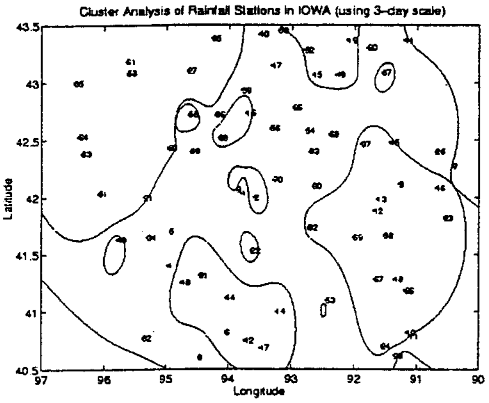
Cluster Analysis of Rainfall Stations in IOWA (using 3-day scale)
43.5
43
42.5
42
41.5
41
40.5
Latitude
97
96
95
94
93
92
91
90
Longitude

Figure 6

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.

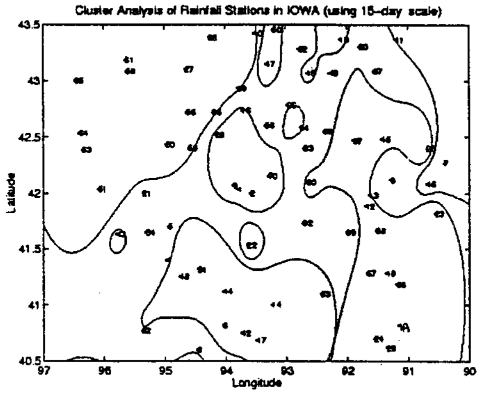
Cluster Analysis of Rainfall Stations in IOWA (using 15-day scale)
43.5
43
42.5
42
41.5
41
40.5
Latitude
97
96
95
94
93
92
91
90
Longitude

Figure 7

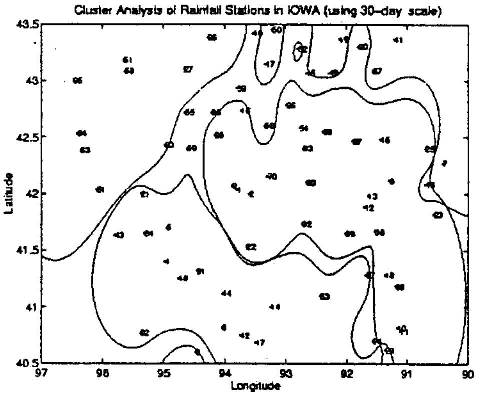
Cluster Analysis of Rainfall Stations in IOWA (using 30-day scale)
43.5
43
42.5
42
41.5
41
40.5
Latitude
97
96
95
94
93
92
91
90
Longitude

Figure 8

6. CONCLUSIONS

We have addressed the necessity of conducting a multiscale study for spatiotemporal'data mining using CWT. The CWT provides an option to consider a 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 study. We review the general clustering methodologies and propose a rotated principal component methodology for clustering spatiotemporal data based on

CWT. Experiment results show that the multiscale clustering can effectively analyze the scale-dependent variances inherent in time series data. Further investigation into comparing RPC to other modern clustering approaches, for example, self-organization maps in artificial neural networks area and its application to more complicated problems are undergoing.

7. REFERENCES

[1] Bigus, J. P., Data Mining with Neural Networks, McGraw-Hill, New York, 1996.

[2] Everitt, B., Cluster Analysis, Wileym Halsted Press, New York, 1980.

[3] Gong, X. and Richman, M. B., 'On the application of cluster analysis to growing season precipitation data in north America east of the Rockies', *J. Climate* , 8, 897-931, 1995.

[4] Meyers, S. D. Kelly, B. G. and O'Brien, J. J., 'An introduction to wavelet analysis in oceanography and meteorology: With application to the dispersion of Yanai Waves', *Mon. Wea. Rev.*, 121, 2858-2878, 1993.

[5] Mallat, S., "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation," IEEE Pattern Analysis and Machine Intelligence, Vol. 11, No. 7, pp. 674-693, 1989.

[6] Morlet, J., Arens, G., Fourgeau, I. and Giard, D., "Wave propagation and sampling theory", 1982.

[7] Van Regenmortel, G., "Regionalization of Botswana Rainfall During the 1980s using Principal Component Analysis", Int'l., J. Climatol, Vol. 15, pp. 313-323, 1995.