Non-Linear Approaches to Assess Water and Soil Quality

Gunnar Lischeid

Abstract Systematic monitoring is indispensable for a thorough water and soil management. However, large data sets with many variables, natural heterogeneities, and a variety of (possible) influencing factors require new approaches for processing and visualization of the data. A variety of advanced techniques has been developed recently in different disciplines. Some of them have been tested for application in water and soil resources management and exhibited very promising results. Two out of these approaches are presented here by application to a data set of shallow groundwater quality that has been complied during a five years period in a small catchment in Northeast Germany. Measured variables of soil or water quality usually reflect effects of various processes. On the other hand, single processes usually affect more than one variable and thus generate a characteristic "fingerprint" that can be used in an inverse approach to identify this process based on observed measured variables. Other processes differ with respect to their "fingerprints" and thus can be differentiated in a large data set. This is the basic idea of applying dimensionality reduction approaches. Every single sample can be ascribed a score of a component that is a quantitative measure for the impact of the respective process on the given sample. Usually, a small number of components (or processes, respectively) accounts for a large fraction of the variance in a data set with many variables. This "dimensionality reduction" helps a lot to gain better understanding of the prevailing processes, of spatial and temporal patterns, and of the reasons for conspicuous data. The larger a given data set, and the larger the number of variables, the more advanced methods of data visualization are required. Modern visualization techniques pave the way for efficient use of the most powerful interface between data stored on a computer and the human brain. A single non-linear projection of high-dimensional data on a two-dimensional graph provides comprehensive information about outliers, clusters, linear and non-linear relationships, spatial patterns, multivariate trends, etc. in the data. This

G. Lischeid (✉)
Leibniz Centre for Agricultural Landscape Research (ZALF) e. V.,
Eberswalder Strasse 84, 15374 Muencheberg, Germany
e-mail: lischeid@zalf.de

approach could usefully be combined with other dimensionality reduction techniques. This chapter can serve only as an appetizer. A variety of sophisticated new methods exist. These techniques still are not part of textbooks of hydrology or soil science. They require an open mind and some initial training. Then a wealth of powerful tools are at hand as a base for thorough water and soil resources management.

Keywords Water quality · Soil quality · Statistical methods

1 Motivation

Scarce resources need to be managed very carefully. Water quantity, water quality and soil quality certainly are scarce and valuable resources in most parts of the world, and are intimately intertwined. None of them could be replaced by any technical means at affordable prices and at the required scale. Water and soil resources management has to grasp the actual state, current and future threats and options, has to implement and perform measures, and, last but not least, has to evaluate the effects and possible harmful side-effects of the adopted measures as well as the efficiency of the management. Thus a well-designed monitoring scheme is indispensable for providing the required data with the necessary spatial and temporal resolution (Muller 2012).

Natural systems usually exhibit considerable heterogeneity in space and time which any monitoring has to account for. This requires a large number of replicates, yielding large data sets that are not easy to handle. Moreover, usually a multitude of different processes act in parallel and at different scales and have an impact on the observed variables. In return, single measured variables usually are prone to various processes that will be reflected by the measurement data, although to unknown quantities. Last but not least, the monitoring scheme itself needs to be evaluated and to be checked with respect to efficiency as well as to possible errors etc.

2 Example Data Set

Here two new techniques will be presented that have been developed recently and have proved successful in a variety of studies on soil and water quality. Both methods explicitly allow for non-linearities that are to the author's experience more the rule rather than the exception for real world data. Statistical analyses and graphs have been performed using the R software environment (R Development Core Team 2006) which is freely available at http://www.r-project.org.

Table 1 Pearson correlation between the measured variables

	NH$_4$	NO$_3$	P	Cl	SO$_4$	pH	Redox potential	EC
NH$_4$		−0.2	0.06	0.03	−0.1	−0.08	0.12	−0.01
NO$_3$	−0.2		−0.14	−0.15	−0.19	−0.08	0.1	−0.1
P	0.06	−0.14		0	−0.08	0.19	−0.12	−0.05
Cl	0.03	−0.15	0		0.31	0.08	0.02	0.62
SO$_4$	−0.1	−0.19	−0.08	0.31		0.04	−0.07	0.84
pH	−0.08	−0.08	0.19	0.08	0.04		−0.01	0.01
Redox potential	0.12	0.1	−0.12	0.02	−0.07	−0.01		−0.04
EC	−0.01	−0.1	−0.05	0.62	0.84	0.01	−0.04	

Both methods have been applied to a data set that was kindly provided by Dr. Uwe Schindler, Leibniz Centre for Agricultural Landscape Research, Müncheberg, Germany. It comprises data from 718 samples of shallow groundwater, sampled between March 2000 and May 2004 by piezometers at 2–4 m depth at seven different sites in the Quillow catchment, about 100 km north of Berlin, Germany. Land use was arable land mostly, but included grassland and forest sites as well. In total, 41 different pipes have been sampled. Number of temporal replicates per pipe varied between 1 and 38.

Eight variables have been determined: Concentration of NH$_4$, NO$_3$, P, Cl, and SO$_4$, pH value (pH), redox potential, and electric conductivity (EC). The data were tested for plausibility prior to the analysis. Values below the level of quantification were replaced by half the value of quantification level. To compensate for different data ranges for different variables, the data were z-normalized for each variable separately (values divided by the respective mean value, and then divided by the standard deviation).

Most variables are only slightly correlated with each other. Only SO$_4$ and Cl concentration are clearly correlated with electric conductivity and are closely related to each other (Table 1). This data set is rather small with respect to the number of samples as well as to the number of measured variables compared to some recent applications (Lischeid and Bittersohl 2008; Lischeid 2009; Schilli et al. 2010). Thus it is well suited to serve as an example. A good textbook that provides in-depth information about these methods and many more related approaches is given, e.g., by Lee and Verleysen (2007).

3 Dimensionality Reduction and Process Identification

Measured variables of soil or water quality usually reflect effects of various processes. E.g., nitrate concentration is affected by fertilisation, uptake by plants, nitrogen release by decomposition of organic matter, oxidation of ammonium, leaching, denitrification, etc. On the other hand, single processes usually affect more than one variable. E.g., denitrification requires reduced oxygen availability

and low redox potential, which often comes along with enhanced Fe(II) concentration or sulphate concentration due to oxidation of sulphides. This interplay generates a characteristic "fingerprint" that can be used in an inverse approach to identify this process based on observed measured variables. Other processes differ with respect to their "fingerprints" and thus can be differentiated in a large data set. This is in fact the basic idea of applying dimensionality reduction approaches.

The term "dimensionality reduction" reflects another aspect of these approaches: Often only a very small number of processes prevail in a given data set. Thus, a small number of "components" suffices to explain a large fraction of the variance of a data set with many variables, that is, a high-dimensional data set. Thus application of dimensionality reduction techniques facilitates interpretation and understanding the basic features of large, high-dimensional data sets.

A well known approach is the Principal Component Analysis (PCA). In a mathematical sense it performs an Eigen value decomposition of the correlation matrix of the measured variables. It is the most efficient approach as long as the data are Gaussian distributed and only linear relationships have to be accounted for. It has often been applied to large multivariate data sets of water quality (Haag and Westrich 2002; Thyne et al. 2004; Fernandes et al. 2006) as well as in soil science and geochemistry (Gupta et al. 2006; Zhang et al. 2007; Weyer et al. 2008; Langer and Rinklebe 2009).

However, the PCA is neither well suited for non-Gaussian distributed data nor for data sets that exhibit pronounced non-linear correlations. To that end a variety of non-linear methods of dimensionality reduction approaches have been developed (Lee and Verleysen 2007). Unlike for the linear case, there is no clear advice when which approach would be superior to others, and a trial-and-error procedure is suggested (Lee and Verleysen 2007). One of these approaches is presented here. The Isometric Feature Mapping (Isomap) has been introduced by Tenenbaum et al. (2000). In a certain sense it is a non-linear extension of the linear PCA approach (Gámez et al. 2004). The theory of that approach is not presented here. To that purpose, the reader is referred to Tenenbaum et al. (2000), Gámez et al. (2004), Mahecha et al. (2007) or Lee and Verleysen (2007).

Here the potential of that approach for soil and water quality monitoring is illustrated. Compared to the linear PCA, the first two Isomap components explain a substantially larger fraction of variance of the data set (82 % compared to 70 %) (Fig. 1). This provides some evidence that non-linear relationship play a role in the given data set. Including the third component yields about 90 % of the variance both for Isomap as for PCA. Additional components increase that fraction only marginally. It can thus be concluded that only three different processes play an important role for the given data set. In the following only the Isomap results will be considered.

The next step aims at identifying these three processes. The Isomap analysis ascribes a value for each of the eight components to each of the 718 samples of the data set. These component values, called "scores", can be considered a measure for the strength of the effect of the respective (yet unknown) process. The relationship of the measured variables with the Isomap scores reveals the "fingerprint"
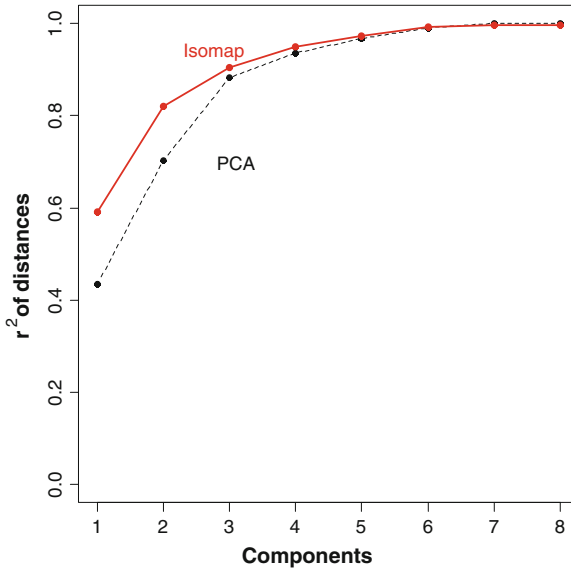
Fig. 1 Cumulative variance explained by eight principal components (*PCA*) and isometric feature mapping (*Isomap*) components

of the respective process. This relationship might be blurred due to the fact that other processes have an impact on the measured variables as well. Those effects can be subtracted from the measured data as follows: A multivariate linear regression of the measured values with all but one Isomap component gives the fraction of the variance of the measured data that is explained by these components. Then the residuals of that regression are compared with the scores of the remaining component.

Figure 2 gives the results for the first component which depicts 59 % of the variance of the data set. It is positively correlated with SO_4 and Cl concentration, pH value and electric conductivity. Negative and partly non-linear correlations are observed for NH_4, NO_3 and P concentration and redox potential. Figure 3 shows that component scores are close to zero at most sites. Only four sites exhibit very large values and substantial temporal variance. These results were unexpected prior to the study. However, both the "fingerprint" as well as the spatial pattern provide strong evidence that upwelling saline groundwater plays a major role at these sites: very high SO_4 and Cl concentration, pH value and electric conductivity are indicative for tertiary deep groundwater that is known to upwell at many sites in East Germany. Low redox potential, and hardly detectable agricultural contaminants like NH_4, NO_3 and P support this hypothesis. Only few piezometers at two different sites were affected. It has been found at other sites that plumes of upwelling saline groundwater can have in fact very small extent.

The second component accounts for 23 % of the data set's variance. The scores of that component increase with NO_3 concentration and redox potential and
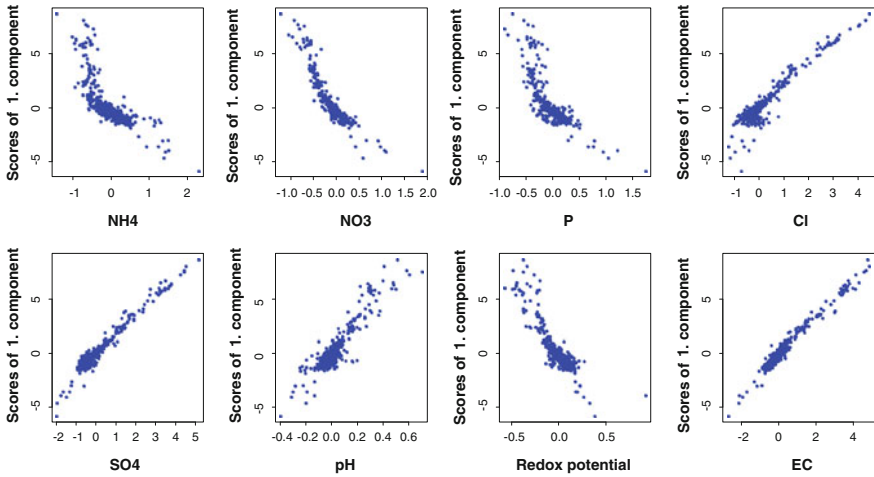
Fig. 2 Scores of the first component versus the residuals of the measured variables
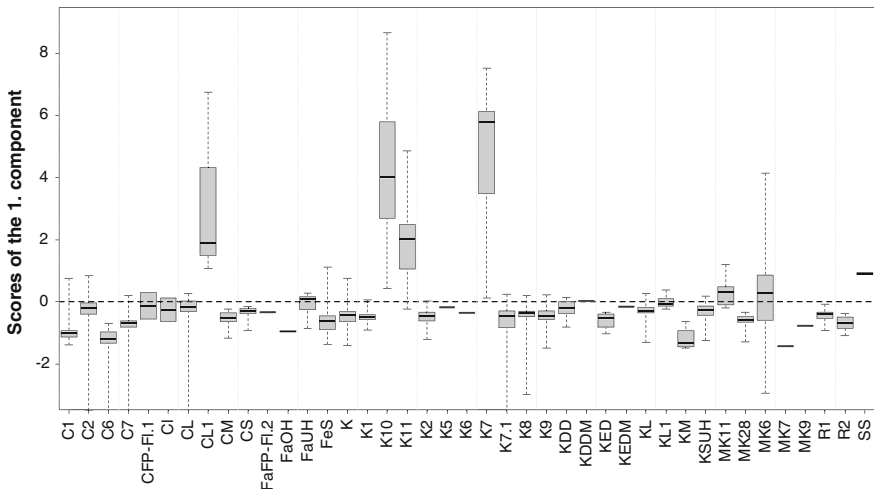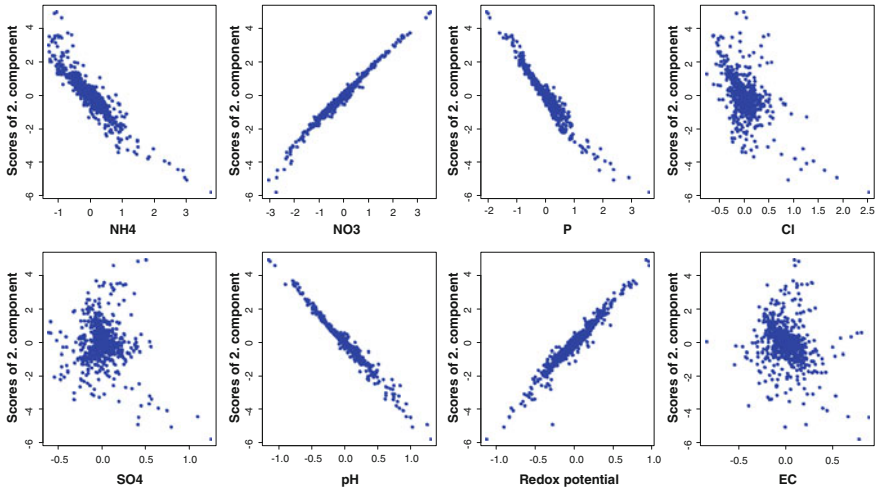

Fig. 3 Scores of the first component for different piezometers. The *boxes* give the 25th, 50th and 75th percentiles, and the *whiskers* the range of the Isomap scores

decrease with NH_4 and P concentration and with pH value. The remaining variables do not show any clear relationship with the second component (Fig. 4). This pattern is consistent with the well known sequence of redox processes. High scores reflect oxic conditions where NO_3 is stable. Being the anion of a strong mineral acid, high NO_3 concentration reduces the pH value. In contrast, low scores of the second component characterise hypoxic conditions where NO_3 is partly denitrified, partly reduced to NH_4, and P is released from sorption to Fe oxides and

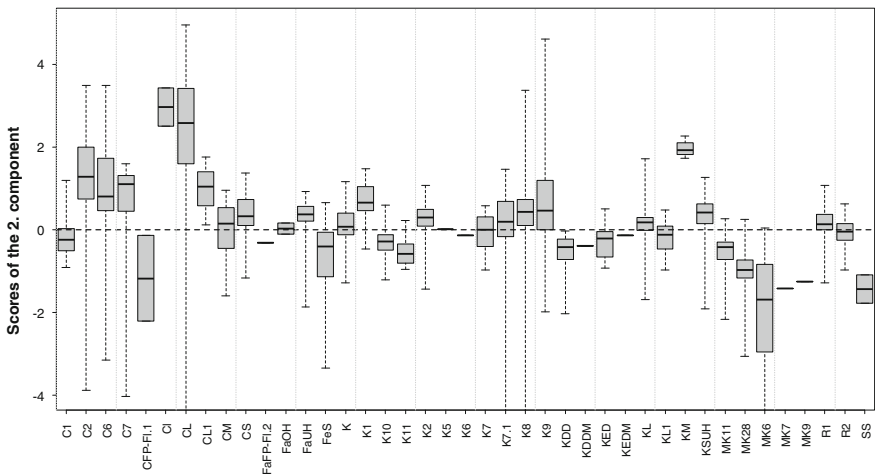**Fig. 4** Scores of the second component versus the residuals of the measured variables

Fig. 5 Scores of the second component for different piezometers. The *boxes* give the 25th, 50th and 75th percentiles, and the *whiskers* the range of the Isomap scores

hydroxides, that become reduced and dissolved under hypoxic conditions. That interpretation is confirmed by the fact that very low scores have been found at the wetland site (MK6, MK28) (Fig. 5).

This component exhibits consistent temporal patterns at many sites. Figure 6 gives time series of the component scores at selected sites. Many sites show very low values during winter 2000/2001, and in the second half of 2002. Precipitation in September 2001 was about four times that of mean monthly precipitation, and
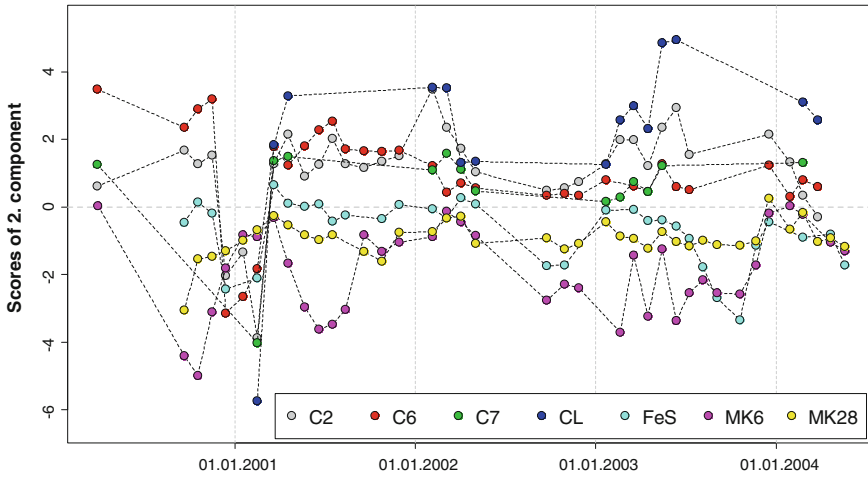
Fig. 6 Time series of scores of the second component for selected sites

continued to exceed the mean until the end of 2002. Thus, soil saturation was higher as usual, which might have promoted oxygen consumption and a decrease of redox potential. It is remarkable that other components exhibited less clear and consistent temporal patterns in this study (not shown).

Another 8.5 % of the variance of the data set is explained by the third component. It is positively correlated with NO_3 and P concentration and pH values, and negatively with NH_4 concentration and redox potential (Fig. 7). Thus this component is determined by the same variables compared to the second
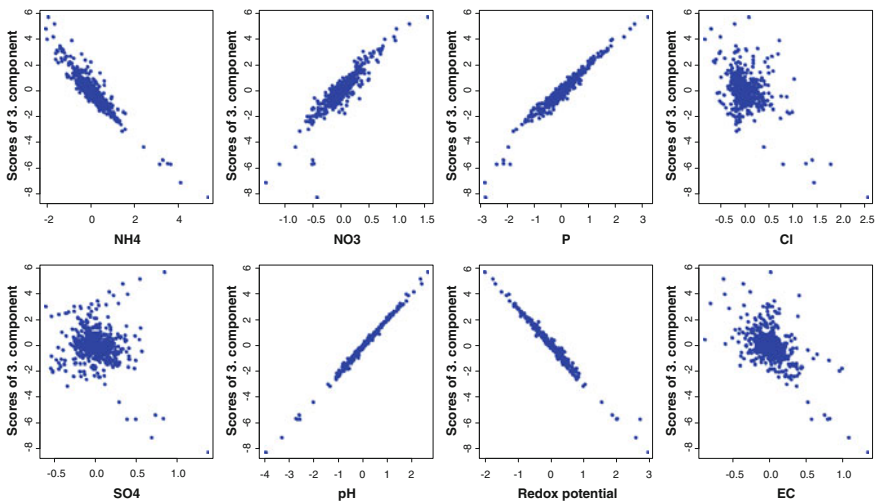

Fig. 7 Scores of the third component versus the residuals of the measured variables
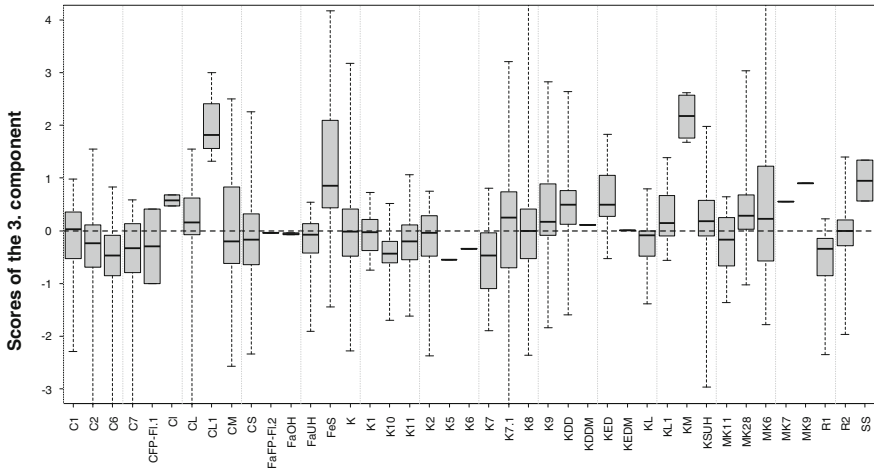
Fig. 8 Scores of the third component for different suction cups. The *boxes* give the 25th, 50th and 75th percentiles, and the *whiskers* the range of the Isomap scores

component, although partly with inverted sign. The third component explains some scatter in the data that has not already depicted by the second component. As will be shown later (Fig. 10), this component mainly differentiates between samples with low scores of the second component, that is, reduced groundwater. Samples with high scores of the third component exhibit higher NO_3 and P concentration and pH value, and lower NH_4 concentration and redox potential than would be expected based on the scores of the second component only. This might be either due to kinetic constraints, or to site-specific peculiarities. In fact, temporarily high scores of the third component have been observed at three sites only (Fig. 8).

Statistics cannot replace the expert's knowledge about biogeochemical processes. The interpretation of the Isomap results might be matter of debate, or might need to be refined and proved by additional measurements. E.g., nitrate isotope data can be used as an independent assessment of denitrification processes. But the Isomap analysis does reveal relationships within the given data set that both challenge and deepen our understanding of the prevailing processes.

4 Visualization

Visualization means more than generating fancy graphs for presentations or papers. Visualization makes use of the most powerful interface between the data stored in a computer and the human brain. The latter is very good in extracting relevant information from large data sets. Within less than a second information from 250 million sensory cells in the man's eyes, that is, a 250 million dimensional
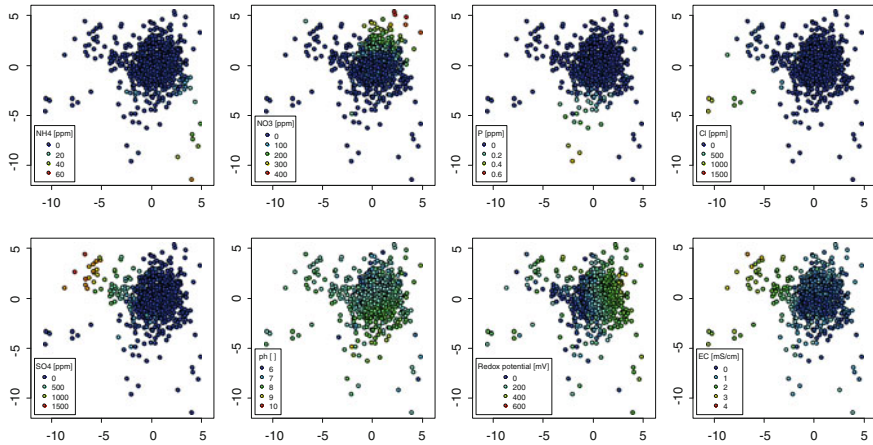
Fig. 9 Measured values of the samples, represented by the SOM-SM

data set, is reduced to a handful of relevant objects. Without this skill man would not have been able to escape predators, enemies, or other road users, and to hunt prey or to find food.

Efficient visualization can help a lot in identifying clusters and outliers, in analysing multivariate trends, or to detect any peculiarities in large multivariate data sets. Here the SOM-SM approach is presented and is applied to the same data set as the Isometric Feature Mapping. It is a combination of Sammon's Mapping (SM; Sammon 1969) and Self-Organizing Maps (SOM), which have been termed Kohonen Feature Mapping as well (Kohonen 2001). Again, the reader is referred to the literature (Sammon 1969; Kohonen 2001; Lischeid 2009) for further details. Here only the application to environmental monitoring data is demonstrated.

The SOM-SM approach yields a two-dimensional graph where every symbol denotes a single sample of the data set. The same graph is used in various ways, where colours indicate different features of the data (Fig. 9). In contrast, the location of the symbols remains the same. The symbols have been allocated in an iterative procedure to ensure that the distance between any two symbols in the graph is proportional to the similarity of the respective two samples with respect to all measured variables. In this study, the two-dimensional graph depicts 89 % of the variance of the data set with eight measured variables. Please note that the scaling at the axes is given for orientation only but has nothing to do with measured data, location of the sampling site in the Quillow catchment, or alike. It is an entirely arbitrary configuration, except for the above stated correlation between distance and similarity.

In a certain sense, any of the graphs given in Fig. 9 can be interpreted analogously to a topographical map where the colour indicates the altitude, with numerous mountain tops and valleys. Similar to a topographic map, a rather smooth "landscape" develops where neighbour points exhibit roughly the same

values. However, in these graphs measured values for the samples are given rather than topographic height. Moreover, different "maps" are shown for different measured variables but for the same symbols, that is, the samples of the data set (Fig. 9).

Ammonium concentration is very low and close to the detection level for most samples (Fig. 9). Only samples that plot in the lower right corner exhibit higher NH_4 concentration values. In contrast, NO_3 concentration of these samples is very low, and the highest NO_3 concentration is found in the upper right corner of the graph. Samples with enhanced P concentration plot close to those of high NH_4 concentration, but are not identical. Similarly, Cl and SO_4 concentration and electric conductivity exhibit similar patterns in the SOM-SM with the highest values on the left side (Fig. 9). The pattern is less clear for pH values. There is a slight tendency of increasing pH values from the top to the bottom of the graph. For the redox potential values increase from the left to the right.

These patterns reflect the correlation structure of the data set: Variables with similar patterns, that is, same directions of steepest increase of measured values, are positively correlated, like Cl, SO_4 and EC (cf. Table 1). In contrast, the directions of steepest increase of P and Cl are approximately perpendicular to each other, that is, the correlation is close to zero (Fig. 9, cf. Table 1). The gradients for P and NO_3 are inverse to each other, which is symptomatic for a negative correlation.

Similar to the measured variables the Isomap scores can be represented in the SOM-SM graph (Fig. 10). There is a more or less monotonic increase from the bottom right to the upper left corner for the first component, which corresponds roughly to that of SO_4, Cl and EC which in fact are closely related to this component (Fig. 2). In contrast, the scores of the second component increase from the bottom to the upper right for the second component, which is parallel or inverse to those of the redox-dependent variables (Fig. 4). Please note that the Isomap scores have not been used for generating the SOM-SM. Rather, organisation of the samples within the SOM-SM makes use of the internal relationships of the data set, similar to that of the Isomap analysis.

The SOM-SM graph can be used for analysing spatial and temporal patterns as well. In Fig. 11 samples from single sites are colour-coded. Although there is some
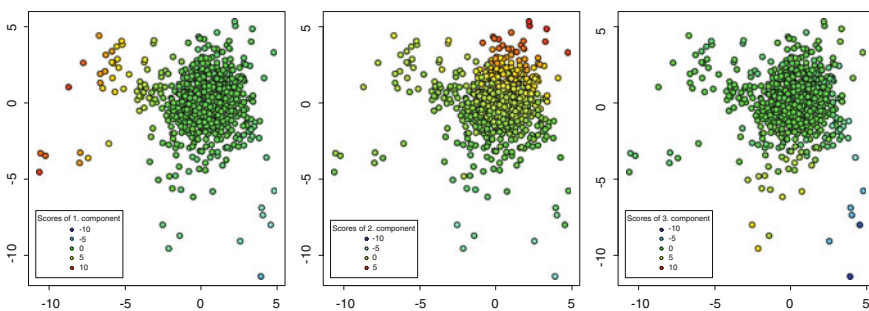

Fig. 10 Isomap scores of the samples, represented by the SOM-SM

overlap, samples of different sites usually cluster in different sub-regions of the graph. E.g., the C6 samples are located in the upper part of the graph, those of site FeS more to the lower right, and those of K10 in the upper left corner. These patterns can be related to those of the measured variables (Fig. 8) or of the Isomap scores (Fig. 9). Moreover, the temporal variance can be investigated for single sites separately. For example, the samples from the MK6 site spread over a larger range compared to those of C6 (Fig. 11). In addition, the MK6 samples scatter more in a direction parallel to the gradient of the second Isomap component rather to that of the first component (cf. Fig. 9). Thus, for this site changes of the redox conditions play a larger role compared to the process which is represented by the first component.

At the K8 and at the C6 site some samples plot far away from the centre of the respective clusters. These "outliers" should be checked for possible erroneous assignment to the sampling site. As an alternative, these deviations from the centre of the respective clusters could be due to extreme conditions during or immediately prior to sampling. This is a common phenomenon especially for stream and river water samples taken during major floods, where deviations from the respective clusters occur roughly in parallel at different sites.

Colour coding in Fig. 11 indicates the year of sampling. This allows investigating trends in the SOM-SM graph, that is, systematic shifts over time. Please note that the SOM-SM represents information about all measured variables. Thus any clear shift in the graph usually is due to a (possibly non-linear) trend of more than one variable. Figure 11 indicates trends only at some of the sites, and none of these is linear. At K8 and K9, the first samples plot close to the right border, those of subsequent years more to the upper left, and those of 2004 more close to the
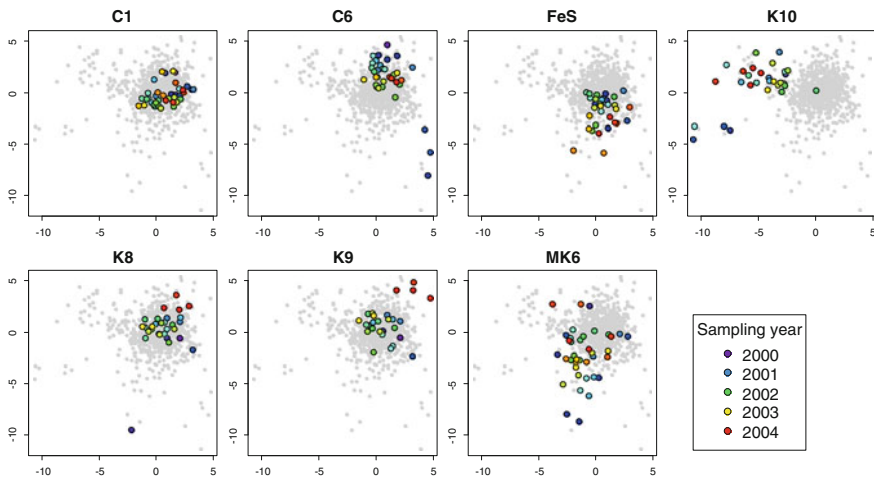

Fig. 11 Time series of soil solution quality at selected sites, represented by the SOM-SM. Samples of the respective site as indicated in the title are *coloured* according to sampling year, all remaining samples are given in *grey*

upper right corner. Thus, this non-linear trend firstly follows an increase of the scores of the first component, and then an increase of those of the second component (cf. Fig. 9), indicating less reduced conditions. In contrast, samples at C6 follow more a spiral from the lower right in anti-clockwise direction and with decreasing radius towards 2004. Correspondingly, the data could be tested for seasonal patterns etc. (not shown).

5 Conclusions

Managing water and soil resources requires sound data that can only be provided by systematic monitoring (Muller 2012). Advanced statistical approaches have been developed to make more efficient use of large data sets in spite of substantial spatial and temporal heterogeneities that are immanent to natural systems. Here only two out of a vast zoo of different approaches have been presented. Natural water resources management as well as science could and should benefit a lot from a variety of approaches that have been developed in other disciplines.

Dimensionality reduction techniques have been applied to time series of groundwater head, lake and river water levels, and stream discharge (Longuevergne et al. 2007; Lewandowski et al. 2009; Lischeid et al. 2010; Lischeid and Kalettka 2012; Thomas et al. 2012). They were successful in differentiating between different effects on the observed dynamics at minor data requirements. The author is firmly convinced that much more can be learned from data by advanced data analysis techniques and will be happy to assist in any related efforts.

References

Fernandes PG, Carreira P, da Silva MO (2006) Identification of anthropogenic features through application of principal component analysis to hydrochemical data from the Sines coastal aquifer, SW Portugal. Math Geol 38:765–780

Gámez AJ, Zhou CS, Timmermann A, Kurths J (2004) Nonlinear dimensionality reduction in climate data. Non-linear Process Geophys 11:393–398

Gupta AK, Sinha S, Basant A, Singh KP (2006) Multivariate analysis of selected metals in agricultural soil receiving UASB treated tannery effluent at Jajmau, Kanpur (India). Bull Environ Contam Toxicol 79:577–582

Haag I, Westrich B (2002) Processes governing river water quality identified by principal component analysis. Hydrol Process 16:3113–3130

Kohonen T (2001) Self-organizing maps. Springer Series in Information Sciences, vol 30, 3rd edn. Springer, Berlin

Langer U, Rinklebe J (2009) Lipid biomarkers for assessment of microbial communities in floodplain soils of the Elbe River (Germany). Wetlands 29:353–362

Lee JA, Verleysen M (2007) Nonlinear dimensionality reduction. Information science and statistics. Springer, Berlin

Lewandowski J, Lischeid G, Nützmann G (2009) Drivers of water level fluctuations and hydrological exchange between groundwater and surface water at the lowland River Spree (Germany): field study and statistical analyses. Hydrol Process 23:2117–2128. doi:10.1002/hyp.7277

Lischeid G (2009) Non-linear visualization and analysis of large water quality data sets: a model-free basis for efficient monitoring and risk assessment. Stoch Env Res Risk Assess 23:977–990. doi:10.1007/s00477-008-0266-y

Lischeid G, Bittersohl J (2008) Tracing biogeochemical processes in stream water and groundwater using nonlinear statistics. J Hydrol 357:11–28. doi:10.1016/j.jhydrol.2008.03.013

Lischeid G, Kalettka T (2012) Grasping the heterogeneity of kettle hole water quality in Northeast Germany. Hydrobiologia 689(1):63–77. doi:10.1007/s10750-011-0764-7

Lischeid G, Natkhin M, Steidl J, Dietrich O, Dannowski R, Merz C (2010) Assessing coupling between lakes and layered aquifers in a complex Pleistocene landscape based on water level dynamics. Adv Water Resour 33:1331–1339. doi:10.1016/j.advwatres.2010.08.002

Longuevergne L, Florsch N, Elsass P (2007) Extracting coherent regional information from local measurements with Karhunen-Loève transform: case study of an alluvial aquifer (Rhine valley, France and Germany). Water Resour Res 43:W04430. doi:10.1029/2006WR005000

Mahecha M, Martínez A, Lischeid G, Beck E (2007) Nonlinear dimensionality reduction as a new ordination approach for extracting and visualizing biodiversity patterns in tropical montane forest vegetation data. Ecol Inf 2:138–149. doi:10.1016/j.ecoinf.2007.05.002

Muller M (2012) From raw data to informed decisions. In: WWAP (World Water Assessment Programme): The United Nations World water development report 4: Managing water under uncertainty and risk. Paris, UNESCO, Chap. 6, pp 157–173

Sammon JW (1969) A nonlinear mapping for data structure analysis. IEEE Trans Comput C-18/5:401–409

Schilli C, Lischeid G, Rinklebe J (2010) What processes prevail? Analyzing long-term soil-solution monitoring data using nonlinear statistics. Geoderma 158:412–420. doi:10.1016/j.geoderma.2010.06.014

R Development Core Team (2006) R: A language and environment for statistical computing. R Foundation for statistical computing. Vienna, Austria. ISBN: 3-900051-07-0, http://www.Rproject.org

Tenenbaum JB, de Silva V, Langford JC (2000) A global geometric framework for non-linear dimensionality reduction. Science 290:2319–2323

Thomas B, Lischeid G, Steidl J, Dannowski R (2012) Regional catchment classification with respect to low flow risk in a Pleistocene landscape. J Hydrol 475:392–402. doi:10.1016/j.jhydrol.2012.10.020

Thyne G, Guler C, Poeter E (2004) Sequential analysis of hydrochemical data for watershed characterization. Ground Water 42:711–723

Weyer C, Lischeid G, Aquilina L, Pierson-Wickmann A-C, Martin C (2008) Mineralogical sources of the buffer capacity in a granite catchment determined by strontium isotopes. Appl Geochem 23:2888–2905

Zhang HB, Luo YM, Wong MH, Zhao QG, Zhang GL (2007) Concentrations and possible sources of polychlorinated biphenyls in the soils of Hong Kong. Geoderma 138:244–251