Clustering and Analyzing Air Pollution Data using Self-Organizing Maps

Minji Maria Lee
Institute of Geography
Berliner Straße 48
Heidelberg, Germany
mariaminjilee@gmail.com

Enrico Steiger
Institute of Geography
Berliner Straße 48
Heidelberg, Germany
enrico.steiger@geog.uni-heidelberg.de

Alexander Zipf
Institute of Geography
Berliner Straße 48
Heidelberg, Germany
Alexander.zipf@geog.uni-heidelberg.de

Abstract

In Geographic Information Science (GIScience), the rise in the availability of spatial data paved ways for increased research in different spatial data mining techniques. This paper presents a Spatial Self-Organizing Map (Spatial SOM) for analysing high-dimensional and complex spatial datasets. The results of the selected case study with air pollution data for the US has demonstrated that Spatial SOMs are an effective and efficient analysis tool with the ability to explore, detect, and visualize latent spatial structures and characteristics within large datasets.

Keywords: Self-Organizing Map, Data Mining, Spatial Analysis, Air Pollution Data

1 Introduction

With the advancement of technology and the widespread use of devices connected to global positioning system, more and more spatial data is being collected and stored in databases. Although offering valuable information and knowledge, there are several challenges in processing spatial big data due to its large size and unique spatial properties such as spatial dependency. In previous research, Kohonen's Self-Organizing Map [1] has been proven effective to explore, analyse, and understand latent structure of multi-dimensional data.

In GIScience, the rise in the availability of spatial data paved ways for increased research in different techniques of spatial data mining. One fundamental property that has been central in spatial data mining research is spatial dependence [2], which postulates that entities near in distance share more similarities than those that are far apart [3]. Although this dependency can be viewed as confounding, they can be valuable sources of underlying geographical phenomena in spatial data [4].

2 Background

Several studies have demonstrated strengths in Self-Organizing Map algorithm (SOM) over other data mining methods [5, 6]. Originated from the Computer Science field, SOM is an unsupervised neural network algorithm for data clustering and analysing. It is a technique used for reducing multi-dimensional data into a lower dimensional map, by mapping input data to neurons in a topologically ordered manner. For this reason, SOM is easy to visualize multi-dimensional complex data, and thus, have been of great interest in GIScience [7]. So far, there have been several attempts to integrate spatial dependence in SOM algorithm. For example a Geographical Hypermap and the Spatial Kangas Map, algorithms which are inspired by [8, 9] and GeoSOM [7] respectively, integrate spatial dependence by giving more weight on geospatial variable of the input data. Hence, although research efforts have been made in recent years in the field of spatial data mining and SOM [10], there is no common approach to tackle the growing amount of spatial data. Especially, the amount of environmental sensor data being collected on air pollution is raising very quickly. Therefore, the goal of the study is to develop an integrated approach that could provide us with an insight to tackling the problems of large complex heterogeneous data we face today.

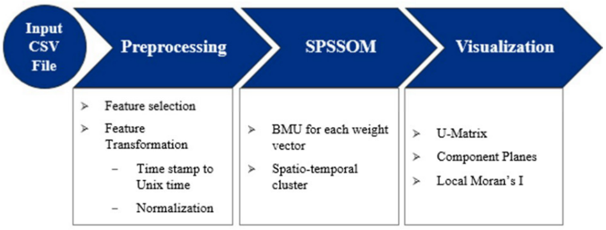
In the study, we aim to answer the following question that remain open in the study of GIScience:
How can we analyse and understand spatial characteristics of given sensor observations using SOMs?

In order to answer this research question, we apply a Spatial SOM and test its performance in terms of both speed and quality in clustering and analysing spatial big data. The study shows that SOM is a promising approach to explore and analyse latent pattern within large air pollution dataset in a distinct use case.

2.1 Implementation of Spatial SOM

Figure 1 depicts the analysis framework used to analyse our air pollution dataset. The SOM can be divided into roughly three parts: 1. initializations, 2. training and collecting new weight values, and 3. updating the weight vectors of neuron. The initialization involves initialization of weight vectors of output neurons and initialization of accumulators. In order to assist faster training time, K number of samples from the input are randomly selected and their weight vectors are used to initialize the neurons. To ensure the algorithm accounts for the latent structures of spatial data, the Spatial SOM takes a similar approach as it is done in Kangas Map [8] and Geo-SOM [7] with the introduction of a geographical tolerance parameter to limit the Best Matching Unit (BMU) search.

Figure 1: Spatial SOM analysis framework.



3 Case Study: Air Pollution Data

In this case study, sensor data of Particulate Matter (PM) are used to test the performance of a Spatial SOM on clustering high-dimensional data with spatial dependence. PM is a mixture of solid and liquid that constitutes air pollution. It is composed of acids (such as nitrates and sulphates), organic chemicals, metals, soil, and other dust particles and allergens from the atmosphere. PM of size smaller than 2.5 micrometres in diameter or smaller are referred to as fine particle pollution or PM2.5. And our focus is PM2.5 data set for this study.

A lot of research on PM2.5 in both United States and abroad demonstrate that there is a strong association between PM2.5 and human health, especially mortality [11]. However, there is no agreement among research on the magnitude of the effect of PM2.5 on human health [12]. This alludes that there are specific aspects of PM2.5 that influences varying results. Among many aspects, it has been suggested in some studies [13, 14] that differing geographical location of sites where PM2.5 were measured may be attributable to varying results of the PM2.5 studies. Recent research [13, 15, 16] show that the varying results of the PM2.5 studies may be due to spatial factor causing chemical composition of PM2.5 to vary, consequently, producing mixed results for different study sites. For this reason, PM2.5 chemical composition data set make an interesting case to study and assess for any hidden spatial relationships in the data set.

3.1 Data Preprocessing

As a first step of the analysis, PM2.5 data set is obtained from the United States Environmental Protection Agency Air Data website (http://www.epa.gov/airdata) and is pre-processed. They are diurnal samples from the period January 2003 until December 2008 selected based on the basis where the data is most complete. Samples are collected in the intervals varying from 3-6 days. Table 1 provides further meta-data of the data set used in the study.

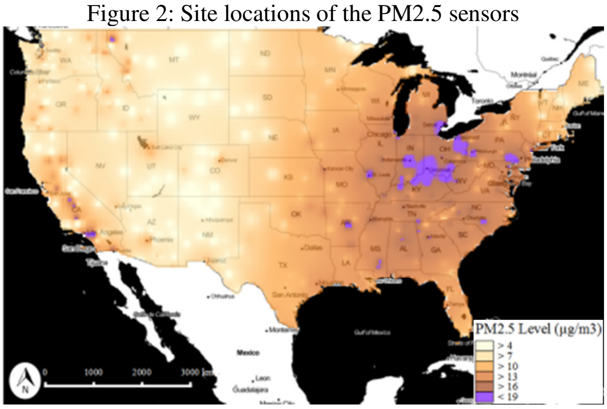
Table 1: Metadata of air pollution data used in the study.

Dataset	Continental United States
Bounding Box (WGS84)	-125.0011, 24.9493, -66.9326, 49.5904
Time Span	1 January 2003 - 31 December 2008
Covered Area	8,080,464.3 km
Number of Objects	221,000

To better understand the retrieved data set, geo-spatial assessment is done for the year 2005, to find out if there is any visible spatial correlation in the data set. Figure 2 illustrates the average PM2.5 concentration levels across the continental United States in 2005. The concentration levels show that PM2.5 level is especially high (highlighted purple) in the Midwest region near the Great Lakes and in Fresno region, California in the West.

To be able to get a comprehensive analysis of such high-dimensional data set with spatial dependence, a multivariate SOM approach is necessary. A better grasp of spatial relationships will enhance our understanding of the interaction between pollutants as well as further human health effects related to exposure to these complex mixtures.

Figure 2: Site locations of the PM2.5 sensors



Source: (base map: Stamen Design CC BY 3.0, Data by OpenStreetMap CC BY SA).

3.2 SOM Result

Figure 3 is a resulting U-matrix of PM2.5 data trained with Spatial SOM. In the U-matrix, several clusters can be detected. The difference in distances are represented by different shades of grey. Neurons shaded in black to dark grey (range = 0.0-0.5) indicate neurons that are close to each other in the input space. Neurons shaded in white to light grey (range = 0.5 - 1.0) show neurons that are far from each other in the input space. Darker regions in the U-Matrix are units with low U_{height} value and therefore they are clusters, and lighter regions are units with high U_{height} value and thus can be interpreted as cluster separators. In general, there are many large loose clusters and some small clusters. This suggests that there are many values that are similar in the data set and relatively fewer values that are more extreme than the others. From the U-Matrix, four clusters are identified for a further analysis. The analysis focused on these four clusters because they had the lowest U_{height} values (< 0.2). These clusters are highlighted in shades of red, blue, yellow, and green in Figure 3.

It can be concluded that the spatial SOM clusters varying PM2.5 chemical species levels over geographical space fairly well. In general, regions with higher levels of total PM2.5 chemical concentration are clustered together (C3 and C4), and lower levels are clustered together as well (C1 and C2). C2 Midwest cluster, more specifically Great Lakes region, is one of the interesting cluster that can be observed in the

Spatial SOM output. The sites included in this group are primarily urban.

Figure 3: U-matrix representation of the SOM analysis on PM2.5 data.

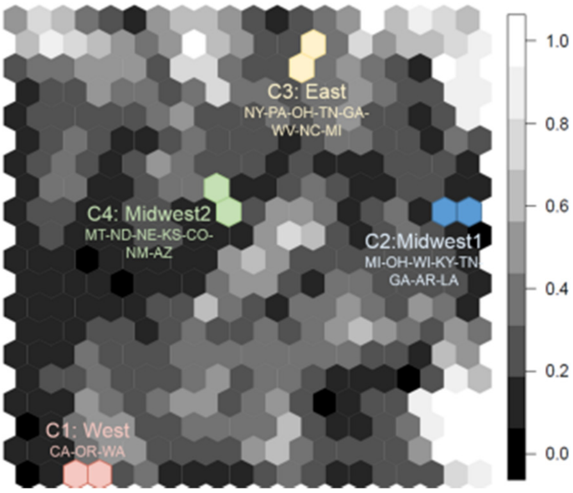
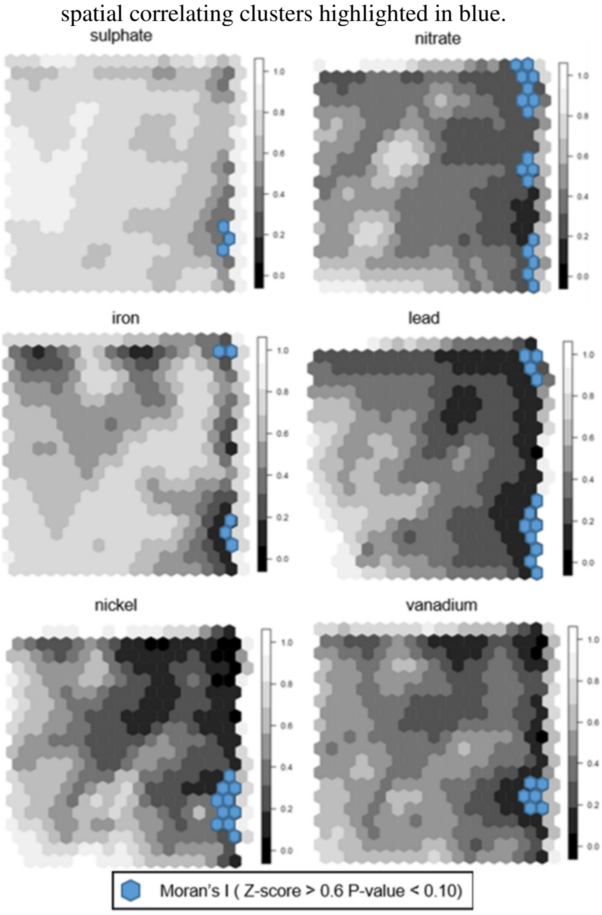


Figure 4: Component planes for selected variables with spatial correlating clusters highlighted in blue.



Component planes (Figure 4) show that there is a relatively stronger tendency that metal elements, such as iron and lead, are clustered close to the right side, which coincides with the C2 cluster's location in the U-Matrix (Figure 3). Thus, it can be implied that these metal variables influence the C2 cluster the most. This phenomenon can be attributed to the fact the Great Lakes region is known as the agglomerate of heavy industry such as iron and automobile industries historically and today. The East cluster (C1) can also be extracted from vanadium and nickel planes. Vanadium and nickel are known chemical pollution that come from ship engine exhaust [17]. The sites in C1 are primarily close to the eastern coasts or major inland body of water, they are New York, North Carolina, Pennsylvania, North Carolina, and Michigan. Thus, the presence of large volume of ship traffics near the coasts and inland waters may explain the patterns of higher vanadium and nickel concentration in these areas. Although it is less clear in comparison to other component planes, sulphate and nitrate component planes also have tendency to cluster on the far right hand side of the plane.

4 Conclusion

In this study, we set out the objective to contribute and provide insights in overcoming the challenges with spatial big data. Specifically, we aimed to study ways in which we can apply and assess spatial SOM algorithms for handling large datasets, and how we can uncover latent structures in big data with spatial dependence.

The detailed analysis of the Spatial SOM result demonstrates that Spatial SOM is an effective tool in detecting cluster with spatial dependence in the data. It is able to detect chemical species variation across continental United States. In conclusion, the case study validates the effectiveness of Spatial SOM as an analysis tool for discovering not only hidden relationships in general attribute features but also for spatial features as well.

Regarding our research question, we demonstrated that Spatial SOMs can detect underlying latent spatial and chemical structures and covariates from pollution data collected from sensors using a case study with air pollution data. In other SOM variants, usually spatial attributes are treated as any other attributes given equal weights. However, in a Spatial SOM, more weight is given to the spatial attribute, so that input vectors or observations are largely aggregated based on their geographical proximity first, and then fine tuned by other chemical features. The Spatial SOM was able to reduce high-dimensional data into two-dimensional visual representation as shown in the U-matrix (Figure 3).

References

[1] T. Kohonen. The self organizing map. In *Proceedings of the IEEE* 78, 9: 1464–1480, 1990.

[2] L. Anselin. Local indicators of spatial association - LISA. *Geographical Analysis*, 27(2):93–115, 2010.

[3] W. Tobler. A computer model simulating urban growth in the Detroit region. *Economic Geography*, 46:234–240, 1970.

[4] H. J. Miller and J. Han. Geographic data mining and knowledge discovery: An overview. In H. J. Miller and J.

Han, editors, *Geographic Data Mining and Knowledge Discovery*, pages 1–48. CRC Press, 2009.

[5] S. Openshaw and C. Wymer. Classifying and regionalizing census data. In S. Openshaw, editor, *Census Users Handbook*, pages 239–268. Cambridge, UK, 1995.

[6] A. Ultsch and C. Vetter. Self-organizing-feature-maps versus statistical clustering methods: a benchmark. *Research Report No 90194*, 1994.

[7] F. Bação, V. Lobo, and M. Painho. The self-organizing map, the Geo-SOM, and relevant variants for geosciences. *Computational Geosciences*, 31:155–163, 2005.

[8] J. Kangas. Temporal knowledge in locations of activations in self-organizing map. In I. Aleksander and J.Taylor, editors, *Artificial Neural Networks*, 2, pages 117–120. North-Holland, Amsterdam, Netherlands, 1992.

[9] T. Kohonen. The hypermap architecture. In T. Kohonen, K. Mäkisara, O. Simula,and J. Kangas, editors, *Artificial Neural Networks*, volume 1, pages 1357–1360. The Nertherlands, 1991.

[10] E. Steiger, B. Resch, A. Zipf. Exploration of spatiotemporal and semantic clusters of Twitter data using unsupervised neural networks. *International Journal of Geographical Information Science*, volume and issue pending, 2015.

[11] J. Schwartz, D. W. Dockery, and L. M. Neas. Is daily mortality associated with fine particles? *Journal of the Air and Waste Management Association*, 46:927–939, 1996.

[12] A. Zanobetti, J. Schwartz, E. Samoli, and A. Gryparis et al. The temporal pattern of mortality response to air pollution: A multicity assessment of mortality displacement. *Epidemiology*, 13, 2002.

[13] M. L. Bell, F. Dominici, K. Ebisu, S. Zeger, and J. Samet. Spatial and temporal variation in pm2.5 chemical composition in the United States for health effects studies. *Environ Health Perspect*, 115.

[14] E. Austin, B. A. Coull, A. Zanobetti, and P. Koutrakis. A framework to spatially cluster air pollution monitoring sites in us based on the pm2.5 composition. *Environment International*, 59:244–254, 2013.

[15] G. Lin, J. Fu, D. Jiang, W.Hu, D. Dong, Y. Huang and M. Zhao. Spatio-temporal variation of pm2.5 concentrations and their relationship with geographic and socioeconomic factors in china. *International Journal of Environmental Research and Public Health*, 11(1):173–186, 2014.

[16] J. P. Pinto, A. S. Lefohn, and D. S. Shadwick. Spatial variability of pm2.5 in urban areas in the United States. *Journal of Air and Waste Management Association*, 54(4):440–449, 2012.

[17] D. Mueller, S. Uibel, M. Takemura, D. Klingelhoefer, and D. A. Groneberg. Ships, ports, and particulate air pollution - an analysis of recent studies. *Journal of Occupational Toxicology*, 6(31), 2011.