Application of SOM to Analysis of Minnesota Soil Survey Data

Sauptik Dhar and Vladimir Cherkassky, Fellow, IEEE.

Abstract— **This paper describes data-analytic modeling of the Minnesota soil chemical data produced by the 2001 metro soil survey. The chemical composition of the soil is characterized by the concentration of many metal and non-metal constituents, resulting in high-dimensional data. This high dimensionality and possible unknown (nonlinear) correlations in the data make it difficult to analyze and interpret using standard statistical techniques. This paper applies Self Organizing Map (SOM), to present the high-dimensional soil data in a 2D format suitable for human understanding and interpretation. This SOM representation enables analysis of the soil chemical concentration trends within the Twin Cities Metropolitan area of Minnesota. These trends are important for various Minnesota regulatory agencies concerned with the concentration of polluting chemical elements due to human activities.**

Index Terms— **Self-organizing maps (SOM), pollution, soil chemical survey data, geological surveying, cluster analysis.**

I. INTRODUCTION

The Minnesota Department of Transport (Mn/DOT) may expand the use of recycled materials as roadway bed or fill material. Reflecting dual goals of environmental stewardship and regulatory compliance, Mn/DOT recognized a need to analyze the chemistry of the surface soils in order to better understand the hazards associated with the use of recycled materials for sound barrier walls. Specifically there is a need to regulate the concentration of the hazardous elements like Arsenic (As), Chromium (Cr), Copper (Cu), Lead (Pb), Nickel (Ni), Tungsten (W), Zinc (Zn) and hence understand the trend in which the concentration of these elements varies throughout the Twin Cities Metropolitan area. For instance, a difference in the soil chemical concentration of these elements in the downtown, suburbs and the rural lands can be attributed to the different land usage in these areas. Apart from that, it is also important to understand how the pollution from the Mn/DOT roads affects the concentration of these elements in the soil samples nearby. However, such an analysis of the soil chemical data is not straightforward. The dimensionality and possible unknown (nonlinear) correlations in the data make it difficult to analyze via standard statistical techniques like Principal Component Analysis (PCA) or Factor Analysis (FA), widely used in soil chemical data analysis (see [1] for a brief survey). Other major issues with these methods are the non-normality of the variables and outliers [2, 3] and the compositional nature of the soil chemical data, which leads to false correlations [4-6] and may even result in unstable or erroneous conclusions [2]. On the other hand, Self Organizing Map (SOM) has shown promising results under such scenarios [1] and has recently gained popularity for soil chemical data interpretation [1], [7-11].

In this paper we apply SOM, to present the high-dimensional soil data characterized by the chemical concentration of the elements As, Cr, Cu, Pb, Ni, W, and Zn, in a 2D format suitable for human understanding and interpretation. This SOM representation will help us to understand,

Goal 1. If there exists a trend in the soil chemical concentration of these elements within the Twin Cities Metropolitan area, based on the region (downtown, suburbs and the rural lands) from which the soil samples were collected. Note that, any observable trend can be attributed to the enrichment of these elements due to different land usage in the downtown, suburbs and the rural lands.

Goal 2. If there exists a trend in the soil chemical concentration of these elements within the Twin Cities Metropolitan area, based on the proximity to the Mn/DOT roads from which the soil samples were collected. Any observable trend can be attributed to the enrichment of these elements due to pollution caused by the Mn/DOT roads.

The paper is organized as follows,

- Section II provides an overview of the data collection and the experimental procedure adopted to answer the Goals 1 and 2.
- Section III describes the SOM modeling results and provides a brief analysis of the results.
- Finally the summary is presented in Section IV.

II. DATA COLLECTION AND EXPERIMENTAL PROCEDURE

The dataset used for this study is the Metro 2001 Soil Survey Data [11]. This data represents the surface soil samples collected from a depth of 0-5 cm at different sites along the Mn/DOT roads within the Minneapolis-St. Paul metropolitan area. Each of the soil samples measures the concentration of the elements As, Cr, Cu, Pb, Ni, W, and Zn in parts per million (ppm). These elements have been

Manuscript received January 17, 2011. This work was supported by Minnesota Department of Transport (Mn/DOT).

Sauptik Dhar is with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis MN 55455 USA. (e-mail: dharx007@umn.edu).

Vladimir Cherkassky is with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis MN 55455 USA. (phone: 612 625-9597; fax: 612 625-4583; e-mail: cherk001@umn.edu).

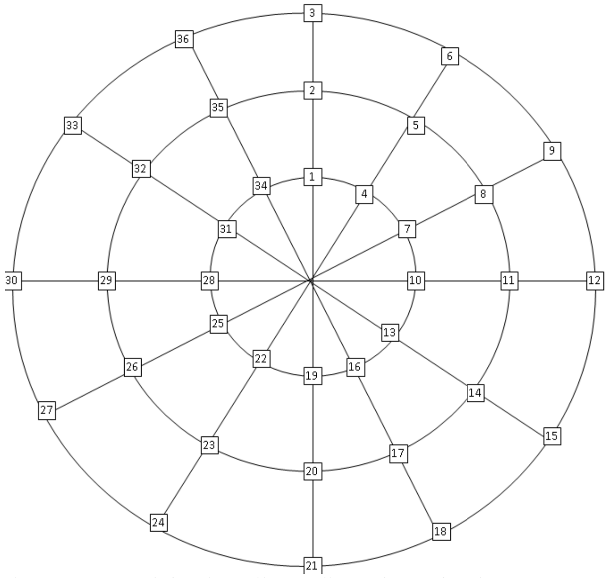


Fig. 1. Map used for the soil sampling scheme for the Metro 2001 Soil Survey Data.

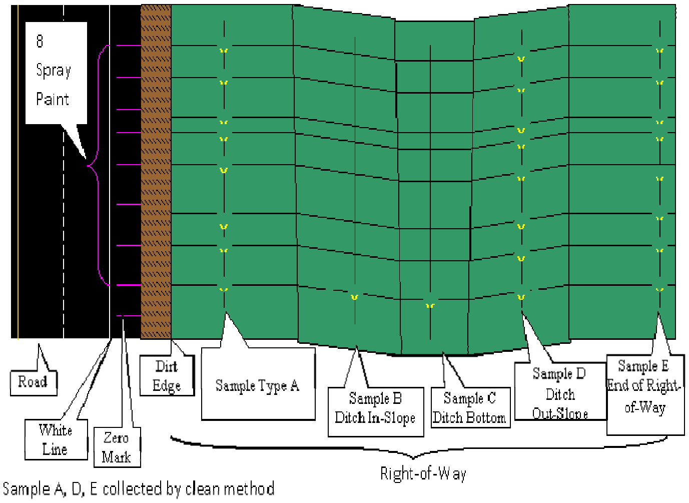


Fig. 2. Map explaining different soil sample types for the Metro 2001 Soil Survey Data.

identified by the Mn/DOT Technical Advisory Panel as the elements of regulatory interests whose presence beyond certain levels can endanger human health. In this section we provide an overview of the soil sampling scheme, the data preprocessing and the procedure for applying SOM to the preprocessed data.

A. Soil Sampling Scheme

The soil sampling scheme for this data set is reproduced below from [12]. This scheme will help us to answer the questions specified in the Goals 1 and 2. The soil sampling scheme uses three concentric circles encompassing the Minneapolis-St. Paul metropolitan area (as shown in Fig. 1). The center corresponds to Minneapolis Post Office, and each circle generally covers the following areas:

- Circle I (inner circle): Downtown
- Circle II (middle circle): First and second ring suburbs.
- Circle III (outer circle): The edge of the second ring suburbs and the rural lands.

The circles are divided by 12 line segments (each 30 degrees apart). And the points at which the lines intersect the circles are identified as sites (numbered from 1-36 shown inside the squares in Fig. 1). For each site three soil samples are collected at different distances from the road. These samples are named as Type-A, Type-D and Type-E, as shown in Fig. 2, where

- Type-A: Represents the soil samples that are nearest to the road.
- Type-D: Represents the soil samples that are relatively further away from the road. It generally indicates ditch out-slope.
- Type-E: Represents the data samples that are farthest from the road. These are basically the samples collected from the end of Right-of-Way.

Apart from this we also have 8 background soils samples for each of the Circle I, II and III. These soil samples were not along Mn/DOT roads and have been collected near a parking lot or minor city road. A brief description of the background samples is provided in Table I. In total we have 130 soil samples of which 108 soil samples (of Types A, D and E) are collected along the Mn/DOT roads and the remaining 24 are background samples.

TABLE I
BACKGROUND SAMPLES SITE DESCRIPTION

Circle	Mn/DOT sample ID	Description of Sites
I	37	Evergreen Park Roseville (Baseball Field)
I	38	Roseville Park (Playground lawn)
II	39	Woodland lawn perhaps in a park
II	40	Bare spot in wooded area
III	41	Forest Lake truck station near woods.
III	42	Carlos Avery Wildlife Mgmnt Area North Metro Office near woods
I	43	Minnehaha Falls wooded lawn near nature trail
I	44	Metro lake wooded lawn
II	45	Wooded lawn with fence
II	46	Wooded lawn residential
III	47	Unknown state park
III	48	Grassy lawn near a farm
I	49	W 44th St. and Lake Harriet Parkway grassy roadside
I	50	City roadside possibly lake area
II	51	Bryant Lake regional park
II	52	Bush Lake Park grassy area
III	53	Wooded lawn residential
III	54	Cleary Lake Regional Park grassy roadside
I	55	Bassett Creek Park wooded lawn
I	56	North Bass Lake Park grassy lawn near building
II	57	Clifton French Regional Park along grassy woodland
II	58	Grassy woodland dirt road
III	59	Elm Creek Park Reserve Horse Camp grassy field
III	60	Brushy grass along dirt road

Further, we use a special naming scheme for the different soil samples. This scheme is designed to capture the information implicit to the soil sampling scheme and will help us to better answer the questions put forward in the Goals 1 and 2. In this naming scheme, we design a tag to identify the different soil samples based on the Circles, Site ID and Sample Type. Here, each soil sample is identified by its: *Circle-Site ID-Sample Type*. For example, a sample with a tag *I-4-A* would belong to,

- Circle = Circle I (Downtown).
- Site ID (for the sample) = 4.
- Sample Type = Type A. (The background samples do not have any soil sample type and hence the sample type will be replaced by '*Back'* indicating a background sample).

Note that any trend in the concentration values of the elements of regulatory interest (As, Cr, Cu, Pb, Ni, W and Zn) for the samples belonging to the same *Circle* or *Sample Type* could answer the questions put forward by the Goals 1 and 2 respectively.

B. *Data Preprocessing*

The data contains a number of missing values. A total of 9 (III-9-A, III-9-D, III-9-E, III-12-A, III-12-D, III-12-E, I-10-D, I-13-D, I-25-E) out of the 130 soil samples have missing values for all the elements As, Cr, Cu, Pb, Ni, W and Zn. These missing data were due to budget cuts and hence no measurements were available for these samples. For this analysis we remove all these 9 samples, and use the remaining 121 soil samples for our analysis. Table II provides the pearson correlation coefficient for this dataset using these 121 soil samples.

TABLE II
PEARSON CORRELATION COEFFICIENT FOR DIFFERENT ELEMENTS

	As	Cr	Pb	Ni	W	Zn	Cu
As	1.00	-0.18	-0.05	-0.04	0.07	-0.06	-0.06
Cr	-0.18	1.00	-0.01	0.38	-0.08	-0.03	0.08
Pb	-0.05	-0.01	1.00	0.22	0.52	0.36	0.25
Ni	-0.04	0.38	0.22	1.00	0.21	0.12	0.27
W	0.07	-0.08	0.52	0.21	1.00	0.23	0.26
Zn	-0.06	-0.03	0.36	0.12	0.23	1.00	0.27
Cu	-0.06	0.08	0.25	0.27	0.26	0.27	1.00

As seen from this table the data is not linearly correlated. However, the possibility of some unknown nonlinear correlation in the dataset motivates us to resort to a nonlinear dimensionality reduction technique like Self Organizing Map (SOM) which has shown promising results under such scenarios [1].

For SOM modeling it is important to uniformly scale the data to a range of [0-1] because the ranges of concentration values for different elements are quite different. This scaling is performed independently for each element according to the standard preprocessing prior to SOM modeling [13].

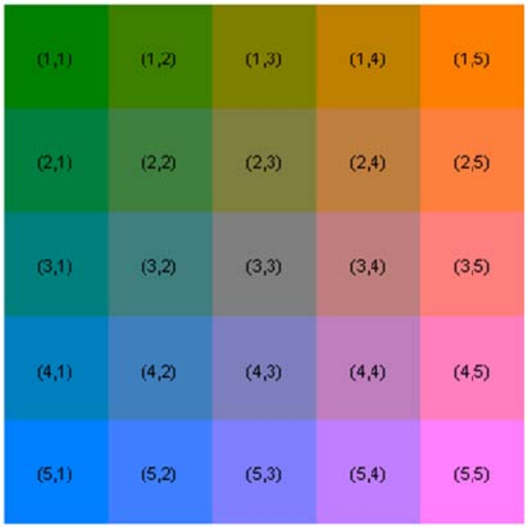


Fig. 3. 2D SOM topological map with 25(5x5) units. The units are numbered as in the coordinate system. The units are colored so that similar color shades represent neighboring units.

C. *SOM Experimental Setup*

In this paper we use the batch version of the SOM algorithm proposed by Kohonen [13]. This algorithm is provided in the SOM package publicly available at [14]. We use a 2-D SOM topological map as shown in Fig. 3 where each unit is represented by a colored box. The parameters for the SOM map used for the analysis is specified below,

- Initial Neighborhood=1, (This parameter specifies the initial neighborhood width to be used during training the SOM map. This package uses a Gaussian neighborhood. This parameter describes the standard deviation of the Gaussian neighborhood, where a value of 1 means a standard deviation roughly equal to the width of the map)
- Final Neighborhood=0.05, (The Gaussian neighborhood width to be used at the end of training)
- Number of Units=25 (This is the number of units to be used for the SOM map. We use a 5x5 rectangular structure).
- Total Iteration=75. (This is the total number of passes through the training set)

These parameters were manually chosen to produce stable clusters.

III. SOM MODELING RESULTS

The SOM modeling result for this data is provided in Fig. 4. For the ease of representation we print the identifiers/tag of the soil samples that belong to a particular SOM unit on the box representing that unit. Hence, the samples that fall on the same box (or the neighboring boxes) are likely to have similar (As, Cr, Cu, Pb, Ni, W and Zn) concentration values in comparison to the samples that lie in the distant

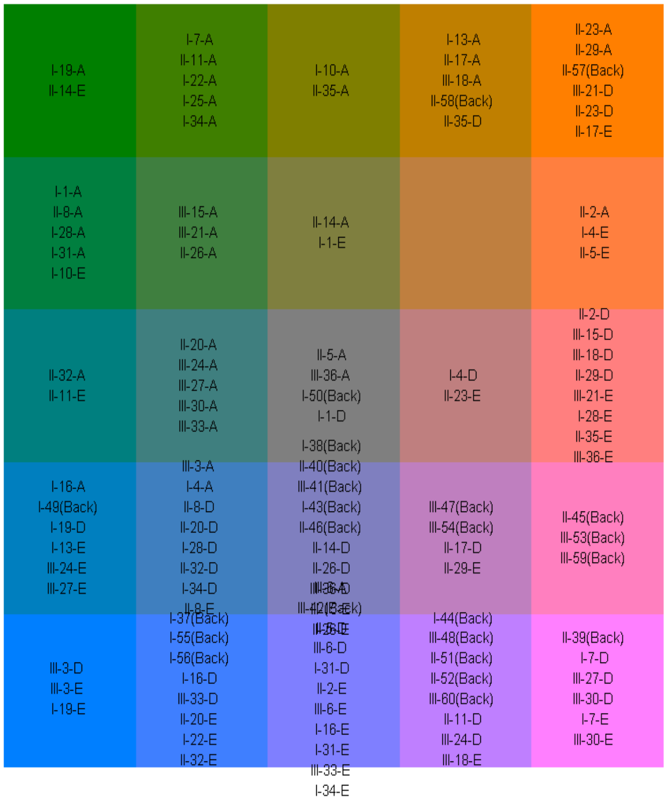
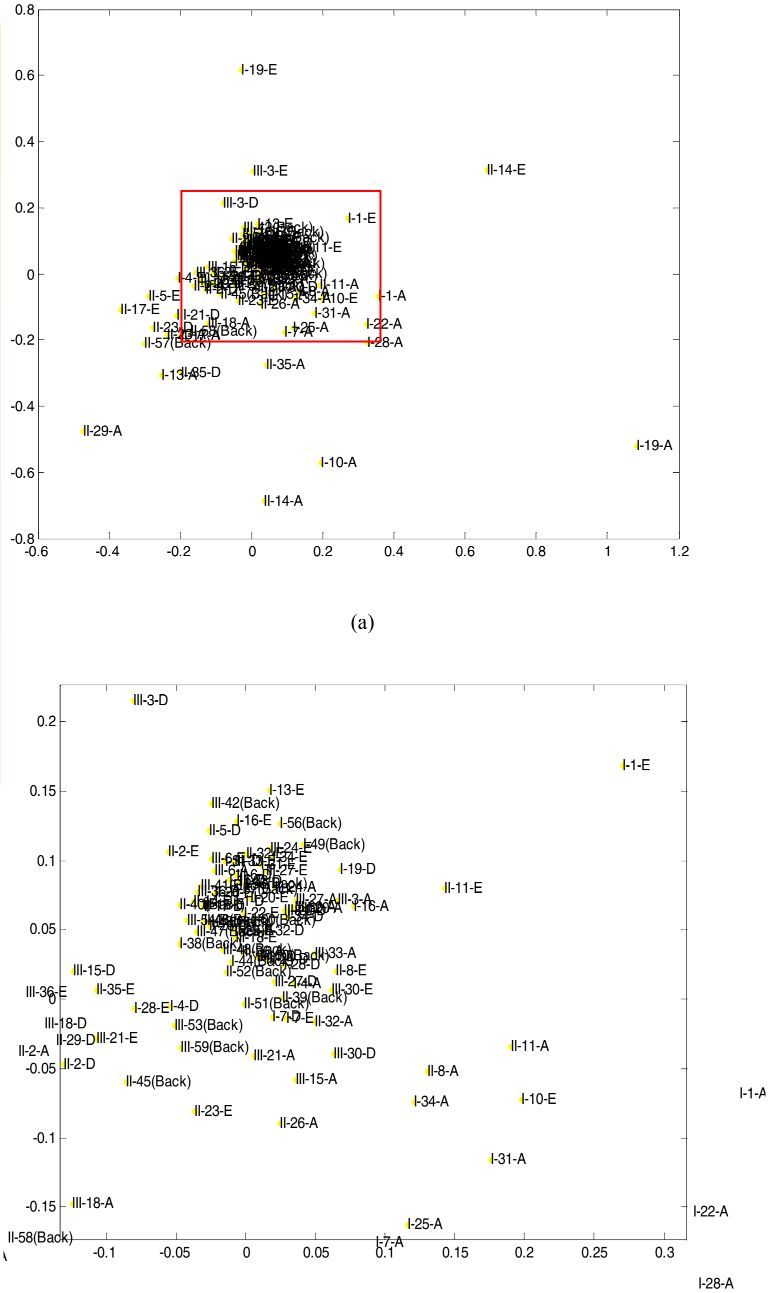


Fig.4. SOM model for Metro Soil Survey Data based on the chemical concentration of As, Cr, Cu, Pb, Ni, W, Zn.

boxes (units). Fig. 4 shows that there is no pattern of clustering based on the *Circle*. However, we do observe some pattern of clustering based on the *Sample Type*. We observe that the Type D & E soil samples and the background samples are mostly clustered at the bottom of the SOM map. On the other hand, the soil samples of Type A are clustered differently mostly at the top portion of the SOM map. Additionally, we have also applied the classical Multi Dimensional Scaling (MDS) [13] to this data and obtained the results shown in Fig. 5. Fig. 5 shows that the soil samples of Type A are mostly clustered on the bottom right and the soil samples of Type D & E and the background samples are mostly clustered on the left and the upper portions of the MDS representation. Such a pattern of clustering indicates that the soil samples of Type A which are closest to the Mn/DOT roads have different concentration values for the elements of regulatory interest than the soil samples of the other types. This provides a better understanding of the Goal 2, where we observe that there could be a possible enrichment of the concentration of the elements As, Cr, Cu, Pb, Ni, W and Zn in the soil data, due to pollution from the Mn/DOT roads.

For a deeper analysis of the Goal 2 we further perform the SOM modeling taking each element separately. This will provide a better understanding of the pattern of clustering based on the concentration of a particular element. Next we discuss the SOM modeling results for each element taken separately,



(a)

(b)

Fig. 5. (a) Output produced by classical MDS representing the pairwise distance between the soil samples. (b) A zoomed in version of the classical MDS representation about the red-box showing a clustering of Type-A samples (on the bottom right) vs. Type D & E and Background samples (mostly on the left and the upper portions of the representation).

- Arsenic (As): The SOM modeling result based on the element As is shown in Fig. 6. As seen from the figure we do not observe any pattern of clustering based on *Circle* or *Sample Type*.

- Chromium (Cr): The SOM modeling result based on the element Cr is shown in Fig. 7. As seen from the figure we do not observe any pattern of clustering based on *Circle* or *Sample Type*.

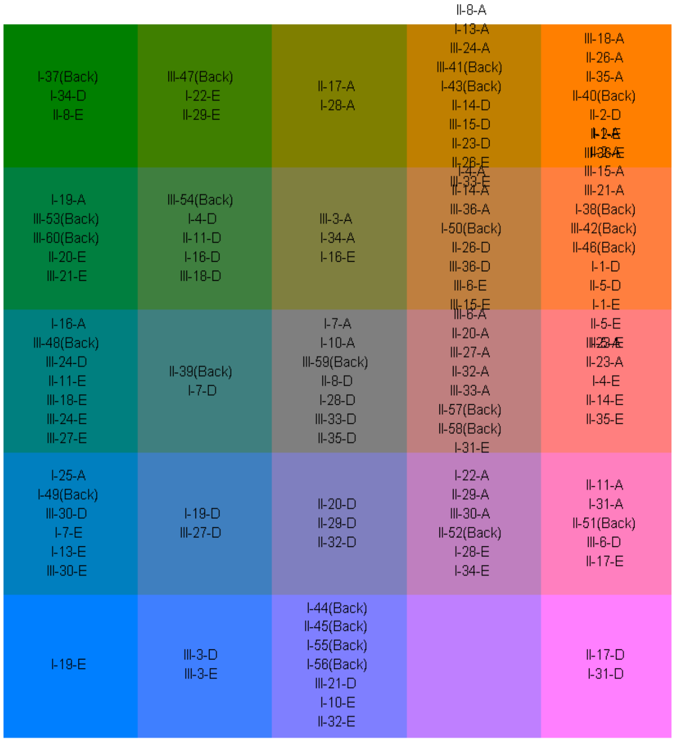


Fig. 6.SOM model for Metro Soil Survey Data based on the chemical concentration of As.

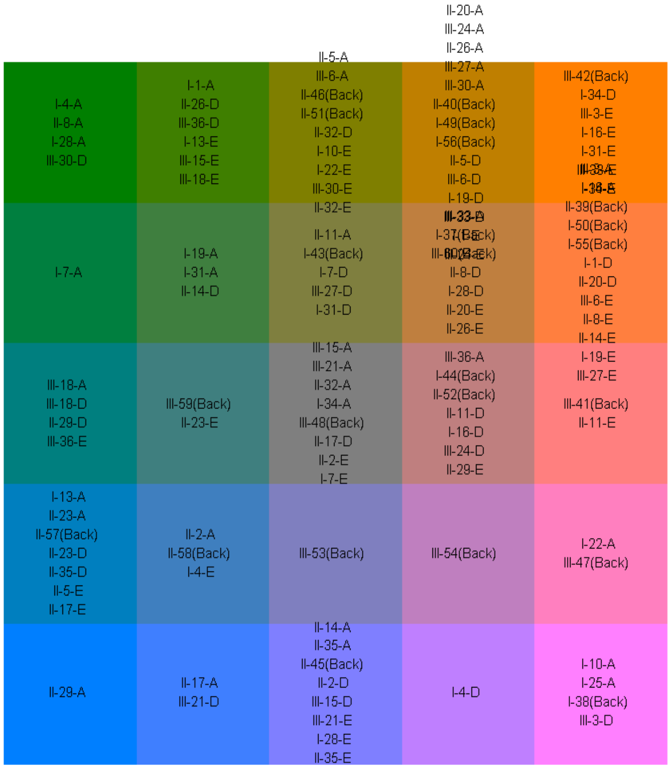


Fig.7. SOM model for Metro Soil Survey Data based on the chemical concentration of Cr.

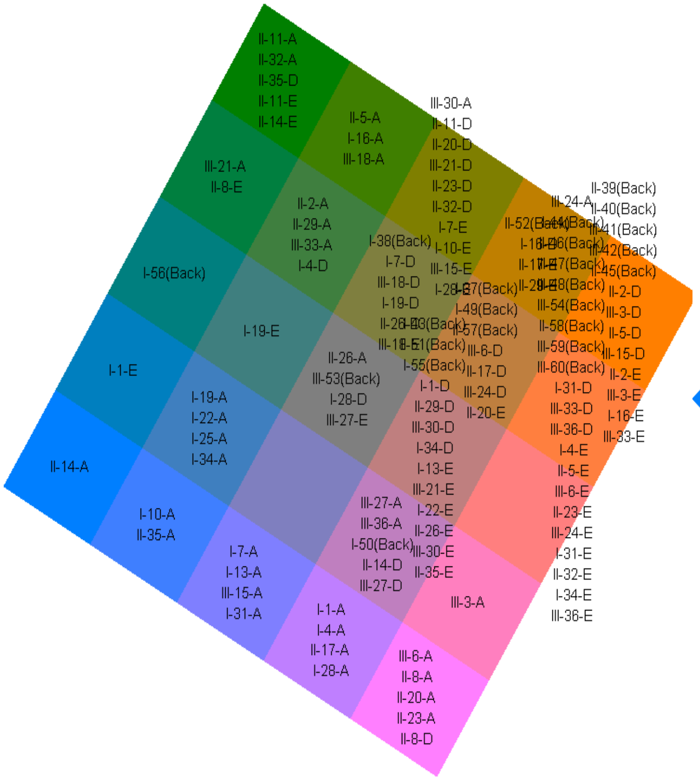


Fig. 8.SOM model for Metro Soil Survey Data based on the chemical concentration of Cu.

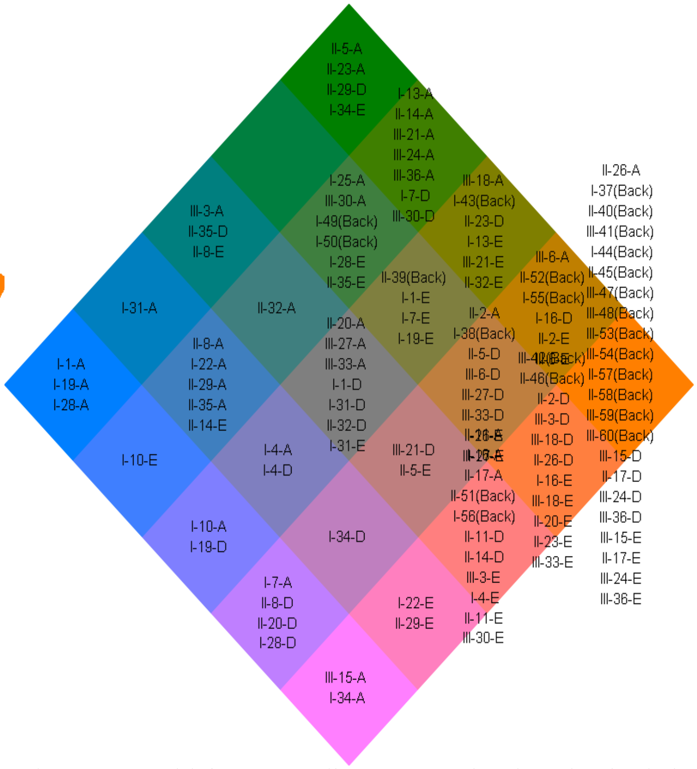


Fig. 9.SOM model for Metro Soil Survey Data based on the chemical concentration of Pb.

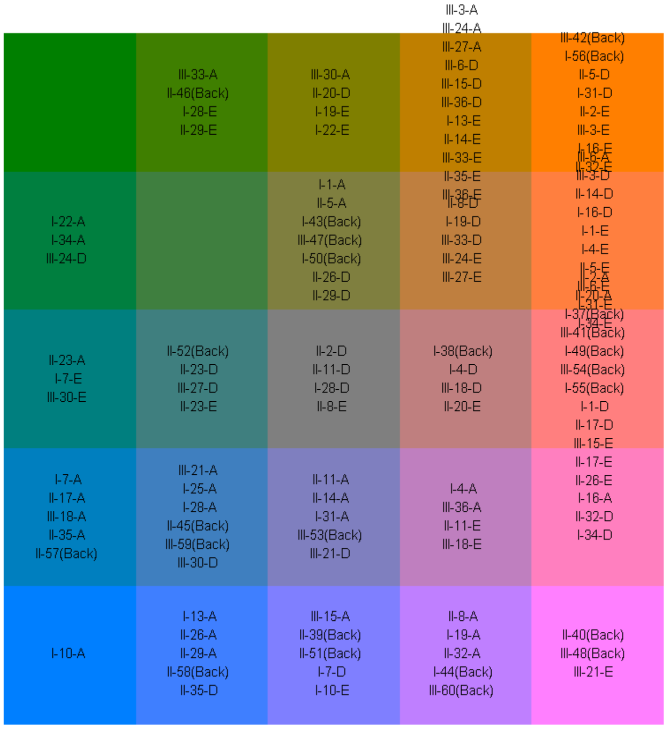


Fig. 10.SOM model for Metro Soil Survey Data based on the chemical concentration of Ni.

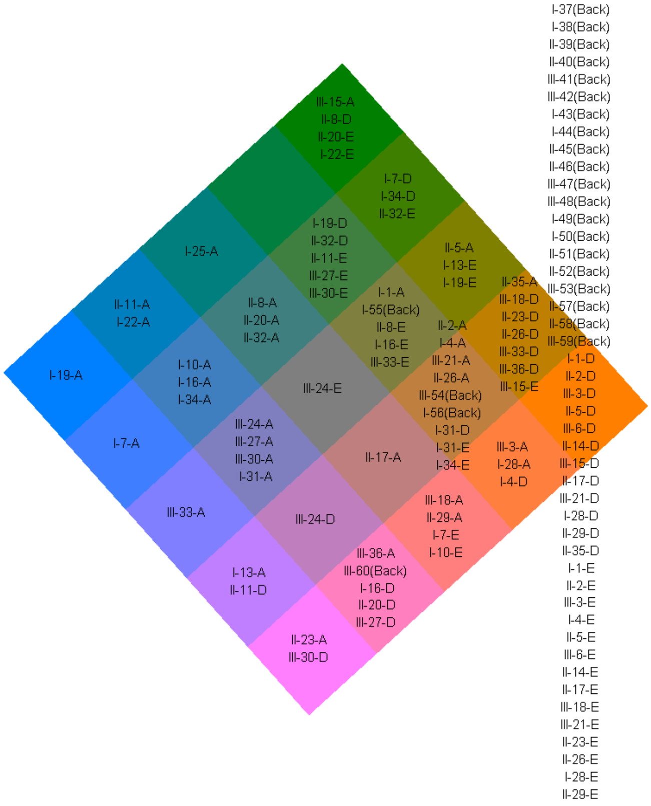


Fig.11. SOM model for Metro Soil Survey Data based on the chemical concentration of W.

– Copper (Cu): The SOM modeling result based on the element Cu is shown in Fig. 8. For better view ability we have rotated the SOM map. As seen from the figure we do not observe any pattern of clustering based on *Circle*. However we do observe a pattern of clustering based on the *Sample Type*. We observe that the samples of Type D & E and the background samples are clustered mostly on the right corner of the map; whereas the samples of Type A are clustered differently. In fact, we have checked that the samples of Type A have higher Cu concentration values than the samples of other types.

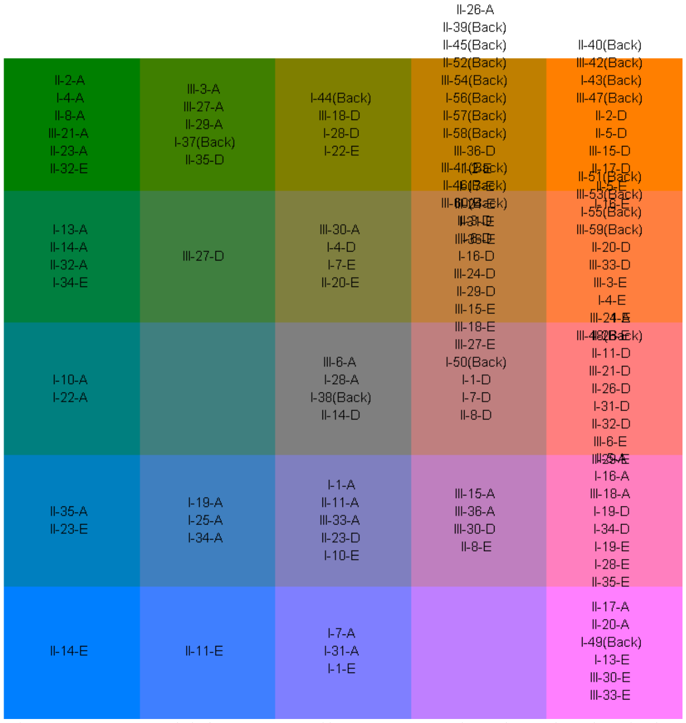


Fig. 12. SOM model for Metro Soil Survey Data based on the chemical concentration of Zn

– Lead (Pb): The SOM modeling result based on the element Pb is shown in Fig. 9. As seen from the figure we do not observe any pattern of clustering based on *Circle*. However, similar to Cu we do observe that the samples of Type D & E and the background samples are clustered mostly on the right corner of the map; whereas the samples of Type A are clustered differently. In fact, we have checked that the samples of Type A have higher Pb concentration values than the samples of other types.

– Nickel (Ni): The SOM modeling result based on the element Ni is shown in Fig. 10. As seen from the figure we do not observe any pattern of clustering based on *Circle* or *Sample Type*.

– Tungsten (W): The SOM modeling result based on the element W is shown in Fig. 11. For better view ability we have rotated the SOM map. As seen from the figure we do not observe any pattern of clustering based on *Circle*. However, as seen for some of the other elements we observe a pattern of clustering based on the *Sample Type*. We observe that the samples of Type D & E and the background samples are strongly clustered on the right corner of the map; whereas the samples of Type A are clustered differently more on the left side of the map. In fact, we have checked that the samples of Type A have relatively higher W concentrations than the Type D &

E samples and the background samples.

- Zinc (Zn): The SOM modeling result based on the element Zn is shown in Fig. 12. As seen from the figure we do not observe any pattern of clustering based on *Circle*. However, we do observe a pattern of clustering based on the *Sample Type.* We observe that the samples of Type D & E and the background samples are clustered mostly on the right side of the map; whereas the samples of Type A are clustered differently. In fact, we have checked that except for the samples II-14-E, II-11-E which are clustered at the bottom left of the map the samples of Type A have relatively higher Zn concentrations than the Type D & E samples and the background samples.

Thus, we observe that there is no distinct trend in the soil chemical concentration of the elements As, Cr, Cu, Pb, Ni, W and Zn based on the region (downtown, suburbs and rural lands) from which the soil samples were collected. However, there is a distinct trend based on the distance from the Mn/DOT roads. A detailed analysis of our results suggests enrichment in the concentration values of the elements Cu, Pb, W and Zn for the soil samples which are close to the Mn/DOT roads. This can be attributed to the pollution caused by the activities in the Mn/DOT roads. Thus, care should be taken while deciding upon the use of recycled materials for the sound barrier walls. Probably materials having lower concentration of Cu, Pb, W and Zn could be a better choice.

IV. Summary

This paper describes data-analytic modeling of the Minnesota soil chemical data within the Twin Cities Metropolitan area. The purpose of the analysis was to understand how the chemical concentration of the elements Arsenic(As), Chromium(Cr), Copper(Cu), Lead(Pb), Nickel(Ni), Tungsten(W), and Zinc(Zn) for the soil samples in the metro area change with the distance from major Mn/DOT roads and region (downtown, suburbs and rural lands) relative to the Minneapolis Main Post Office. SOM clustering based on the overall concentration of these elements of regulatory interest indicates that the Type A samples (collected close to the road) have similar characteristics, whereas the Type D & E samples (collected far away from the roads) have similar characteristics. Moreover, the background samples collected near parking lots or minor city roads have chemical characteristics similar to the Type D & E samples. This analysis indicates that the enrichment of the concentration of these elements of regulatory interest in the soil data is due to the proximity to the major Mn/DOT roads in the metro area. Further, a detailed analysis shows that these Type A soil samples have high concentration values for the elements Copper (Cu), Lead (Pb), Tungsten (W), and Zinc (Zn). These results could be helpful to determine the suitability of certain materials for usage as roadway bed or fill-in materials, at a particular location.

Acknowledgment

We thank Dr. Robert Edstrom from the Minnesota Department of Transport for the numerous discussions regarding the goals of data modeling and interpretation of results. We also gratefully acknowledge contributions from Mr. Steven Hennes from the Minnesota Pollution Control Agency for providing the list of elements of regulatory interest and for helping to interpret the pattern of clustering for these elements. Finally we thank Mr. Feng Cai who performed pre-processing of the soil survey dataset, including the analysis of missing data.

References

[1] G. Žibret, and R. Šajn, “Hunting for Geochemical Associations of Elements: Factor Analysis and Self-Organising Maps,” *Mathematical Geosciences*, vol. 42, pp. 681–703, June 2010.

[2] C. Reimann, P. Filzmoser, and R.G. Garrett “Factor analysis applied to regional geochemical data: problems and possibilities,” *Applied Geochemistry*, vol. 17, pp.185–206, March 2002.

[3] P. Filzmoser , and K. Hron, “Outlier Detection for Compositional Data Using Robust Methods,” *Mathematical Geosciences*, vol. 40, pp. 233–248, 2008.

[4] V. Pawlowsky-Glahn, and J.J Egozcue, “Geometric approach to statistical analysis on the simplex,” *Stochastic Environmental Research and Risk Assesment*, vol. 15, pp. 384-398, 2001.

[5] V. Pawlowsky-Glahn, and J. J. Egozcue, “Compositional data and their analysis: an introduction”, *Geological Society*, vol. 264, pp. 1-10, 2006.

[6] P. Filzmoser, K. Hron, C. Reimann, and R. Garrett, “Robust factor analysis for compositional data,” *Computers & Geosciences* ,vol. 35, pp. 1854-1861, September 2009.

[7] R. Arias, A. Barona, G. Ibarra-Berastegi, I. Aranguiz, and A. Elías, “Assessment of metal contamination in dredged sediments using fractionation and Self-Organizing Maps,” *Journal of Hazardous Materials*, vol. 151, pp. 78-85, February 2008.

[8] P.M. Mele, and D.E Crowley, “Application of self-organizing maps for assessing soil biological quality,” *Agriculture, Ecosystems & Environment*, vol. 126, pp. 139-152, July 2008.

[9] S. Tsakovski , B. Kudlak , V. Simeonov , L. Wolska, and J. Namiesnik “Ecotoxicity and chemical sediment data classification by the use of self-organising maps,” *Analytica Chimica Acta*, vol. 631, pp. 142-152, January 2009.

[10] A. Samecka-Cymerman, A. Stankiewicz, K. Kolon, and A. J. Kempers, “Self-organizing feature map (neural networks) as a tool to select the best indicator of road traffic pollution (soil, leaves or bark of Robinia pseudoacacia L.),” *Environmental Pollution*, vol. 157, pp. 2061-2065, July 2009.

[11] S. Dhar, V. Cherkassky, “Statistical Analysis of the Soil Chemical Survey Data”, Report no. Mn/DOT 2010-22, June 2010.

[12] HDR Engineering, Inc., “Soil Sampling Plan, Minneapolis: HDR Engineering, Inc.”, June 2003.

[13] V. Cherkassky, and F. Mulier, *Learning from Data Concepts: Theory and Methods*, 2nd ed. NY: Wiley, 2007.

[14] V. Cherkassky, “Self Organizing Map”, [Online], Available: http://www.ece.umn.edu/users/cherkass/ee8591/software/Software_and_Datasets/SOM.htm ,[Accessed: Feb 03,2011].