# Comparison of Kohonen's Self-Organizing Map algorithm and principal component analysis in the exploratory data analysis of a groundwater quality dataset

Luk Peeters[1] and Alain Dassargues[1,2]

[1] Applied Geology and Mineralogy, KULeuven, Celestijnenlaan 200E, Heverlee, Belgium luk.peeters@geo.kuleuven.be

[2] Hydrogeology and Environmental Geology, ULG, GEOMAC B52/3, Lige, Belgium alain.dassargues@ulg.ac.be

**Summary.** Regional monitoring of groundwater chemistry yields large, multivariate data sets. Summarizing available data, extracting useful information and formulating hypotheses for further research are the key aspects in the exploratory data analysis of these data sets. Traditionally multivariate statistical techniques such as principal component analysis (PCA) are applied for this purpose. In PCA a linear dimensionality reduction of the original, high dimensional data set is carried out in order to identify orthogonal directions (principal components) of maximum variance in the dataset based on linear combinations of correlated variables [3].

In this study, PCA is compared to the Self-Organizing Map (SOM) algorithm. The SOM-algorithm is a neural network designed to carry out a non-parametric regression process in order to represent high-dimensional, nonlinearly related data items in a topology-preserving, often two-dimensional display, and to perform unsupervised classification and clustering [11].

PCA and SOM are applied to a groundwater chemistry data set from a regional monitoring network in two sandy, phreatic aquifers in Central Belgium. The 47 monitoring wells are each equipped with three well screens at different depths, in which 14 variables are measured.

Both techniques succeed in distinguishing between both aquifers and reveal the apparent relationships between variables. The main advantage of PCA is the expression of each variable in terms of the principal components and the quantification of the amount of variance explained by each component. The visualization of the SOM-analysis has the advantage of allowing a straightforward interpretation of the structure of the data set in which even non-linear relationships can be identified. Additionally, the SOM-algorithm can handle a limited amount of missing values in the data set, contrary to PCA.

**Key words:** Principal Components Analysis, Self-Organizing Maps, Groundwater chemistry

## 1 Introduction

Monitoring of groundwater chemistry typically yields a large number of samples, analyzed for numerous chemical and physical parameters thus creating larger, high-dimensional data sets. A wealth of methods exist for analyzing and interpreting such data sets, ranging from univariate and bivariate data analysis [17, 1] to graphical techniques such as Piper and Stiff-diagrams [14, 18] and multivariate statistical techniques such as K-means clustering, R-mode factor analysis and principal component analysis [5].

Principal component analysis is a widely applied technique in hydrogeological research [15, 9, 7, 2] in which a linear dimensionality reduction of the original,high dimensional data set is carried out by identifying orthogonal directions or principal components (PC) of maximum variance in the data sets based on linear combinations of correlated variables. Projections of the original data into the subspace spanned by the principal components can be used to identify groups in the data and to reveal relationships between variables [3].

The Self-Organizing Map algorithm developed by Kohonen is an artificial neural network designed for visualizing and analyzing high dimensional data sets by a grouping of data based on similarity and visualizing them on a, typically two-dimensional, grid [11]. Although this method is frequently used in a.o. financial, medical, chemical and biological research (an overview is presented in [10]), there are few cases in which the SOM-algorithm is applied in hydrogeological research. Hong & Rosen [8] applied the technique to diagnose the effect of storm water infiltration on groundwater quality variables and to capture complex nonlinear relationships between the measured parameters. Sanchez-Martos et al. [16] used a SOM-analysis in the classification of a hydrochemical data set from a detritic aquifer in a semi-arid region into distinct classes of different chemical composition. In [6], self-organizing maps are used in combination with a radial basis neural network for the prediction of time series of groundwater level. Peeters et al. [13] uses a modified SOM-algorithm to create a spatially coherent grouping of groundwater samples

In this study PCA and SOM are applied to a hydrochemical data set of a regional groundwater chemistry monitoring network in two phreatic, sandy aquifers in Central Belgium. The goals of the data analysis are (1) visualization of the data set, (2) identification of groups of similar chemical composition in both aquifers and (3) exploring relationships between groundwater quality variables. Performance of principal component analysis and self-organizing map analysis in achieving these goals will be evaluated and both strong and weak points of the techniques will be highlighted.

# 2 Methods

## 2.1 Data set

The hydrochemical data set is obtained from a monitoring network of the Flemish Government in two phreatic, sandy aquifers, made available through Databank Ondergrond Vlaanderen [4]. 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.

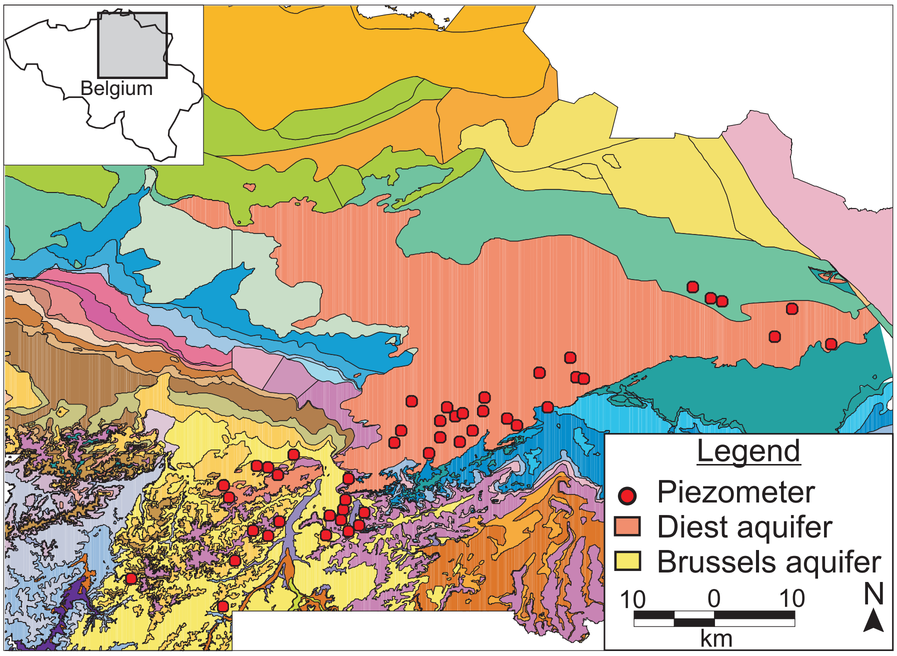

**Fig. 1.** Study area (after [4])

The first aquifer, the Diest sands aquifer is of Late Miocene age and consists of coarse, glauconiferous sands and sandstones [12]. 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 [12]. 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. Fig. 1 shows the geological map of the study area and location of piezometers used. A sampling campaign was carried out in the spring of 2005. From the 20 measured variables, a subset of 14 variables, including depth of well screen and thickness of unsaturated zone are considered in this analysis.

### 2.2 Principal Component Analysis

Principal component analysis projects a data set in a new coordinate system in such a way that the maximum variability in the data set is projected along the axes. This projection is carried out by transforming a set of correlated variables into a set of uncorrelated, orthogonal variables, the principal components, which are ordered by reducing variability. These uncorrelated variables are the eigenvectors of the variance-covariance matrix and can be expressed as linear combinations of the original variables. The eigenvalue of each eigenvector expresses the amount of variance explained by the eigenvector. There are as many eigenvectors as there are original variables in the data set. Since the last of these principal components only account for a negligible amount of the variation in the data set, these can be omitted and a dimensionality reduction is achieved. A detailed description of principal component analysis can be found in [3] and [21].

### 2.3 Self-Organizing Map algorithm

The Self-Organizing Map algorithm is an unsupervised neural network technique which classifies data according to their similarity by plotting the high-dimensional data onto a two-dimensional grid in a topology preserving way, so that the structure of the data set is rendered in the visualization [11]. The network architecture is shown in Fig.2.

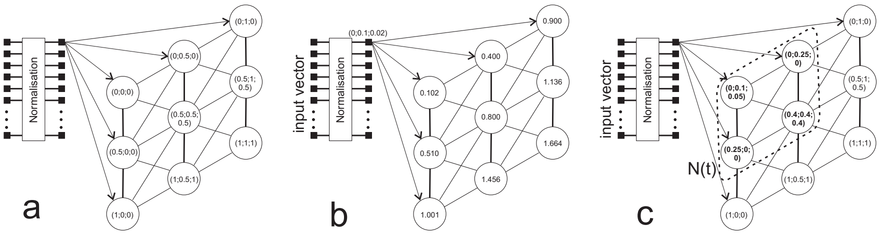

**Fig. 2.** SOM-algorithm a) Initialization reference vectors b) Calculation of Euclidean distance between input vector and reference vectors c) Assignment of input vector to its BMU and update of reference vectors within neighborhood
(after [13])

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 [11, 10]:

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 and the Euclidean distances between the considered input vector $X$ and all of the reference vectors $M_i$ are calculated according to:
$$\|X - M\| = \sqrt{\sum_{i=1}^{n}(x_i - m_i)^2} \tag{1}$$
4. The best matching unit $M_c$ is chosen according to:
$$\|X - M_c\| = \min_i\{\|X - M_i\|\} \tag{2}$$
This step is illustrated in Fig. 2b.
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 closer to the input vector (Fig. 2c). The factor $\alpha(t)$ controls the rate of change of the reference vectors and is called the learning rate.
$$M_i(t+1) = \begin{cases} M_i(t) + \alpha(t)[X(t) - M_i(t)] & \forall \in N(t) \\ M_i(t) & \forall \notin N(t) \end{cases} \tag{3}$$
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 $\alpha$ 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 [20]. 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 also represented by a color code. Each of the component planes visualizes the distribution of one variable in the data set [19]. By visually comparing those maps, variables with similar distributions can be detected and it helps in visually finding relationships between variables.

## 3 Results

### 3.1 Principal Component Analysis

Since PCA is based on the calculation of eigenvalues of the covariance matrix, missing values are not allowed in the data matrix. 11 samples containing missing value were omitted for analysis, reducing the data set to a matrix of 131 observations. A normalization per variable of the data is carried out due to the differences in measuring units and ranges by rescaling each variable in the $[0, 1]$ range.

Fig. 3 shows the scree plot of the analysis for the first 10 PC's, representing the percentage of variance explained by each of the principal components. Only the PC's explaining more than 10 % of the variance are retained in this study. The first three principal components explain more than 10% of the variance and together they explain approximately 67 % of the variance. Table 1 represents the contribution of each variable to these three principal components together with their eigenvalues and the percentage of variance explained. PC1 explains most of the variance (29.3 %) and is dominated by pH, calcium and alkalinity and to a lesser extent by magnesium and sulphate. The second component explains 23.2 % of the variance and is dominated by oxygen, nitrate and depth of well screen. The third component includes oxygen, depth and thickness of the unsaturated zone and accounts for 14.5 % of the total variance.

Bivariate plots of the scores of the PC's are shown in Fig. 4 and visualize the differences between samples. The plot of PC1 vs PC2 enables the differentiation between samples from the Brussels and the Diest aquifer, based on principal component 1. The Diest samples have lower values for pH, calcium and alkalinity compared to samples from the Brussels aquifer. Samples from Quaternary deposits are plotted in the upper left corner, reflecting a

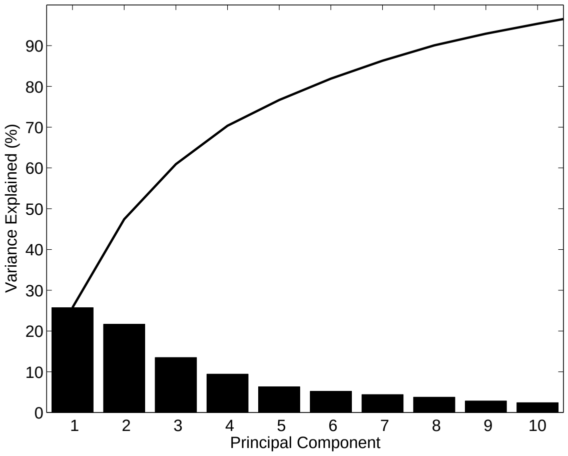

**Fig. 3.** Scree plot

**Table 1.** Loadings of first three principal components

| Variable | PC1 | PC2 | PC3 |
|---|---|---|---|
| pH | 0.32 | -0.2 | 0.2 |
| $O_2$ (mg/l) | 0.17 | 0.65 | 0.43 |
| $Na^+$ (mg/l) | 0.05 | 0.22 | -0.17 |
| $K^+$ (mg/l) | -0.15 | 0.14 | -0.26 |
| $Mg^{2+}$ (mg/l) | 0.3 | 0.11 | -0.22 |
| $Ca^{2+}$ (mg/l) | 0.59 | -0.09 | -0.09 |
| $Fe^{2+/3+}$ (mg/l) | -0.09 | -0.09 | -0.06 |
| $Mn^{2+}$ (mg/l) | -0.14 | -0.2 | -0.26 |
| $Cl^-$ (mg/l) | 0.14 | 0.15 | -0.09 |
| $SO_4^{2-}$ (mg/l) | 0.27 | -0.05 | -0.38 |
| $HCO_3^-$ (mg/l) | 0.51 | -0.22 | 0.04 |
| $NO_3^-$ (mg/l) | 0.03 | 0.49 | -0.21 |
| depth | -0.13 | -0.29 | 0.31 |
| unsat | 0.09 | -0.02 | 0.51 |
| Variances | 0.21 | 0.17 | 0.1 |
| % explained | 29.2 | 23.2 | 14.5 |
| Cumulative | 29.2 | 52.5 | 66.8 |

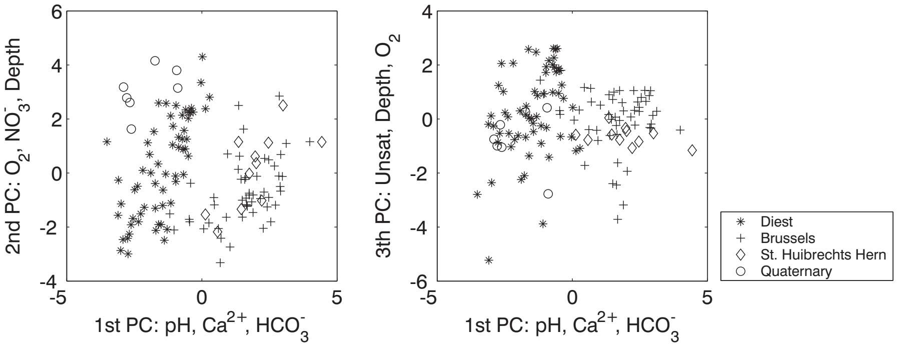

**Fig. 4.** Bivariate plots of PC1 vs PC2 and PC1 vs PC3

combination of high nitrate and oxygen concentrations and low pH, calcium and alkalinity values. Samples of the St. Huibrechts Hern formation are indistinguishable from the Brussels sands samples. Fig. 4 also shows the plot of PC1 vs PC3, which shows that in both aquifers samples are present with a thin unsaturated zone, situated at shallow depth and having low oxygen concentrations.

### 3.2 Self-Organizing Map

Since the number of missing values is less than 10 %, the entire data set of 141 samples and 14 variables is used in the SOM-analysis [10]. The same normalization as for the PCA-analysis is applied to the data set.

A hexagonal, toroid grid of 18 by 14 nodes is chosen for the SOM-design in which the reference vectors are initialized randomly. A Gaussian function is used as neighborhood function. During the training phase the data set is shown 500 times to the SOM and the radius of the neighborhood function as well as the learning rate decrease linearly during training.

Fig. 5a and b show the resulting U-matrix and component planes of the SOM-analysis. In the U-matrix, each node is labeled with the geology of the input vector assigned to it. It can be noted that not all nodes are assigned an input vector and that some nodes are assigned multiple input vectors.

The quality of the SOM-analysis is expressed by two error measures; the quantization error ($qe$) and the topological error ($te$). The quantization error is a measure of the resolution of the map and is calculated as the average Euclidean distance between an input vector and its best matching unit. The topological error assesses the ability of the SOM to represent the topology of the data set and is defined as the percentage of input vectors for which the best matching unit and its second best matching unit are not neighboring nodes on the grid. For this SOM-analysis the quantization error and topological error is 0.28 and 0.05 respectively.

Fig. 5a shows the ability of the SOM to distinguish between samples of the

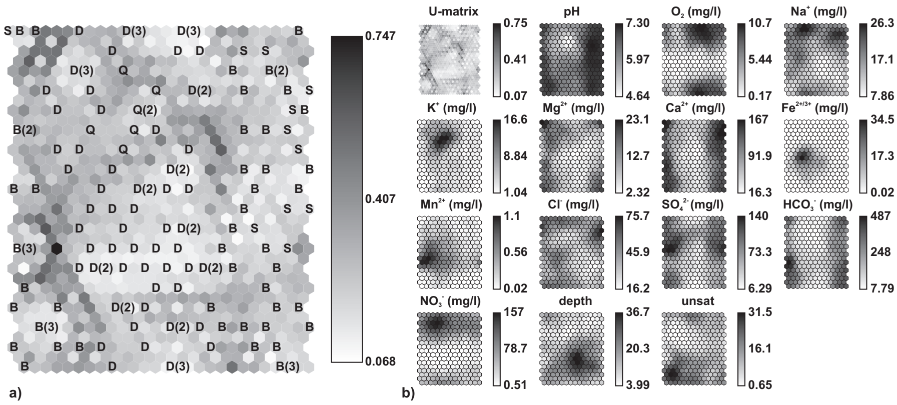

**Fig. 5.** a)U-matrix (Q=Quaternary, D=Diest aquifer, S=St. Huibrechts Hern aquifer, B=Brussels aquifer) b)Component Planes

Diest and Brussels aquifer. From Fig. 5b it can be seen that the difference between both groups lies in the concentrations of pH, calcium and alkalinity. Sulphate and magnesium have a similar distribution to these parameters, although distinct differences exist.

In the zone dominated by samples of the Diest aquifer a subgroup can be delineated with predominantly samples from Quaternary deposits. Besides the shallow depth and small unsaturated zone, they are characterized by very low pH-values and high potassium concentrations. Similarly as in the PCA-analysis, the samples of the St. Huibrechts Hern aquifer cannot be distinguished from the samples of the Brussels aquifer.

The component planes of oxygen, nitrate, iron and manganese show a central, horizontal band of samples with low oxygen and nitrate concentrations and elevated iron and manganese values. In this zone there is a dominance of samples from deeper filters. Although this subdivision appears to be independent of geology of the sample, it can be noted, that when iron is present in the groundwater, the concentrations are higher in the Diest aquifer than in the Brussels aquifer. The only exception is a cluster of three Brussels-aquifer samples situated on the central-left part of the grid. These samples originate from one and the same monitoring well and present anomalous high iron and manganese concentration together with a shallow sampling depth.

## 4 Discussion

The first goal of the data analysis was to visualize the data set. Both techniques succeed in providing summarizing plots of the information contained in the data set. While the position of each sample in the bivariate plots of the principal components (Fig. 4) is based on a linear combination of the original variables and thus gives an exact representation of the sample, in the visualization of the SOM (Fig. 5) a difference exists between the input vector and its best matching unit representing the input vector in the visualization. The interpretation of the SOM on the other hand is more straight forward compared to bivariate plots of PCA.

The delineation of groups of samples with a similar chemical composition is in the principal component analysis limited to the distinction between three groups, namely Diest samples, Brussels samples and Quaternary samples. The visualization of the self-organizing map gives a more detailed view on the structure of the data and allows distinguishing more groups in the data set. The SOM-analysis even reveals the presence of a monitoring well with outlying values in the Brussels sands aquifer, which remained undetected in the PCA-analysis. Additionally the characteristics of each group and the differences between groups can directly be derived from the visualizations.

The last goal of the exploratory data analysis is to explore relationships between variables. Since principal component analysis is based on the covariance matrix, the contribution of each variable in the principal components gives a direct quantification of the relationship between variables. It has to be noted however that principal component analysis is restricted to detecting linear relationships. Exploring relationships between variables in the SOM is based on a visual comparison of the distribution of the values of the variables on the component planes. This gives rise to a certain degree of subjectivity, although this allows for the detection of non-linear relationships.

## 5 Conclusions

Principal component analysis and the self-organizing map algorithm are used in the exploratory data analysis of a hydrochemical data set in order to evaluate and compare both techniques.

The main advantages of PCA are: (1) the ability to group correlated variables in principal components and thus reducing dimensionality of the data set, (2) the revelation of global structures in the data set and (3) the quantitative measure of variance explained by the PC's and the contribution of each variable in these PC's. The disadvantages include the difficulty to express the results of the principal component projection in a straight forward manner and the limitation of only taking in account linear relationships between variables.

Self organizing maps prove to be able to visualize the entire data set in terms of the original variables and to provide a measure of similarity between samples allowing grouping of samples. The visualization also enables the detection of nonlinear relationships between variables. The main disadvantages lies in the subjective nature of defining clusters and establishing relationships between variables.

## 6 Acknowledgements

The authors wish to thank AMINAL for providing data through their website [4]

## References

1. Antonopoulos VZ, Papamichail DM and Mitsiou KA, (2001) Statistical and trend analysis of water quality and quantity data for the Strymon River in Greece. Hydrology and Earth System Sciences 5:679-691
2. Cruz JV and Franca Z (2006) Hydrogeochemistry of thermal and mineral water springs of the Azores archipelago (Portugal). Journal of Volcanology and Geothermal Research 151:382-398
3. Davis JC (1986) Statistics and data analysis in geology. John Wiley & Sons, Inc, New York
4. Databank Ondergrond Vlaanderen: http://dov.vlaanderen.be, 2006
5. Güler C, Thyne GD and McCray JE, (2002) Evaluation of graphical and multivariate statistical methods for classification of water chemistry data. Hydrogeology Journal 10:455-474
6. Gwo-Fong Lin L-HC, (2005) Time series forecasting by combining the radial basis function network and the self-organizing map. Hydrological Processes 19:1925-1937
7. Helena B, Pardo R, Vega M, Barradeo E, Ferndandez JM and Fernandez L, (2000) Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga River, Spain) by principal component analysis. Water Research 34:807-816
8. Hong Y-S and Rosen MR (2001) Intelligent characterisation and diagnosis of the groundwater quality in an urban fractured-rock aquifer using an artificial neural network. Urban Water 3:193-204
9. Join J-L, Coudray J and Longworth K, (1997) Using principal component analysis and Na/Cl ratios to trace groundwater circulation in a volcanic island: the example of Reunion. Journal of Hydrology 190:1-18
10. Kaski S, (1997) Data exploration using Self-Organizing Maps. Acta Polytechnica Scandinavica: Mathematics, computing and management in engineering, Series No 82 57
11. Kohonen T (1995) Self-organizing maps. Springer series in information sciences, vol 30. Springer, Berlin
12. Laga P, Louwye S and Geets S, (2001) Paleogene and neogene lithostratigraphic units (Belgium). Geologica Belgica 4:135-152
13. Peeters L, Bação F, Lobo V and Dassargues A, (in press) Exploratory data analysis and clustering of multivariate spatial hydrogeological data by means of GEO3DSOM, a variant of Kohonens Self-Organizing Map. Hydrology and Earth System Sciences Discussions
14. Piper AM, (1944) A graphic procedure in the geochemical interpretation of water-analyses. Transactions-American Geophysical Union 25:914-923
15. Rouhani S and Wackernagel H, (1990) Multivariate geostatistical approach to space-time data analysis. Water Resources Research 26:585-591

16. Sanchez-Martos F, Aguilera PA, Garrido-Frenich A, Torres JA and Pulido-Bosch A, (2002) Assessment of groundwater quality by means of self-organizing maps: application in a semi-arid area. Environmental Management 30:716-726
17. Spruill TB, (2000) Statistical evaluation of effects of riparian buffers on nitrate and groundwater quality. Journal of Environmental Quality 29:1523-1538
18. Stiff H, (1951) The Interpretation of Chemical Water Analysis by Means of Patterns. Journal of Petroleum Technology 3:15-17
19. Ultsch A and Herrmann L (2005) The architecture of emergent self-organizing maps to reduce projection errors. ESANN2005: 13th European Symposium on Artificial Neural Networks, Bruges, Belgium, 1-6
20. Vesanto J, Himberg J, Alhoniemi E and Parhankangas J (1999) Self-organizing map in Matlab: the SOM Toolbox. Matlab DSP Conference, Espoo, Finland, 35-40
21. Wackernagel H (2003) Multivariate Geostatistics Third edition. Springer, Paris, 387