Exploratory data analysis and clustering of multivariate spatial hydrogeological data by means of GEO3DSOM, a variant of Kohonen's Self-Organizing Map

L. Peeters, F. Bação, V. Lobo, A. Dassargues

Hydrol. Earth Syst. Sci. Discuss., 3, 1487–1516, 2006
www.hydrol-earth-syst-sci-discuss.net/3/1487/2006/

Hydrology and Earth System Sciences Discussions

Papers published in *Hydrology and Earth System Sciences Discussions* are under open-access review for the journal *Hydrology and Earth System Sciences*

Exploratory data analysis and clustering of multivariate spatial hydrogeological data by means of GEO3DSOM, a variant of Kohonen's Self-Organizing Map

L. Peeters[1], F. Bação[2], V. Lobo[2,3], and A. Dassargues[1,4]

[1]Applied Geology and Mineralogy, KULeuven, Leuven, Belgium
[2]Instituto Superior de Estatística e Gestão de Informação, Universidade Nova de Lisboa, Campus de Campolide, Lisboa, Portugal
[3]Portuguese Naval Academy, Alfeite, Almada, Portugal
[4]Hydrogeology and Environmental Geology, University of Liege, Belgium

Received: 17 May 2006 – Accepted: 5 June 2006 – Published: 11 July 2006

Correspondence to: L. Peeters (luk.peeters@geo.kuleuven.be)

Abstract

The use of unsupervised artificial neural network techniques like the self-organizing map (SOM) algorithm has proven to be a useful tool in exploratory data analysis and clustering of multivariate data sets. In this study a variant of the SOM-algorithm is proposed, the GEO3DSOM, capable of explicitly incorporating three-dimensional spatial knowledge into the algorithm. The performance of the GEO3DSOM is compared to the performance of the standard SOM in analyzing an artificial data set and a hydrochemical data set. The hydrochemical data set consists of 141 groundwater samples collected in two detritic, phreatic, Cenozoic aquifers in Central Belgium. The standard SOM proves to be more adequate in representing the structure of the data set and to explore relationships between variables. The GEO3DSOM on the other hand performs better in creating spatially coherent groups based on the data.

1 Introduction

Regional monitoring of groundwater quality often yields large multidimensional data sets in which each sampling location is characterized by its geographic coordinates, longitude, latitude and height. Exploratory data analysis (EDA) and clustering can help in summarizing available data, extracting useful information and formulating hypothesis for further research.

Traditionally multivariate techniques like principal component analysis (PCA) and factor analysis (FA) are used in the process of exploratory data analysis and clustering (e.g. Güler et al., 2002; Lambrakis et al., 2004; Love et al., 2004). Both PCA and FA are based on linear combinations of the original variables in order to reduce the dimensionality of the data set (Davis, 1986).

Recently, artificial neural network techniques, such as Kohonen's Self-Organizing Map (SOM), have also been used in EDA. The Self-Organizing Map may be used to project multidimensional data onto a two dimensional grid in a topology preserving way,

HESSD

3, 1487–1516, 2006

capturing complex, non-linear relationships between variables (Kohonen, 1995).

Although this method is frequently used in financial, medical, chemical and biological research (an overview is presented in Kaski, 1997), there are only a few cases in which the SOM-algorithm is used for hydrogeological research.

Hong and Rosen (2001) applied the technique to diagnose the effect of storm water infiltration on groundwater quality variables and to capture the complex nonlinear relationships between groundwater quality variables. Sanchez-Martos et al. (2002) used SOM in the classification of a hydrochemical data set from a detritic aquifer in a semi-arid region, into distinct classes of different chemical composition. Lischeid (2003) applied the self-organizing map algorithm to an intensively monitored watershed to investigate spatial and temporal trends in water quality data.

A major item in the exploratory data analysis and clustering of geo-referenced data is to include the spatial ordering in the clustering or classification algorithm. Bação et al. (2005) discusses this topic in relation to self-organizing map analysis and presents the GEOSOM, a modified version of the SOM-algorithm, designed to explicitly incorporate spatial information. Application of the GEOSOM on two artificial data sets and a real-world demographic data set revealed the ability of the GEOSOM to create a clustering both based on the geographic density of the samples and the similarity of the samples.

The GEO-SOM as presented by Bação et al. (2005) is limited to two-dimensional geo-referenced data. In this study, the GEOSOM is extended to incorporate three-dimensional geo-referenced data, hence the name GEO3DSOM. A thorough discussion on the algorithms proposed is presented in the next section. Comparison between the standard SOM and the GEO3DSOM is carried out by applying both techniques to a theoretical data set and a hydrochemical data set from two phreatic, sandy aquifers in Central Belgium.

2 Methods

2.1 Standard SOM

Artificial Neural Networks (ANN) are computer algorithms, inspired by the functioning of the nervous system of the human brain, capable of learning from data and generalizing. This learning process can be described as supervised or unsupervised learning. In the supervised learning process, the ANN is shown several input-output patterns during training to enable the trained ANN to make generalizations based on the training data and to correctly produce output patterns based on new input (Jain et al., 1996). Neural networks are widely applied in hydrologic research (e.g. ASCE, 2000), especially in time-series prediction (e.g. Alvisi et al., 2006; Coppola et al., 2003)

The SOM-algorithm is based on unsupervised learning, which means that the desired output is not known a priori. The goal of the learning process is not to make predictions, but to classify data according to their similarity. In the neural network architecture Kohonen proposed (Kohonen, 1995), the classification is done by plotting the data in n-dimensions onto a, usually, two-dimensional grid of units in a topology-preserving manner. The former means that similar observations are plotted in each others neighborhood on the 2-D-grid. The network architecture and the learning algorithm are illustrated in Fig. 1.

The neural network consists of an input layer and a layer of neurons. The neurons or units are arranged on a rectangular or hexagonal grid and are fully interconnected. Each of the input vectors is also connected to each of the units. The learning algorithm applied to the network can be divided into six steps (Kohonen, 1995; Kaski, 1997):

1. An $m \times n$ matrix is created from the data set with m rows of samples and n columns of variables. The matrix thus consists of m input vectors of length n. The classification of the input vectors is based on a similarity measurement, for instance Euclidean distance. In order to avoid bias in classification due to differences in measuring unit or range of the variables, a normalization is carried out. This can

be done by setting mean equal to zero and variance equal to 1 or by rescaling the range of each variable in the [0, 1] interval.

2. Each unit is randomly assigned an initial weight or reference vector with a length equal to the length of the input vectors (n).

3. An input vector is shown to the network; the Euclidean distances between the considered input vector $\boldsymbol{X}$ and all of the reference vectors $\boldsymbol{M}_i$ are calculated according to:

$$\boldsymbol{X} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$$
$$M = (m_1, m_2, \ldots, m_n) \in \mathbb{R}^n$$
$$\|\boldsymbol{X} - M\| = \sqrt{\sum_{i=1}^{n} (x_i - m_i)^2} \quad (1)$$

4. The best matching unit $\boldsymbol{M}_c$, the unit with the greatest similarity with the considered input vector, is chosen according to:

$$\|\boldsymbol{X} - \boldsymbol{M}_c\| = \min_i \{\|\boldsymbol{X} - \boldsymbol{M}_i\|\} \quad (2)$$

This step is illustrated in Fig. 1b, where the Euclidean distance between the input vector (0;0.1;0.02) and the reference vectors is calculated. The best matching unit is the upper left unit (distance = 0.102).

5. The weights of the best matching unit and the unit within its neighborhood $N(t)$ are adapted so that the new reference vectors lie henceforth closer to the input vector. The factor $\alpha(t)$ controls the rate of change of the reference vectors and is called the learning rate.

$$\boldsymbol{M}_i(t+1) = \begin{cases} \boldsymbol{M}_i(t) + \alpha(t)[\boldsymbol{X}(t) - \boldsymbol{M}_i(t)] & \forall \in N(t) \\ \boldsymbol{M}_i(t) & \forall \notin N(t) \end{cases} \quad (3)$$

HESSD
3, 1487–1516, 2006

This is illustrated in Fig. 1c where the weights of the upper left unit and the units within the neighborhood $N(t)$ with radius r, indicated by the dashed line, are adapted. The rate of adaptation of the units is controlled by the neighborhood function h, which decreases from one at the winning unit to zero at units located farther away than radius r. The most common used functions are bell-shaped (Gaussian) or square (bubble).

6. Steps 3 until 5 are repeated until a predefined maximum number of iterations is reached. During these iterations both α and $N(t)$ decrease, forcing the network to converge.

After training, each of the input vectors is assigned to his best matching unit and the grids can be visualized. There are two types of grids commonly used to visualize and analyze the result of the SOM procedure: component planes and U-matrix (Vesanto et al., 1999). The U-matrix or distance matrix shows the Euclidean distance between neighboring units by means of a grey scale. Typically darker colors represent great distances and lighter shades represent small distances. In this visualization method clusters are represented by a light area with darker borders, meaning that the reference vectors in a cluster and the input vectors assigned to them are more similar to each other than to reference vectors outside the cluster. Additionally the labels of the input vectors can be plotted onto the U-matrix to identify the input vectors forming a cluster.

The component planes are the second visualization technique. In these maps the component values of the weight vectors are represented by a color code. Each of the component planes visualizes the distribution of one variable in the data set (Ultsch and Herrmann, 2005). By visually comparing those maps, variables with similar distributions can be detected and it helps in visually finding correlations between variables.

In the standard SOM-algorithm geographic coordinates included in the data set are considered as any other variable. The importance of the spatial variables during training of the map can be adjusted by assigning a weighting factor to these variables during the preprocessing stage. This procedure can be used to incorporate spatial information

in the algorithm, although it has to be noted that samples located far from the center of the data set are ill represented in the SOM. In order to overcome this problem, Bação et al. (2005) proposed the GEOSOM.

2.2 GEOSOM and GEO3DSOM

In the GEOSOM the spatial information of the data samples is explicitly included in the algorithm by altering the selection of the best-matching unit during the training into a two-step process. Firstly the unit is selected which lies geographically closest to the input vector. This means that the best-matching unit is searched based only on the geographic variables.

Secondly the unit with the smallest Euclidian distance, based on the complete input vector, within a predefined neighborhood of the geographically closest unit is chosen as best-matching unit. Subsequently the weight vector $\boldsymbol{M}_c$ of the best-matching unit and the weight vectors $\boldsymbol{M}_i$ within the neighborhood $N(t)$ are updated.

The size of the neighborhood to choose the best-matching unit from the units surrounding the geographically closest unit is determined by the variable k, the geographical tolerance. If k equals zero, the best-matching unit is the geographically closest unit. Setting k greater than zero, results in the search of the best-matching unit among the units within a radius k in output space of the geographically closest unit. If k approaches the size of the map, the result is equal to the standard SOM-algorithm.

The GEOSOM is only capable of including two geographic coordinates in the selection of the geographically closest unit, namely the X and Y coordinate. The GEO3DSOM is an extended version of the GEOSOM, capable of incorporating the third dimension, Z, in the selection of the geographically closest unit.

In order to give each geographic coordinate equal weight in the training process, each of the coordinate variables is rescaled so that their ranges are comparable, e.g. between $[0, 1]$.

HESSD
3, 1487–1516, 2006

3 Results

In the following section the standard SOM and the GEO3DSOM are applied to a theoretical data set and a real world hydrochemical data set.

3.1 Theoretical data set

The dataset consists of 1000 points of a cube, regularly spaced with an interval of 0.1 between [0.1, 1]. A variable D was added to this dataset with a value of 0 or 1. The distribution of variable D is shown in Fig. 2. This distribution results in 8 clusters (Table 1).

This data set is analyzed with both the standard SOM and the GEO3DSOM. The parameters used in the analysis are summarized in Table 2.

In both SOM-analysis the grid consists of 20 by 15 units, hexagonally ordered on a toroid shape. The use of rectangular array with a large number of units allows a good representation of the topology of the data set, while the hexagonal ordering provides more neighbors to each unit and border effects are avoided by using the continuous, finite shape of a toroid (Ultsch and Herrmann, 2005).

Figures 3 and 4 show respectively the results of the standard SOM-analysis and the GEO3DSOM-analysis. The U-matrices show the Euclidean distances between the reference vectors of the units by means of a gray scale (black: large distance, white: small distance). The units of the U-matrices are labeled with the cluster number of the sample assigned to the unit, according to Table 1.

Visual inspection of the U-matrices shows that both SOM-analysis are able to extract the clusters from the data. In the U-matrix of standard SOM-analysis a clear separation between groups is only visible between the clusters with D=1 and the clusters with D=0. Although the samples are grouped according to the defined clusters, no distinct borders between clusters are present. The U-matrix of the GEO3DSOM-analysis on the other hand, clearly shows that each cluster is separated from an other by a zone of high Euclidian distance between reference vectors.

The accompanying component planes can be used to explore the differences between the clusters. From both Fig. 3 and Fig. 4 it can be seen that the area with the samples of cluster 4 assigned to it, is characterized by $X>0.55$, $Y>0.55$, $Z<0.55$ and $D=1$.

In order to asses the quality of the SOM-analysis in representing the data set, three quality measures can be computed; the quantization error (qe), the topographic error (te) and the geographic error (ge). The quantization error measures the resolution of the SOM and is calculated as the average total Euclidian distance between an input vector and the reference vector of its best matching unit (Kohonen, 1995). The topographic error quantifies the preservation of the topology of the data by calculating the proportion of all data vectors for which first and second best matching unit are not adjacent units (Kohonen, 1995). Finally, the geographic error is a measure for the ability of the SOM to represent the geographic distribution of the data samples. It is calculated as the geographic distance between an input vector and its best matching unit. Table 3 summarizes the quality measures for the standard SOM-analysis and the GEO3DSOM-analysis.

The quantization error for the standard SOM is lower than for the GEO3DSOM, meaning that the representation of samples is better in the standard SOM. The GEO3DSOM, on the other hand, scores better in terms of topographic and geographic error. The representation of data by the GEO3DSOM is thus better capable of capturing the topology of the data and the geographic information included in the data.

3.2 Hydrochemical data set

The hydrochemical data set is obtained from a monitoring network of the Flemish Government in two regional, sandy, phreatic aquifers, made available through Databank Ondergrond Vlaanderen (DOV, 2006). The data set consists of 47 observation wells, each equipped with three well screens at different depths, resulting in a data set of 141 samples. Facilities in the monitoring well are designed to allow independent sampling of discrete depth intervals without mixing of groundwater of different depths.

The first aquifer, the Diest sands aquifer is of Late Miocene age and consists of coarse, glauconiferous sands and sandstones (Laga et al., 2001). The Brussels sands aquifer is of Middle Eocene age and is a heterogeneous formation consisting of an alteration of highly and poorly calcareous sands, which are locally silicified (Laga et al., 2001). Locally the Brussels sands are overlain by the younger sandy formations of Lede (Middle Eocene) and St. Huibrechts Hern (Early Oligocene). Both aquifers are covered with Quaternary eolian deposits consisting mainly of sands in the north and loam in the south.

Figure 5 shows the geological map of the study area and location of piezometers used in this study.

A sampling campaign was carried out in the spring of 2005 and from the 20 measured variables, a subset of 12 variables are considered in this analysis. Geographic coordinates, X, Y and the Z position above sealevel of the filter of each sample are included in the data set.

Histograms of the variables (Fig. 6) show that most of the variables are not normally distributed, but rather have a bimodal (Ca^{2+}, pH and HCO_3^-), skewed (e.g. K^+, Mg^{2+},O_2 and NO_3^-) or even a lognormal distribution ($Fe^{2+/3+}$ and Mn^{2+}). In order to avoid bias in the normalization or to make assumptions regarding the distribution of the variables, all parameters, including X, Y, Z are rescaled to a [0, 1] interval, according to:

$$x_{\text{new}} = \frac{x_{\text{old}} - \min(X)}{\max(X) - \min(X)} \tag{4}$$

A standard SOM analysis and a GEO3DSOM-analysis are carried out on the normalized data set. The parameters used in both analysis are summarized in Table 4.

The geographic tolerance k is set to 4 for the GEO3DSOM-analysis. This implies that the search of the BMU is restricted to the units lying within a radius of 4 units surrounding the geographically closest unit. The results of both analysis are depicted in Fig. 7 (standard SOM) and Fig. 8 (GEO3DSOM). The visualized results are (a) component planes, (b) U-matrix labeled with geology (B: Brussel sands, S: St. Huybrechts

Hern sands, D: Diest sands, Q: Quaternary deposits), (c) grouping of the SOM based on both the U-matrix and the component planes and finally (d) spatial distribution of the groups. The spatial distribution of the groups is organized per well screen, with screen 1 being the shallowest screen and screen 3 the deepest.

Table 5 renders the quality measures for both analysis. For the standard SOM the quantization error is slightly lower than the qe of the GEO3DSOM. The topologic error on the other hand is significantly lower for the standard SOM than for the GEO3DSOM. The performance of the standard SOM in capturing and representing the structure of the data set is higher than the performance of the GEO3DSOM. This is noticeable on the component planes (Fig. 7a and Fig. 8a), where distributions of the variables on the component planes of the standard SOM are rather smooth compared to those of the GEO3DSOM. This is mainly due to the better topological representation of the standard SOM.

The geographic error of the GEO3DSOM, however, is 15% smaller than in the standard SOM, implying that the geographic representation of the GEO3DSOM resembles the data set more closely than the data set.

On the U-matrices (Fig. 7b and Fig. 8b),it is also noticeable that the U-matrix of the GEO3DSOM divides the SOM in a large number of well separated groups, while the number of groups in the standard SOM is smaller and the borders between groups are less distinct. Based on U-matrices and component planes, both the standard SOM and the GEO3DSOM are manually divided in groups or clusters. This results in 6 clusters for the standard SOM and 8 clusters for the GEO3DSOM. Tables 6 and 7 list the main distinguishing features of respectively the standard SOM-based and the GEO3DSOM-based clustering.

Both SOM-variants succeed in distinguishing between samples originating from the Diest and the Brussels aquifers, as can be deduced from the geology-labeled U-matrices (Fig. 7b and Fig. 8b). The component planes reveal that pH and concentrations of calcium and bicarbonate are relatively high in the Brussels aquifer (groups 1 and 2 in the standard SOM and groups 1 to 3 in the GEO3DSOM). This difference

HESSD
3, 1487–1516, 2006

is due to the presence of calcite in the Brussels sands, while calcite is almost absent in the Diest sands (Laga et al., 2001; Lagrou et al., 2004). Although the difference in chemical composition of groundwater from the Brussels and the St. Huijbrechts Hern formation is rather small, the GEO3DSOM succeeds in grouping the samples from the St. Huijbrechts Hern aquifer, contrary to the standard SOM analysis, where the samples of the St. Huijbrechts Hern aquifer are distributed among the samples of the Brussels aquifer. Table 7 shows that the magnesium concentrations in the St. Huijbrechts Hern aquifer are slightly lower than in the Brussels aquifer.

Similarly, the GEO3DSOM succeeds in identifying a separate group containing the the samples of the Quaternary deposits on top of the Diest sands (group 8). These samples differ from the other samples in the Diest aquifer by their very low pH and higher potassium concentrations. Since these well screens are rather shallow and the Quaternary deposits consist of sands and gravels, the composition of the groundwater is very close to the composition of rain water, hence the lower pH. Within the Diest aquifer another division can be made, represented by groups 3 and 5 in the standard SOM and groups 5 and 7 in the GEO3DSOM. Groups 3 (standard SOM) and 7 (GEO3DSOM) have a slightly higher pH and alkalinity and are located in the south of the Diest aquifer.

Further subdivision of the SOM's is possible based on the concentrations of oxygen, nitrate, iron and manganese. In both aquifers there are zones with low oxygen concentrations and elevated iron and manganese (groups 2 and 4 in the standard SOM and groups 2, 3 and 6 in the GEO3DSOM). These groups consist of the deeper samples (Figs. 7d and 8d) and nitrate concentrations are on average lower in these groups. Due to the ubiquitous presence of iron and manganese bearing minerals like glauconite and iron-oxides in the Diest aquifer (Lagrou et al., 2004), the iron concentrations are rather high in the Diest aquifer when oxygen concentrations are low. Group 3 in the GEO3DSOM-analysis consists of the three well screens of the same well and represents an outlier in the data set. The three well screens are characterized by an absence of oxygen and very elevated iron and manganese concentrations. The same outlying

values can be observed on the U-matrix of the standard SOM (Fig. 7b), but they are not clearly separated from the other samples in group 2. In the standard SOM analysis, group 6 is very similar to group 5, but the relatively high manganese concentrations indicate that this group is also related to group 4.

Comparison of the spatial distribution of the groups of the standard SOM and the GEO3DSOM shows that the GEO3DSOM results in spatially more coherent groups. This effect is especially noticeable for groups 5 and 7 of the GEO3DSOM and the equivalent groups 5 and 3 of the standard SOM. Groups 5 and 3 of the standard SOM are not clearly separated geographically, while group 7 of the GEO3DSOM is distinctly separated from other groups and situated south of group 5.

4 Conclusions

The self-organizing map algorithm has proved to be a very valuable tool in the visualization and interpretation of large, multivariate data sets.

To incorporate spatial information in a self-organizing map analysis, GEO3DSOM is developed and its performance in clustering of both an artificial and a real life data set is compared to the standard SOM.

The performance of the standard SOM in correctly representing the structure of the data set and in minimizing the error between the input vectors and its best matching unit is higher than the performance of the GEO3DSOM on these criteria. The standard SOM is therefore very suitable for an exploratory data analysis in order to capture relationships between variables and the structure of data.

The GEO3DSOM on the other hand outperforms the standard SOM in providing a grouping of the data in a spatially coherent way. Analysis of both the artificial and the real life data sets showed that the GEO3DSOM is capable of a more detailed grouping of both regularly and irregularly distributed spatial data, compared to the standard SOM with geographical coordinates included in the data set.

Acknowledgements. The authors wish to thank AMINAL for providing data through their website (DOV, 2006).

References

Alvisi, S., Mascellani, G., Franchini, M., and Bardossy, A.: Water level forecasting through fuzzy logic and artificial neural network approaches, Hydro. Earth Syst. Sci., 10(1), 1–17, 2006. 1490

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology: Artificial neural networks in hydrology. II: Hydrologic applications, J. Hydrol. Eng., 5(2), 124–137, 2000. 1490

Bação, F., Lobo, V., and Painho, M.: The self-organizing map, the Geo-SOM, and relevant variants for geosciences, Comp. Geosci., 31(2), 155–163, 2005. 1489, 1493

Coppola, E., Szidarovsky, F., Poulton, M., and Charles, E.: Artificial neural network approach for predicting transient water levels in a multilayered groundwater system under variable state, pumping and climate conditions, J. Hydrol. Eng., 8(6), 348–360. 2003. 1490

Davis, J. C.: Statistics and data analysis in geology, John Wiley & Sons, Inc, New York, 1986. 1488

Databank Ondergrond Vlaanderen: http://dov.vlaanderen.be, 2006. 1495, 1500, 1513

Güler, C., Thyne, G. D., and McCray, J. E.: Evaluation of graphical and multivariate statistical methods for classification of water chemistry data, Hydrogeol. J., 10(4), 455–474, 2002. 1488

Hong, Y. S. and Rosen, M. R.: Intelligent characterisation and diagnosis of the groundwater quality in an urban fractured-rock aquifer using an artificial neural network, Urban Water. 3(3), 193–204, 2001. 1489

Jain, A. K., Mao, J., and Mohiuddin, K.: Artificial Neural Networks: a tutorial, IEEE Computer, 26(3), 31–44, 1996. 1490

Kaski, S.: Data exploration using Self-Organizing Maps. Acta Polytechnica Scandinavica: Mathematics, computing and management in engineering, Series No 82, 57, 1997. 1489, 1490

Kohonen, T.: Self-organizing maps. Springer, Berlin, 1995. 1489, 1490, 1495

Laga, P., Louwye, S., and Geets, S.: Paleogene and Neogene lithostratigraphic units (Belgium), Geol. Belgica 4(1–2), 135–152, 2001. 1496, 1498

Lagrou, D., Dreesen, R., and Broothaers, L.: Comparative quantitative petrographical analysis of Cenozoic aquifer sands in Flanders (N Belgium): overall trends and quality assessment, Materials Characterization 53, 317–326, 2004. 1498

Lambrakis, N., Antonakos, A., and Panagopoulos, G.: The use of multicomponent statistical analysis in hydrogeological environmental research, Water Res., 38(7), 1862–1872, 2004. 1488

Lischeid, G.: Taming awfully large data sets: using self- organizing maps for analyzing spatial and temporal trends of water quality data, Geophys. Res. Abstr., 5(01879), 2p., 2003. 1489

Love, D., Hallbauer, D., Amos, A., and Hranova, R.: Factor analysis as a tool in groundwater quality management: two southern African case studies, Phys. Chem. Earth, 29(15–18), 1135–1143, 2004. 1488

Sanchez-Martos, F., Aguilera, P. A., Garrido-Frenich, A., Torres, J. A., and Pulido-Bosch, A.: Assessment of groundwater quality by means of self-organizing maps: application in a semi-arid area, Environ. Manage., 30(5), 716–726, 2002. 1489

Ultsch, A. and Herrmann, L.: The architecture of emergent self-organizing maps to reduce projection errors. In: ESANN2005 13th European Symposium on Artificial Neural Networks, Bruges, Belgium, 1–6, 2005. 1492, 1494

Vesanto, J., Himberg, J., Alhoniemi, E., and Parhankangas, J.: Self-organizing map in Matlab: the SOM Toolbox, in: Matlab DSP Conference, Espoo, Finland, 35–40, 1999. 1492

Table 1. Theoretical data set.

X	*Y*	*Z*	*D*	Cluster
0.1–0.55	0.1–0.55	0.1–0.55	1	1
0.55–1	0.1–0.55	0.1–0.55	0	2
0.1–0.55	0.55–1	0.1–0.55	0	3
0.55–1	0.55–1	0.1–0.55	1	4
0.1–0.55	0.1–0.55	0.55-1	0	5
0.55–1	0.1–0.55	0.55–1	1	6
0.1–0.55	0.55–1	0.55–1	1	7
0.55–1	0.55–1	0.55–1	0	8

HESSD
3, 1487–1516, 2006

Table 2. Parameters of SOM and GEO3DSOM-analysis.

Parameter	Standard SOM	GEO3DSOM
size	20×15	20×15
grid	hexagonal	hexagonal
type	toroid	toroid
h	bubble	bubble
training mode	sequential	sequential
rough training		
epochs	50	50
r_{initial}	15	15
α_{initial}	0.7	0.7
fine training		
epochs	50	50
r_{initial}	4	4
α_{initial}	0.1	0.1
k	–	2

Table 3. Quality measures for the theoretical data set.

Quality measure	standard SOM	GEO3DSOM
qe	0.115	0.145
te	0.128	0.070
ge	0.100	0.097

HESSD
3, 1487–1516, 2006

Table 4. Parameters of SOM and GEO3DSOM-analysis.

Parameter	Standard SOM	GEO3DSOM
size	20×15	20×15
grid	hexagonal	hexagonal
type	toroid	toroid
h	bubble	bubble
training mode	sequential	sequential
rough training		
epochs	500	500
r_{initial}	10	10
α_{initial}	0.5	0.5
fine training		
epochs	500	500
r_{initial}	2	2
α_{initial}	0.2	0.2
k	–	4

Table 5. Quality measures for the hydrochemical data set.

Quality measure	standard SOM	GEO3DSOM
qe	0.127	0.139
te	0.076	0.160
ge	0.028	0.024

HESSD
3, 1487–1516, 2006

Table 6. Average concentrations per variable and geology of the standard SOM groups.

Group	1	2	3	4	5	6
pH	7.02	7.02	6.06	6.16	5.31	5.16
O_2 (mg/l)	4.86	0.24	8.32	0.53	3.53	2.48
Na^+ (mg/l)	14.55	13.49	17.95	11.07	16.7	14.75
K^+ (mg/l)	15.06	11.29	11.63	26.83	13.4	6.26
Mg^{2+} (mg/l)	10.82	11.08	6.34	4.58	9.11	13.13
Ca^{2+} (mg/l)	129.74	136.2	41.97	31.42	34.97	41.85
$Fe^{2+/3+}$ (mg/l)	0.22	2.98	3.27	15.32	0.44	0.03
Mn^{2+} (mg/l)	0.08	0.73	0.13	0.61	0.19	0.56
Cl^- (mg/l)	39.96	34.13	45.9	35.76	33.39	21.5
SO_4^{2-} (mg/l)	80.35	90.5	24.5	33.48	47.39	91.5
HCO_3^- (mg/l)	279.8	336.88	40.65	76.62	16.44	16.75
NO_3^- (mg/l)	41.46	1.53	74.71	0.55	85	72
Geology	Brussels & St. H. Hern	Brussels	Diest	Diest	Diest & Quaternary	Diest

Table 7. Average concentrations per variable and geology of the GEO3DSOM groups.

Group	1	2	3	4	5	6	7	8
pH	6.99	7.03	6.87	7.16	5.54	6.09	6.1	4.58
O_2 (mg/l)	5.56	0.3	0.57	3.12	3.69	1.12	9.66	8.35
Na^+ (mg/l)	13.75	13.14	15.23	17.23	17.4	12.84	16.06	13.58
K^+ (mg/l)	16.09	18.55	5.39	9.39	12.52	23.72	14.15	4.74
Mg^{2+} (mg/l)	12.02	5.46	19.23	7.91	7.93	5.64	6.37	13.5
Ca^{2+} (mg/l)	129.86	120.1	162.13	131.15	34.58	37.33	41.32	28.03
$Fe^{2+/3+}$ (mg/l)	0.13	1.46	5.98	0.22	1.67	13.51	0.08	0.02
Mn^{2+} (mg/l)	0.06	0.28	1.22	0.17	0.23	0.54	0.1	0.23
Cl^- (mg/l)	37	37.14	35	49.08	37	34.92	46.18	27
SO_4^{2-} (mg/l)	74.42	115.29	40.33	97	38.19	46.69	15.18	59.25
HCO_3^- (mg/l)	298.77	220.71	527	225.67	20	73.73	35.73	4.25
NO_3^- (mg/l)	39.48	2.9	1.03	54.19	75.15	15.75	73.36	125.75
Geology	Brussels	Brussels	Brussels	St. H. Hern	Diest	Diest	Diest	Quaternary

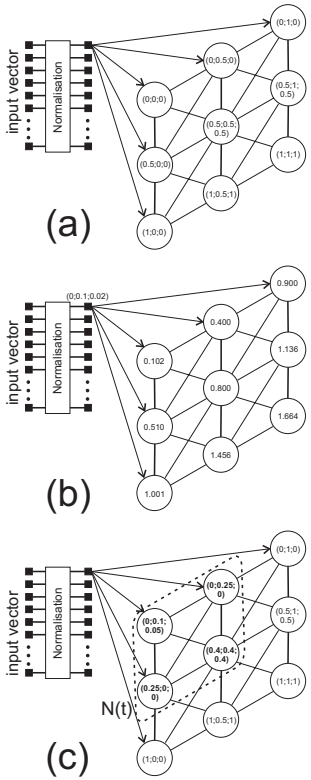

Fig. 1. SOM-algorithm.
(a) Initialization reference vectors.
(b) Calculation of Euclidean distance between input vector and reference vectors.
(c) Assignment of input vector to its BMU and update of reference vectors within neighborhood N(t).

HESSD
3, 1487–1516, 2006

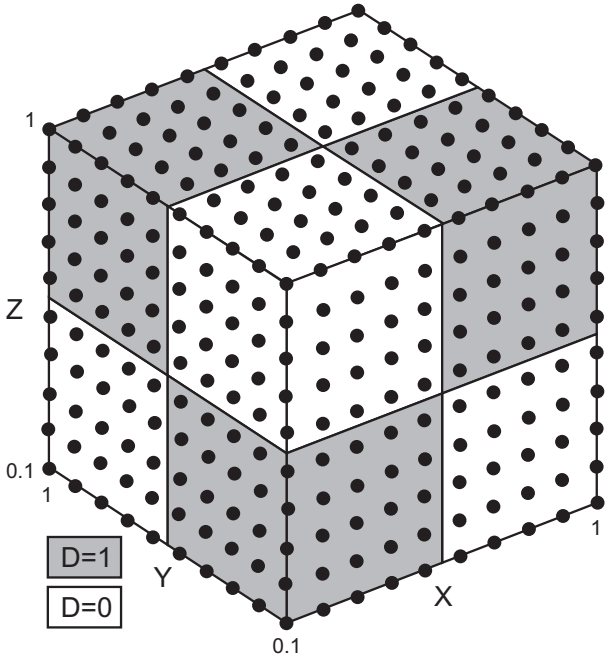

Fig. 2. Theoretical data set.

HESSD
3, 1487–1516, 2006

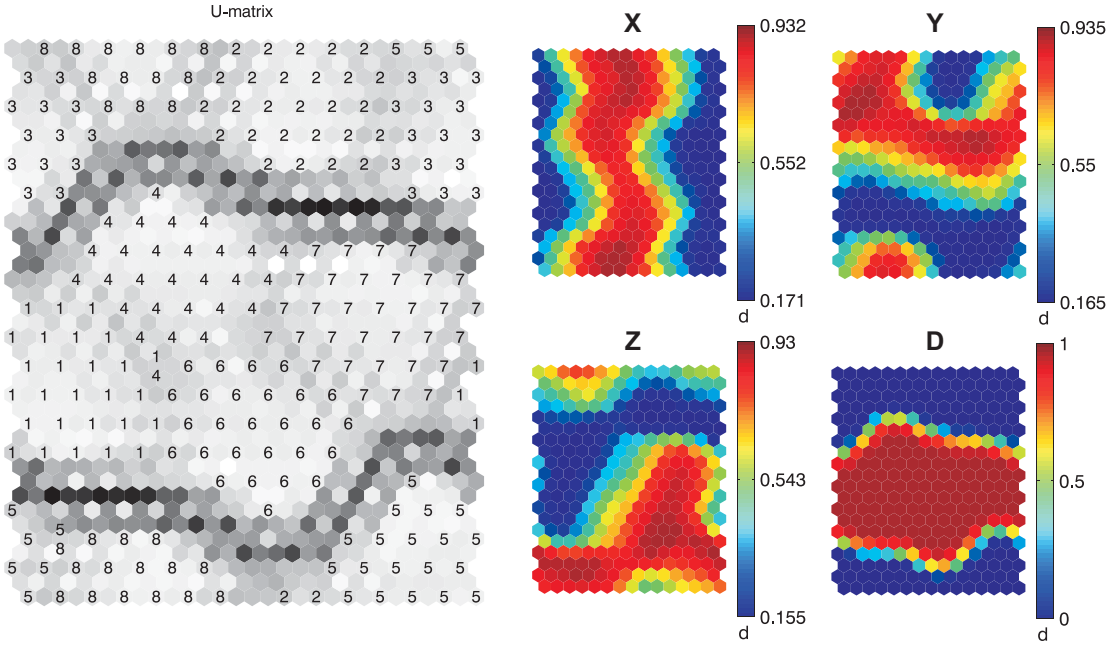

Fig. 3. U-matrix(left) and component planes (right) of the standard SOM-analysis of the theoretical data set.

HESSD
3, 1487–1516, 2006

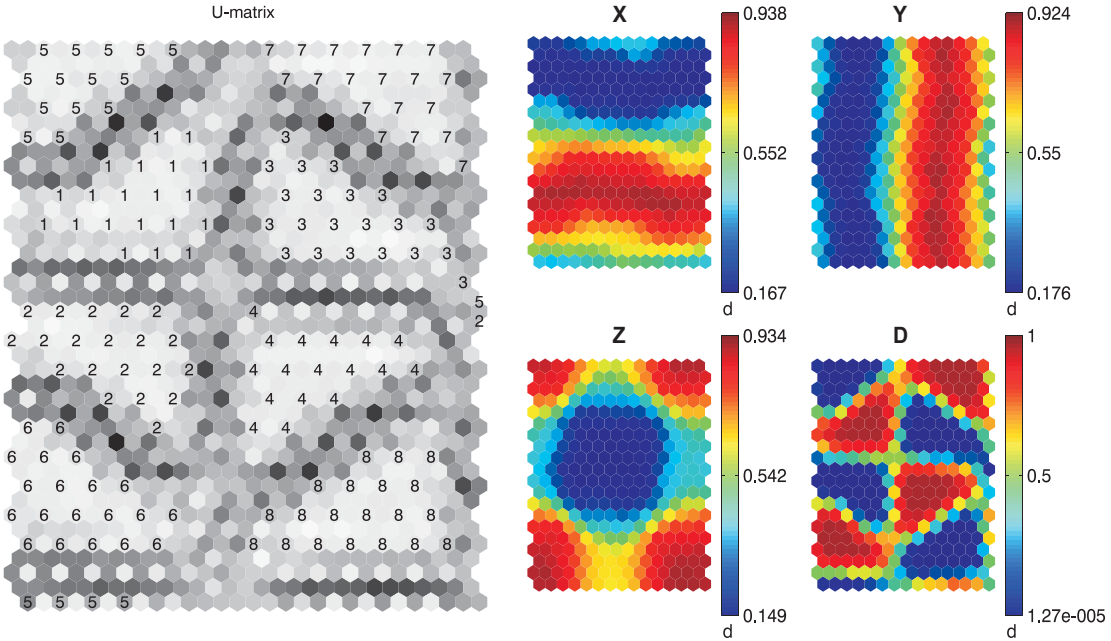

Fig. 4. U-matrix(left) and component planes (right) of the GEO3DSOM-analysis of the theoretical data set.

HESSD

3, 1487–1516, 2006

Analysis of spatial hydrogeologic data using GEO3DSOM

L. Peeters et al.

EGU

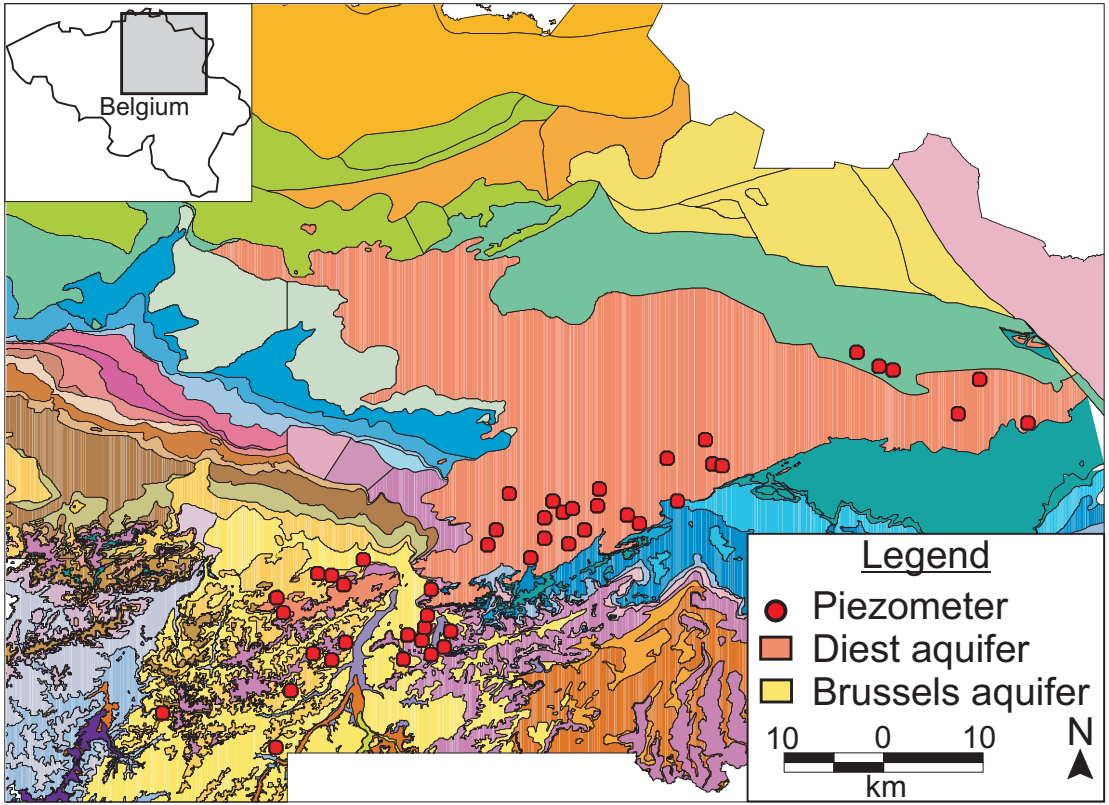

Fig. 5. Study area (after DOV, 2006).

HESSD
3, 1487–1516, 2006

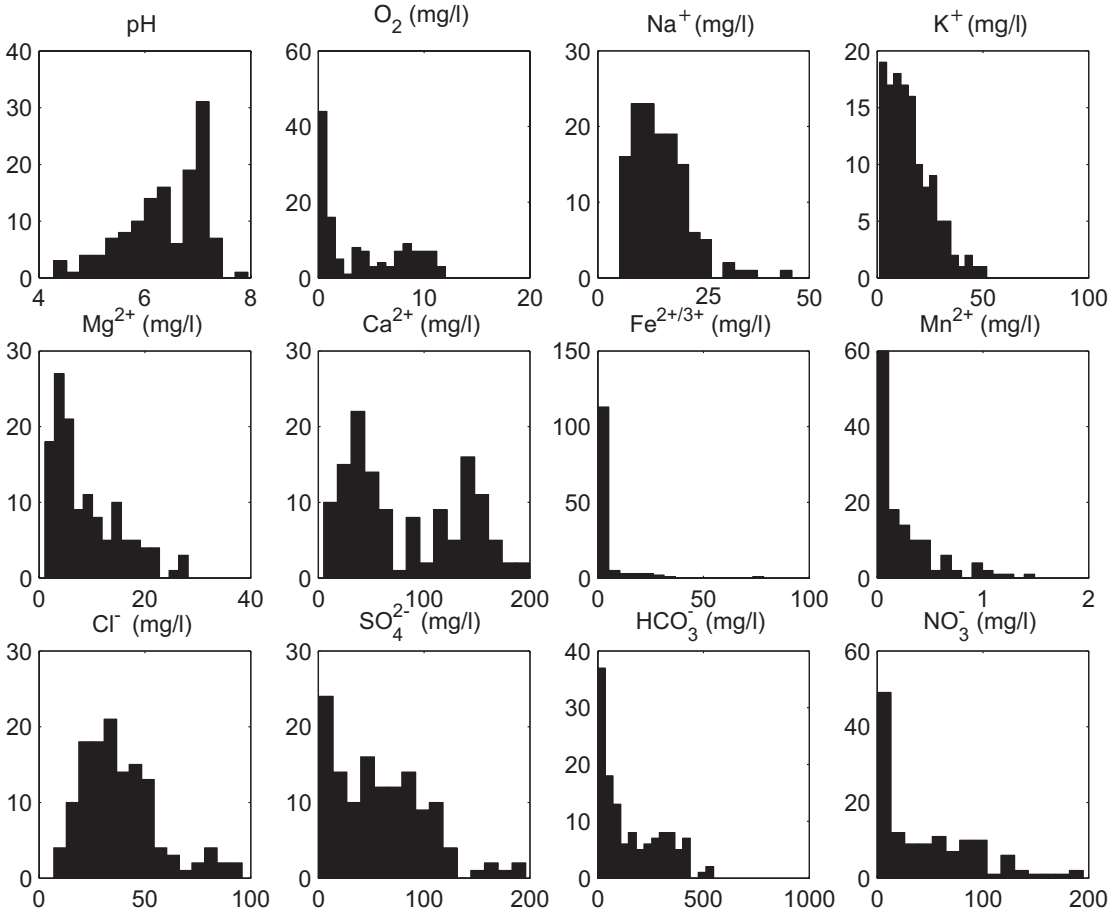

Fig. 6. Histograms of the hydrochemical data set.

HESSD
3, 1487–1516, 2006

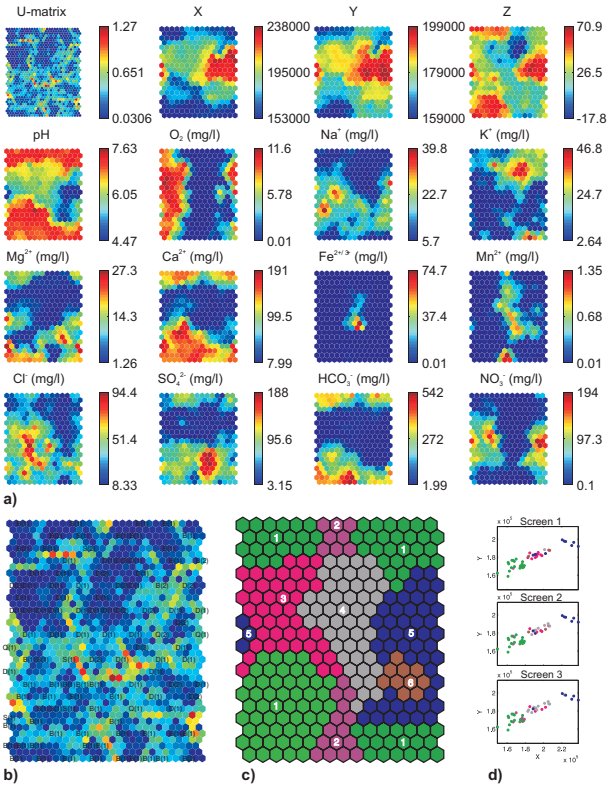

Fig. 7. Results of the SOM-analysis of the hydrochemical data set.
(a) U-matrix and component planes;
(b) labeled U-matrix;
(a) grouping of U-matrix;
(d) spatial distribution of groups, per well screen.

HESSD
3, 1487–1516, 2006

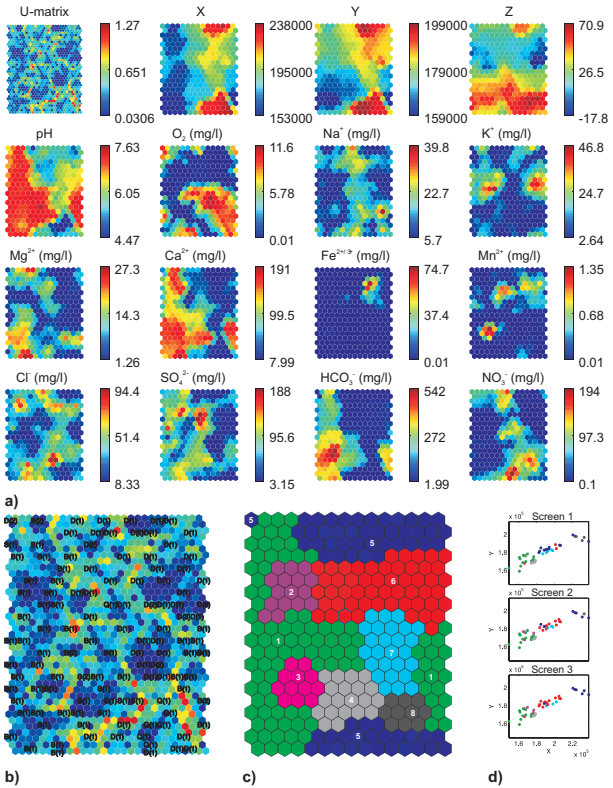

Fig. 8. Results of the GEO3DSOM-analysis of the hydrochemical data set.
(a) U-matrix and component planes;
(b) labeled U-matrix;
(c) grouping of U-matrix;
(d) spatial distribution of groups, per well screen.

HESSD
3, 1487–1516, 2006

EGU