Visual Data Mining for Identification of Patterns and Outliers in Weather Stations' Data

Jose Roberto Motta Garcia[1], Antonio Miguel Vieira Monteiro[2] and Rafael Duarte Coelho dos Santos[2]
[1]Programa de Mestrado ou Doutorado em Computação Aplicada – CAP
Instituto Nacional de Pesquisas Espaciais – INPE

[2]Laboratório Associado de Computação e Matemática Aplicada – LAC
Instituto Nacional de Pesquisas Espaciais – INPE

roberto.garcia@cptec.inpe.br, miguel@dpi.inpe.br,
rafaeldcsantos@gmail.com

Abstract. *Quality control of climate data of weather stations is essential to ensure reliability of research and services. A way to do it is comparing data of one station with data of close stations which somehow are expected to have similar behavior. The purpose of this work is to evaluate some visual data mining techniques to identify groupings/outliers of weather stations using historical precipitation data in a specific time interval. We present and discuss the techniques' details, variants, results and applicability on this problem.*

Keywords: visual data mining, clustering, self-organizing map, fuzzy c-means

1. Introduction

Observational data obtained from weather stations are important due to its use on generating weather and climate numeric predictions, evaluating models results and making climatic research [1] and reliable data is essential to have accurate research. However, this data is not fully reliable: some weather stations are still human operated, which often are subject to reporting errors; and even the automatic ones depend on hardware and network communication which can pollute the data [2]. A quality control system is clearly required to verify the data's quality.

At National Center for Weather Prediction and Climate Studies (CPTEC) part of the Brazilian National Institute for Space Research's (INPE) there is a 3-level quality control system for weather stations data. All of them use specific upper and lower limits to classify the data. The first approach verifies the limits for each data, not considering from where it is, the second considers arbitrary geographic rectangular regions and the third one is specific for a weather station. These controls aim to reject spurious data and classify suspicious data [3].

But nature does not work on rectangular regions and, moreover, if an unexpected event happens and it generates higher/lower values than what is established, the data can be rejected or wrongly classified. Also, the number of rejections increases the work by

creating the need to analyze all of them. Clearly, some other ways to interpret right or wrong data are required. This work presents some algorithms and implementations based on visual data mining that aims to help to fulfill this task using human perception.

Section 2 presents some visual data mining concepts relevant to this work. Section 3 presents the data used in this study and its relevant features. Sections 4 and 5 present two algorithms which implement the visual data mining tasks. Section 6 presents some conclusions and directions for future work.

2. Visual Data Mining

VDM can be defined as the set of techniques and approaches used to extract and understand the information encoded in data sets using the human visual perception system as part of the data processing task [4, 5]. VDM may help uncover or highlight data features, may make the understanding of the features easier and faster and may be used to cross-validate conclusions obtained through other methods [6, 7].

The differences between VDM and traditional data visualization are somehow blurred: for our purposes we consider that VDM tasks involve the processing of the data with one or more data mining algorithms and that visualization is done over the original data together with information obtained from those algorithms. VDM may also be considered one of the components of exploratory data analysis (EDA) and strongly related to visual analytics [8, 9].

VDM tools and approaches are used in several knowledge domains, e.g. analysis of environmental, geophysical and atmospheric data [6, 10, 11], astronomy and astrophysics data [12], health data [13], education evaluation [14], performance evaluation of companies [15], network traffic analysis [16, 17], fraud detection [18], security and forensics [19], etc.

3. Data

The data used to mine and visualize is inherently spatiotemporal: time series from sensors with geographical coordinates associated. It was collected from a daily precipitation dataset for all Brazil. It was 115 million records and some weather stations with more than 100 years of recorded data.

In order to evaluate the visual data mining techniques we've selected only data from stations of São Paulo state, and created for each station a time series containing the monthly accumulated precipitation. Only stations that could yield at least 25 readings in a month were considered. The final dataset had data from 1.341 weather stations with a time series with 84 entries, corresponding to monthly accumulated precipitation.

Several visualization techniques can be used to get some basic information about the behavior of this type of data. The most frequently used are time series plots [10] or parallel coordinates plots [20] (with each horizontal axis mapped to a time coordinate). Figure 1 shows a plot of all data from the 1.341 weather stations. From Figure 1 we can see monthly and global extreme and get a feeling of the behavior of the whole set of time series: there are periods with more and less accumulated precipitation which are more or less correspondent to the wet and dry seasons.

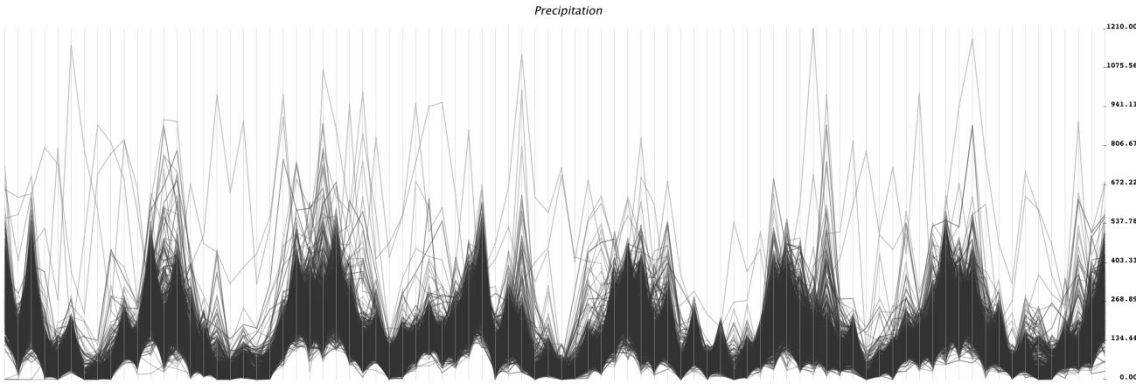

Figure 1. Time series plot of the data used in this work.

At the same time we can observe that there isn't a clear global maximum or minimum and visual identification of which station is providing data outside of a range is not trivial. Although this kind of plot provides some general information about the series it does not conveys any information of behavior similarity (or anomaly) between stations – in other words we cannot infer whether two or more stations have similar or different behaviors nor identify geographically close weather stations – this is important because we expect that weather patterns have a geographic extent and that would influence data from stations that are close to this pattern.

Since the data is inherently geographic (stations' coordinates are points in space), it is natural to visualize them overlaid in a map. The problem is that the data associated to each point in the map is a time series – in order to visually identify behavior similarity we would need to reduce the time dimension so it can also be plotted in the map and used in visual comparison with other points in the map.

In this work we investigate two approaches to remap the time series: the first is based on a common clustering algorithm and the other on a well-known dimensionality reduction algorithm. These techniques and results are presented in Sections 4 and 5.

4. Fuzzy C-Means Clustering

Fuzzy C-Means is a clustering algorithm which partitions data in a set of groups or clusters [21–23]. This iterative algorithm attempts to minimize an objective function J defined as:

$$J = \sum_{k=1}^{n} \sum_{i=1}^{c} \mu_{ik}^{m} |x_k - v_i|^2 , m > 1 \quad (1)$$

Where $\boldsymbol{x_k}$ is the k-th data vector, $\boldsymbol{v_i}$ is the i-th cluster center vector, $\boldsymbol{c}$ is the number of clusters,$\boldsymbol{n}$ is the number of points in the data, $\boldsymbol{\mu}$ is the membership values' table or matrix which contains the membership values for all points in all clusters; which indicates to which degree or extent the data vector $\boldsymbol{x_k}$ belongs to the cluster vi, and m is a fuzziness value. The membership values are subject to the conditions $\boldsymbol{0 \leq u_{ik} \leq 1}$ and $\sum_{i=1}^{c} \boldsymbol{u_{ik}} = \boldsymbol{1}$ for all k.

One advantage of this algorithm is its ability to create a membership table which can be used in a *defuzzification* step to determine the best cluster for each data vector. The

same membership table can also be used to identify equal or almost equal membership of a data point in two or more clusters, which may indicate a data vector that can't be satisfactorily assigned to a cluster.

One problem with the Fuzzy C-Means algorithm is the definition of its parameters. In particular two parameters are often defined experimentally: the number of clusters C and the fuzziness factor m. Of those, determination of a suitable m is relatively easy: when m is close to 1 the algorithm behaves like the non-fuzzy K-Means; while when m is large enough all data may have equal membership in all clusters. For some applications empirical values of m between 1.5 and 2.5 are suggested [24, 25].

C, the number of clusters, is often empirically determined, although some metrics of cluster validity may be employed to find a suitable value for C. Three of those metrics are the partition coefficient, the partition entropy and the compactness and separation metrics [23]. These metrics are calculated after the data is clustered with several values of C and the best metric for a particular C can be used (maximum value for partition coefficient; minimum value for the others).

In order to use the Fuzzy C-Means algorithm to map the time series into a fixed number of clusters we've executed experiments with some values of m and several values of C to determine the best values for clustering. Five arbitrary values were used for m (1.01, 1.125, 1.25, 2.5, 5), while values from 2 to 25 were used for C. Other parameters for the algorithm that control the number of interactions were left large enough to ensure convergence. From this experiment we've concluded that for large values of m (# 2.5) the results were practically indistinguishable, which led us to use m = 1.25. Determination of C was harder since there wasn't a single value of C that was a clear minima or maxima for the validity measures. We've used C = 7 as it seemed slightly better than other possible values accordingly to the compactness and separation metric.

The Fuzzy C-Means algorithm will yield two results we want to use to reduce the time dimension in our data for visual mining: a discrete cluster number that Visual Data Mining for Weather Stations' Data 5 will be used to select distinct colors for plotting and the maximum membership value for each data vector. This value is obtained from the rows in the matrix μ and ranges from 1/C to 1, where higher values indicate stronger membership in a given cluster, and can be considered an indicator of quality of clustering for a particular data vector. The maximum membership value for each data vector was used to determine the size of the point to be plotted over the map.

Figure 2 shows the map with the data points plotted over it. Colors are arbitrary, points assigned (through *defuzzification*) to the same cluster have the same color. Sizes of the plotted points are relative to the maximum membership value for the point: smaller points have smaller maximum membership for all clusters. It must be pointed that the coordinates for the stations were not used in the clustering process itself, but only for the map generation.

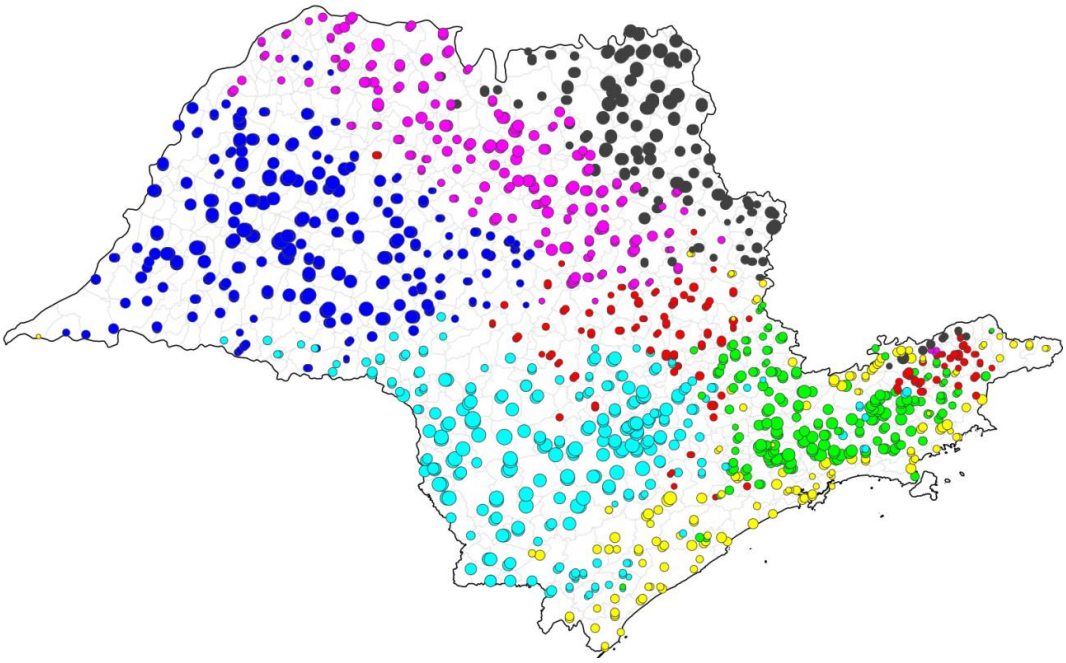

Figure 2. Visualization using the Fuzzy C-Means results.

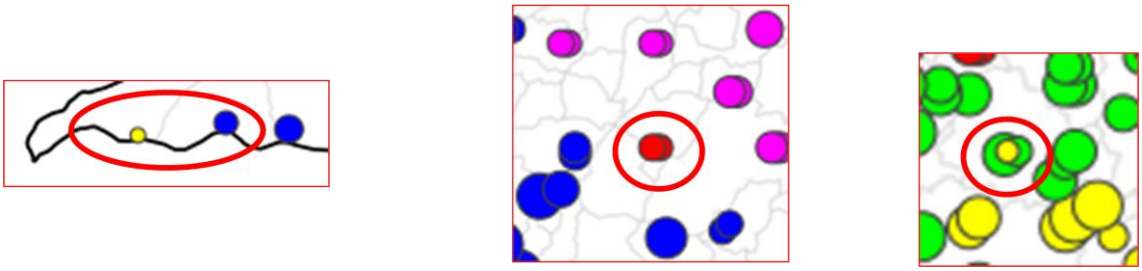

Figure 3. Details of Figure 2.

The map shown in Figure 2 and its details (Figure 3) presents clusters that are mostly contiguous in space, confirming our expectative that weather phenomena (in this case, precipitation) have a moderate spatial correlation. Outliers (points that were assigned to clusters different from points nearby) are also easily identified due to the use of different colors, and the point size can also be used to identify data vectors that were strongly or weakly assigned to their clusters (e.g. southern large point in the middle figure, small sub-clusters on the right).

5. Self-Organizing Maps

Another way to reduce the 84-dimension time vector to some few variables that can be plotted in a map is using the Self-Organizing Map [26] or SOM. This well-known neural network-based algorithm is able to reduce the dimension of a data set while preserving its topology: in other words, data vectors which were close in the original feature space will appear close in the new topology. The SOM uses as input the data and some parameters, the most important being the number of neurons that will be organized in a regular grid, and gives as output the best matching neuron for a particular data vector.

The topology-preserving feature of the SOM is particularly interesting for us: if we use

an adequate representation for the neurons we can plot points that will appear similar if the clusters are similar. One natural choice for graphical representation of the neurons is to use a hue-based color system and map the hue and saturation values to the neurons in such a way that neurons that are topologically close have similar colors associated with them. Some of those mappings for a 2-dimensional, squared-lattice SOM are shown in Figure 4 (from the left to the right: mappings on a 3x3, 9x9 and 15x15 SOM).

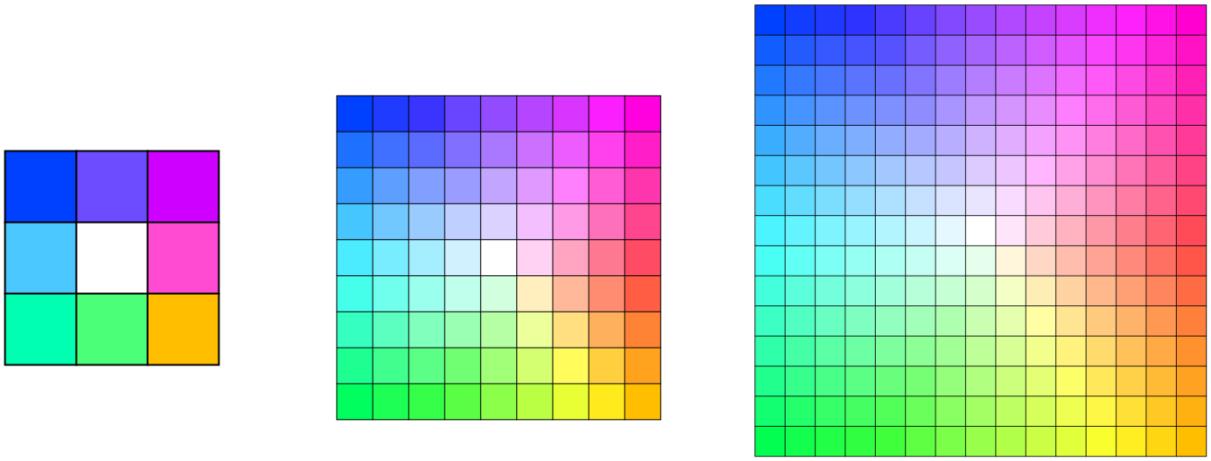

Figure 4. Mapping of colors to neurons in SOMs with different sizes.

Some authors (e.g. [27]) suggest that the number of neurons in a SOM should be a function of the number of input vectors. For our purposes we tried to visualize the data points with colors chosen from the color tables similar to those shown in Figure 4 and several numbers of neurons. The best results for visualization were achieved with SOMs of 9x9 and 15x15 neurons, but surprisingly, even SOMs of size 3x3 yielded easily interpretable results: data corresponding to weather stations with behavior different from the geographic neighbors were 7 plotted in different colors. The map created with the processing of the data with a 15x15 SOM is shown in Figure 5 and 6, that shows the general visual clustering structure for the data (data from weather stations geographically close have similar colors) and also some outliers (e.g. left part of Figure 6). One advantage of using a topology preserving algorithm is that we can perceptually evaluate how much a data point is different from the others.

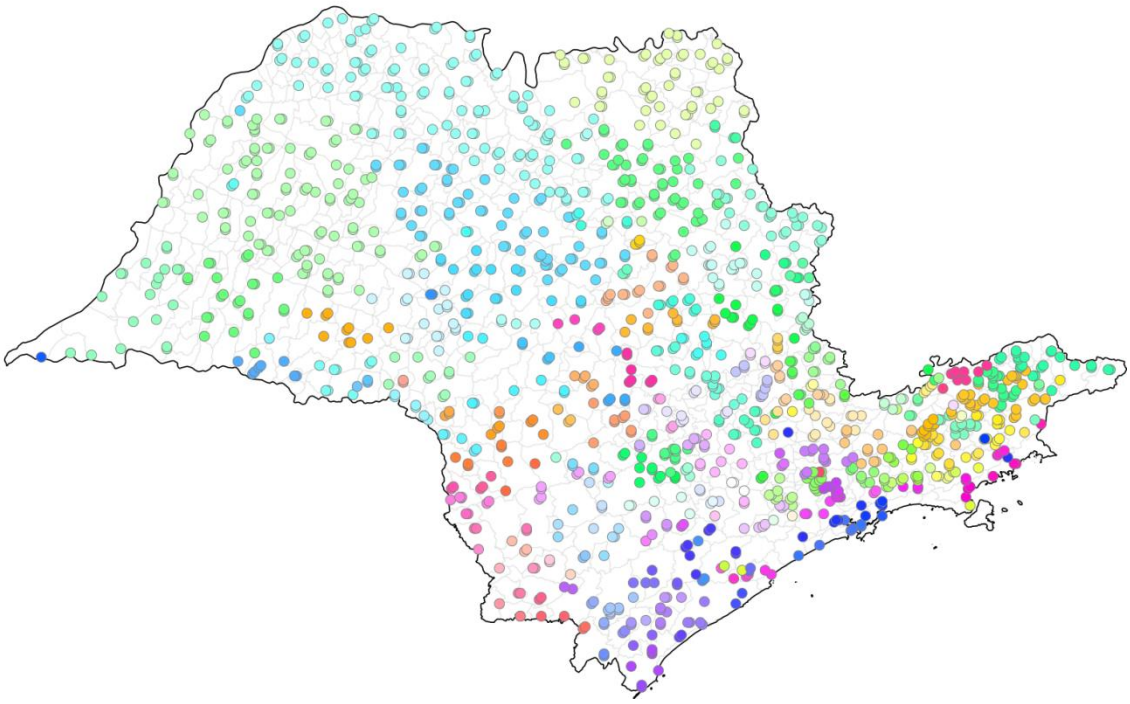

Figure 5. Visualization using the SOM results.

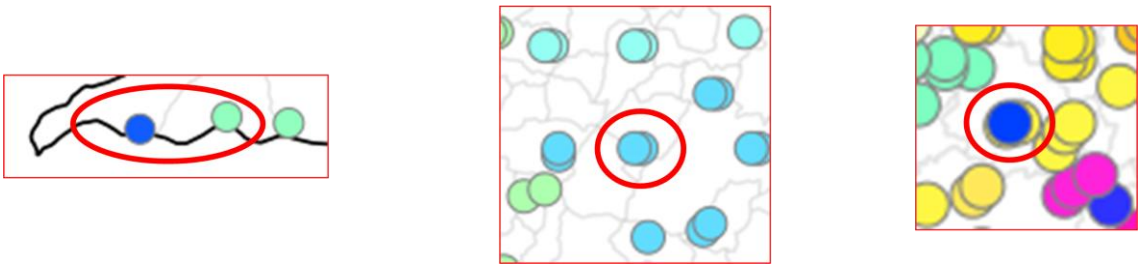

Figure 6. Details of Figure 5.

6. Conclusions and Future Work

In this paper we evaluated two techniques for dimension reduction or mapping in order to create visual representation of time series over geographic coordinates. Ultimately these techniques may be incorporated on the data collection systems at INPE's CPTEC to help identify potentially problematic data subsets.

Of the two techniques the one based on the SOM was considered to be more easily interpretable since it is possible to identify data points that are different from a cluster and at the same time perceptually evaluate how much it is different. Both techniques have been used by researchers in several domains, but due to its features SOM-based techniques are more prevalent, particularly for analysis of data with spatial components (e.g. [28, 29]).

References

Kalnay, E. (2003), Atmospheric Modeling, Data Assimilation and Predictability. 1st edn. Cambridge Press. [1]

Expert Team on Requirements of Data from Automatic Weather Stations: Final report. (2002), http://www.wmo.int/pages/prog/www/OSY/Meetings/ET-AWS1-2002/Final-Report.pdf [2]

Garcia, J.R.M., Carvalho, L.S.M., Júnior, H.C., Sanches, M.B.: Bdc - banco de dados climatológico. (2006), In: Proceedings do XIV Congresso Brasileiro de Meteorologia. [3]

Simoff, S., Böhlen, M., Mazeika, A. : Visual data mining: An introduction and overview. In Simoff, S., Böhlen, M., Mazeika, A., eds.: Visual Data Mining. Volume 4404 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg 1–12 [4]

Keim, D., Panse, C., Sips, M. : Visual data mining of large spatial data sets. In Bianchi-Berthouze, N., ed.: Databases in Networked Information Systems. Volume 2822 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg 201–215 [5]

Macêdo, M., Cook, D., Brown, T. Visual data mining in atmospheric science data. Data Mining and Knowledge Discovery 4 69–80 [6]

Kopanakis, I., Pelekis, N., Karanikas, H., Mavroudkis, T.: Visual techniques for the interpretation of data mining outcomes. In Bozanis, P., Houstis, E., eds.: Advances in Informatics. Volume 3746 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg 25–35 [7]

Andrienko, N., Andrienko, G. (2006) Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach. Springer. [8]

Huang, M., Nguyen, Q.: Context visualization for visual data mining. In Simoff, S., Böhlen, M., Mazeika, A., eds.: Visual Data Mining. Volume 4404 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg 248–263 [9]

Andrienko, G., Andrienko, N., Gatalsky, P.: Visual mining of spatial time series data. In Boulicaut, J.F., Esposito, F., Giannotti, F., Pedreschi, D., eds.: Knowledge Discovery in Databases: PKDD 2004. Volume 3202 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg 524–527 [10]

Watanabe, C., Touma, E., Yamauchi, K., Noguchi, K., Hayashida, S., Joe, K.: Development of an interactive visual data mining system for atmospheric science. In Labarta, J., Joe, K., Sato, T., eds.: High-Performance Computing. Volume 4759 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg 279–286. [11]

Feigelson, E.D., Babu, G.J., Cook, D.: Interactive and dynamic graphics for data analysis: A case study on quasar data. In: Statistical Challenges in Astronomy. Springer New York 255–264 [12]

Fisher, P., Koua, E., Kraak, M.J.: Integrating computational and visual analysis for the exploration of health statistics. In: Developments in Spatial Data Handling. Springer Berlin Heidelberg 653–664 [13]

Ertek, G.: Visual data mining for developing competitive strategies in higher education. In Cao, L., Yu, P.S., Zhang, C., Zhang, H., eds.: Data Mining for Business

Applications. Springer US 253–266 [14]

Liu, H., Eklund, T., Back, B., Vanharanta, H.: Visual data mining: Using self-organizing maps for electricity distribution regulation. In Ariwa, E., El-Qawasmeh, E., eds.: Digital Enterprise and Information Systems. Volume 194 of Communications in Computer and Information Science. Springer Berlin Heidelberg 631–645. [15]

Hao, M., Dayal, U., Hsu, M.: Visual data mining for business intelligence applications. In Lu, H., Zhou, A., eds.: Web-Age Information Management. Volume 1846 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg 3–14. [16]

Han, W., Wang, J., Shaw, S.L.: Visual exploratory data analysis of traffic volume. In Gelbukh, A., Reyes-Garcia, C., eds.: MICAI 2006: Advances in Artificial Intelligence. Volume 4293 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg 695–703 [17]

Cox, K.C., Eick, S.G., Wills, G.J., Brachman, R.J.: Brief application description; visual data mining: Recognizing telephone calling fraud. Data Mining and Knowledge Discovery 1 225–231 [18]

Francia, G., Trifas, M., Brown, D., Francia, R., Scott, C.: Visual data mining of log files. In Sobh, T., ed.: Innovations and Advanced Techniques in Computer and Information Sciences and Engineering. Springer Netherlands 531–535 [19]

Inselberg, A. (2009), Parallel Coordinates – Visual Multidimensional Geometry and Its Applications. Springer. [20]

Bezdek, J.C. (1987), Pattern Recognition with Fuzzy Objective Function Algorithms. 1st edn. Plenum Press. [21]

Bezdek, J.C., Pal, S.K. (1992), Fuzzy Models for Pattern Recognition. 1st edn. IEEE Press. [22]

Chi, Z., Yan, H., Pham, T. (1996), Fuzzy Algorithms with Applications to Image Processing and Pattern Recognition. World Scientific Publishing. [23]

Yang, M.S., Wu, K.L. (2006), Unsupervised possibilistic clustering. Pattern Recogn. 39(1), 5–21. [24]

Wu, K.L. (2012), Analysis of parameter selections for fuzzy c-means. Pattern Recogn. 45(1), 407–415. [25]

Kohonen, T. (1997), Self-Organizing Maps. 2nd edn. Springer. [26]

Barreto, G.: Time series prediction with the self-organizing map: A review. In Hammer, B., Hitzler, P., eds.: Perspectives of Neural-Symbolic Integration. Volume 77 of Studies in Computational Intelligence. Springer Berlin / Heidelberg 135–158 [27]

Friedel, M.J.: Climate Change Effects on Ecosystem Services in the United States – Issues of National and Global Security. In Baba, A., Tayfur, G., G¨und¨uz, O.J., Howard, K.W., Friedel, M.J.W.F., Chambel, A., eds.: Climate Change and its Effects on Water Resources. Volume 3 of NATO Science for Peace and Security Series C: Environmental Security. Springer Netherlands 17–24. [28]

Koua, E., Kraak, M.J.: Geovisualization to support the exploration of large health and demographic survey data. International Journal of Health Geographics 3 1–13.[29]