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

José Roberto M. Garcia, Antônio Miguel V. Monteiro, and Rafael D.C. Santos

Brazilian National Institute for Space Research,
Av dos Astronautas, 1.758, Jd. Granja - CEP 12227-010,
São José dos Campos – São Paulo – Brasil

Abstract. Quality control of climate data obtained from weather stations is essential to ensure reliability of research and services based on this data. One way to perform this control is to compare data received from one station with data from other stations which somehow are expected to show similar behavior. The purpose of this work is to evaluate some visual data mining techniques to identify groupings (and outliers of these groupings) 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 type of problem.

Keywords: Visual data mining, clustering, self-organizing map, fuzzy C-means.

1 Introduction

Observational data obtained from weather stations is important due to its use on generating weather and climate numeric predictions, evaluating models results and making climatic research [1], so having reliable data is an essential issue to make reliable research and applications. However, the data is not completely 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 the Brazilian National Institute for Space Research's (INPE) CPTEC (Brazilian National Center for Weather Prediction and Climate Studies) there is a 3-level quality control system for weather stations data. The first approach verifies whether the data is inside upper and lower limits to the variable; the second uses arbitrary geographic rectangular regions and limits on the variables on these regions, and the third one uses limits for each variable and specific weather station. These controls aims to reject spurious data and classify suspicious data [3].

The problem with these approaches is that the bounds used (for the variable and geographic regions) are not natural and can filter important data from datasets for analysis. Moreover, the number of rejections increases the work of the data administrator that need to analyze, one by one, all the rejected data.

H. Yin et al. (Eds.): IDEAL 2012, LNCS 7435, pp. 245–252, 2012.

Clearly some other methods that help meteorologists interpret the vast amount of data in search for potential suspicious data are required. This work presents some possible algorithms and implementations based on visual data mining to accomplish this task.

This paper is divided in the following sections: 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 presents two of the algorithms that were selected for implementation of the visual data mining tasks. Section 6 presents some conclusions and directions for future work.

2 Visual Data Mining

Visual data mining 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]. Visual data mining may help uncover or highlight data features, may make the understanding of the features easier and speedier and may be used to cross-validate conclusions obtained through other methods [6,7].

The difference between visual data mining and traditional data visualization is somehow blurred: for our purposes we consider that visual data mining 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. Visual data mining may also be considered one of the components of exploratory data analysis (EDA) and strongly related to visual analytics [8,9].

Visual data mining tools and approaches are used in several knowledge domains, including analysis of environmental, geophysical and atmospheric data [6,10,11], which are relevant to our interests.

3 Data

The data we want to mine and visualize is inherently spatio-temporal: time series collected from sensors with geographical coordinates associated. The data was selected from a database containing daily precipitation data for all Brazil, with a total of 115 million records and with some weather stations having more than 100 years of recorded data.

In order to evaluate the visual data mining techniques we've selected only data from stations on the state of São Paulo, 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 database contained data from 1.341 weather stations (including geographic coordinates and altitude) with a time series with 84 entries corresponding to monthly accumulated precipitation. The data for each weather station covers the same period in time.

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 [12] (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 extrema 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. At the same time we can observe that there isn't a clear global maxima or minima, 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 anomalies) 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.

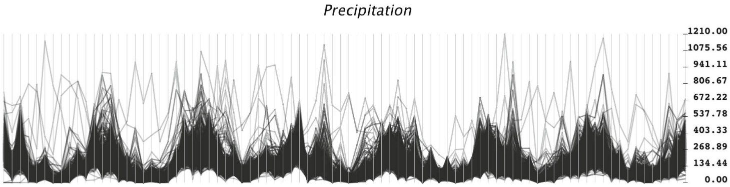

Fig. 1. Time series plot of the data used in this work

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 multidimensional time series – in order to visually identify behavior similarity we would need to reduce the time dimension or extract fewer features from it so it can also be plotted in the map and used for visual comparison with other points in the map.

In this work we investigate two approaches to extract features from the time series that can be used in for visual data mining: 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 [13,14] is an iterative algorithm that attempts to minimize an objective function J defined as:

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

Where x_k is the k-th data vector, v_i is the i-th cluster center vector, c is the number of clusters, n is the number of points in the data, μ 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 x_k belongs to the cluster v_i, and m is a fuzziness value. The membership values are subject to the conditions $0 \leq u_{ik} \leq 1$ and $\sum_{i=1}^{c} u_{ik} = 1$ for all k.

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 [15,16].

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 [14]. 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 $C = 13$ 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 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.

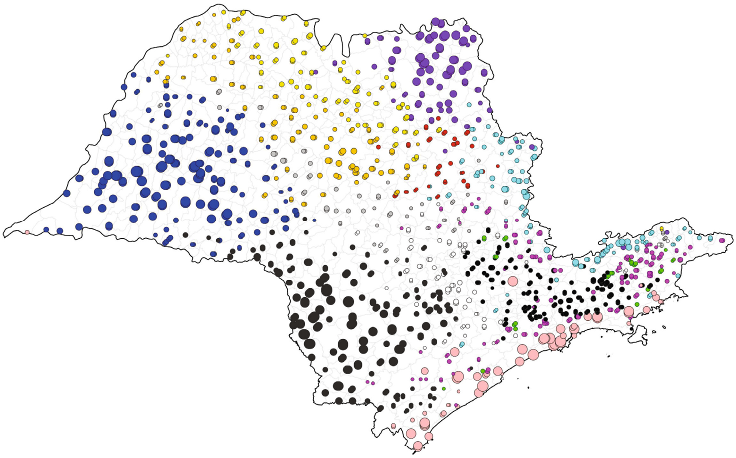

Fig. 2. Visualization using the Fuzzy C-Means results

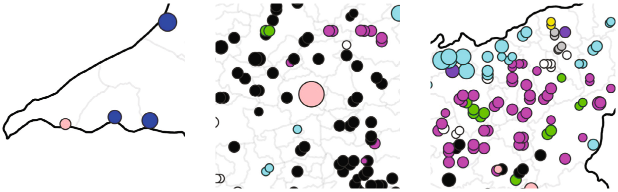

Fig. 3. Details of Figure 2

The map shown in Figure 2 and its details (Figure 3) presents clusters that are mostly contiguous in space, confirming our expectatives 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. central 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 [17] 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 visually 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 3×3, 9×9 and 15×15 SOM).

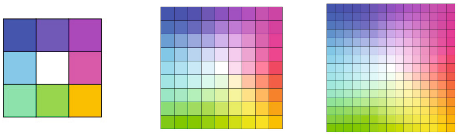

Fig. 4. Mapping of colors to neurons in SOMs with different sizes

Some authors (e.g. [18]) 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 9×9 and 15×15 neurons, but surprisingly, even SOMs of size 3×3 yielded easily interpretable results: data corresponding to weather stations with behavior different from the geographic neighbors were plotted in different colors. The map created with the processing of the data with a 15×15 SOM is shown in Figure 5, with some details shown in Figure 6.

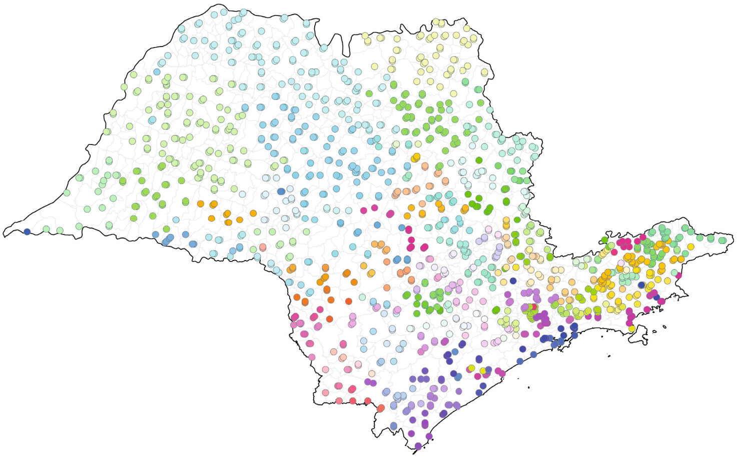

Fig. 5. Visualization using the SOM results

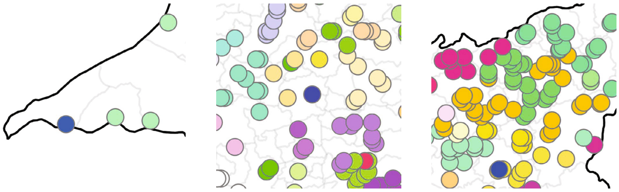

Fig. 6. Details of Figure 5

Figures 5 and 6 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.

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. We must point that we did not cluster time series for classification or characterization, which, depending on the metric used, can be considered meaningless [19,20]. In this research we used the metrics extracted from different clustering algorithms to visualize spatio-temporal data.

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 has 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. [21]).

References

1. Kalnay, E.: Atmospheric Modeling, Data Assimilation and Predictability, 1st edn. Cambridge Press (2003)
2. 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`
3. Garcia, J.R.M., Carvalho, L.S.M., Júnior, H.C., Sanches, M.B.: BDC - banco de dados climatológico. In: Proceedings do XIV Congresso Brasileiro de Meteorologia (2006)
4. Simoff, S.J., Böhlen, M.H., Mazeika, A.: Visual Data Mining: An Introduction and Overview. In: Simoff, S.J., Böhlen, M.H., Mazeika, A. (eds.) Visual Data Mining. LNCS, vol. 4404, pp. 1–12. Springer, Heidelberg (2008)

5. Keim, D., Panse, C., Sips, M.: Visual Data Mining of Large Spatial Data Sets. In: Bianchi-Berthouze, N. (ed.) DNIS 2003. LNCS, vol. 2822, pp. 201–215. Springer, Heidelberg (2003)
6. Macêdo, M., Cook, D., Brown, T.: Visual data mining in atmospheric science data. Data Mining and Knowledge Discovery 4, 69–80 (2000)
7. Kopanakis, I., Pelekis, N., Karanikas, H., Mavroudkis, T.: Visual Techniques for the Interpretation of Data Mining Outcomes. In: Bozanis, P., Houstis, E.N. (eds.) PCI 2005. LNCS, vol. 3746, pp. 25–35. Springer, Heidelberg (2005)
8. Andrienko, N., Andrienko, G.: Exploratory Analysis of Spatial And Temporal Data: A Systematic Approach. Springer (2006)
9. Huang, M.L., Nguyen, Q.V.: Context Visualization for Visual Data Mining. In: Simoff, S.J., Böhlen, M.H., Mazeika, A. (eds.) Visual Data Mining. LNCS, vol. 4404, pp. 248–263. Springer, Heidelberg (2008)
10. Andrienko, G., Andrienko, N., Gatalsky, P.: Visual Mining of Spatial Time Series Data. In: Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) PKDD 2004. LNCS (LNAI), vol. 3202, pp. 524–527. Springer, Heidelberg (2004)
11. 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.) ISHPC 2006 and ALPS 2006. LNCS, vol. 4759, pp. 279–286. Springer, Heidelberg (2008)
12. Inselberg, A.: Parallel Coordinates – Visual Multidimensional Geometry and Its Applications. Springer (2009)
13. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms, 1st edn. Plenum Press (1987)
14. Chi, Z., Yan, H., Pham, T.: Fuzzy Algorithms with Applications to Image Processing and Pattern Recognition. World Scientific Publishing (1996)
15. Yang, M.S., Wu, K.L.: Unsupervised possibilistic clustering. Pattern Recogn. 39(1), 5–21 (2006)
16. Wu, K.L.: Analysis of parameter selections for fuzzy c-means. Pattern Recogn. 45(1), 407–415 (2012)
17. Kohonen, T.: Self-Organizing Maps, 2nd edn. Springer (1997)
18. Barreto, G.: Time Series Prediction with the Self-Organizing Map: A Review. In: Hammer, B., Hitzler, P. (eds.) Perspectives of Neural-Symbolic Integration. SCI, vol. 77, pp. 135–158. Springer, Heidelberg (2007)
19. Keogh, E., Lin, J.: Clustering of time-series subsequences is meaningless: implications for previous and future research. Knowledge and Information Systems 8, 154–177 (2005)
20. Chen, J.: Making subsequence time series clustering meaningful. In: Fifth IEEE International Conference on Data Mining, 8 p. (November 2005)
21. 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 (2004)