Soil Use and Management

Soil Use and Management, March 2015, **31,** 121–131 doi: 10.1111/sum.12169

Exploring soil databases: a self-organizing map approach

D. Rivera[1], M. Sandoval[2] & A. Godoy[3]

[1]*Department of Water Resources, Laboratory of Comparative Policies in Water Resources, University of Concepcion, CONICYT/FONDAP-15130015, Vicente Mendez 595, Chillán, Chile,* [2]*Department of Soil and Natural Resources, University of Concepcion, Vicente Mendez 595, Chillán, Chile, and* [3]*Facultad de Ingeniería, Universidad del Desarrollo, Av. Plaza 700, San Carlos de Apoquindo, Las Condes, Chile*

Abstract

A soil quality database (SQDB) is a collection of soil samples described by a given set of parameters, allowing farmers, scientists and other stakeholders to make informed decisions about practices, processes and policies for soil use and management. If each parameter is considered as a dimension of the space spanned by the SQDB, extracting information becomes a difficult task when the number of parameters is >3. A widely used approach to explore multidimensional data sets is the self-organizing map (SOM) method, which is suitable for clustering, visualization and extraction of information from multidimensional data. We applied the SOM method as an exploratory technique to an unlabelled SQDB to extract knowledge – data patterns and data associations – from the data set (the time and location of each sample were unknown). The SQDB used in this study is a set of 1240 unlabelled samples within the Central Valley of Chile, covering ca 7500 km^2. The predominant soils are Andisols with a large organic matter content (7–12%), small bulk densities (0.6–1.0 g/cm^3) and large water-holding capacity. We identified three patterns: (i) isolated region within the map with close neurons (smooth transitions), (ii) two or more regions with predominantly large or small values and (iii) homogeneous map with small values with an isolated region of large values. These patterns show that the data set represented more than two groups that were not necessarily related. For pH, no important associations with other investigated parameters were observed. Previous studies carried out by the local agricultural research station showed that pH values below 5.5 constrain nutrient uptake. Thus, locations presenting pH<5,5 should be subject to seasonal monitoring to assess management practices that mitigate soil acidity. The component plane for organic matter indicates that ca. 50% of the soil samples had contents <8% related to soil series characteristics and management practices. As the *k*-means is initialized by random partitions, the two-step approach (clustering the map representing the input data) is less sensitive to variations in the input data (subsamples) than is the direct application of *k*-means to the input data, but it also reduces the computational cost. The ability of SOMs to visualize multidimensional data sets helps gain an understanding of the data in the exploratory phase, such as the association and integration of physical, chemical and biological parameters.

Keywords: soil quality databases, self-organizing map, data mining

Introduction

Soil monitoring allows stakeholders (e.g. farmers, scientists) and decisions-makers to make informed evaluations about practices, processes and policies for soil use and management. Each soil sample can be described by a set of parameters, and to characterize a given location in time or space, a certain number of samples need to be taken. Soil laboratories store results of soil samples coming from different times, fields, soils and clients, generating soil quality databases (SQDB).

Soil quality is a multidimensional concept (Villamil *et al.*, 2008), considering physical, chemical and biological parameters. Therefore, it is not possible to define the whole (physical, biological, chemical) condition of the soil by a single parameter, but rather by a set of parameters, each one describing an important characteristic of the soil. When considering soil samples – multiple parameters for a given time and location – as vectors, a SQDB can be defined as a

Correspondence: D. Rivera. E-mail: dirivera@udec.cl
Received May 2014; accepted after revision December 2014

matrix $\mathbf{Q}_{s\times p}$ with p columns (parameters) and s rows (samples), which spans a p-dimensional space (Webster, 2001). There are some issues about how to extract and visualize the information contained in $\mathbf{Q}$. For $P \leq 3$, it is sufficient to analyse scatter plots and 'checklists' for each parameter. However, for $P > 3$, a dimensional reduction method is needed, such as the widely used principal component analysis (PCA) (Webster, 2001; Villamil *et al.*, 2008).

Data mining seeks to answer questions or solve problems based on available data, following, in an iterative way, four methodological steps (Vesanto, 2002): (i) data exploration, to gain understanding (ii) preprocessing, (iii) model(s) construction and assessment and (iv) knowledge consolidation and deployment. Confirmatory data analysis answers questions such as 'Do the data confirm the hypothesis of the study?' whereas data mining tends to ask 'What can the data tell me about this relationship?' (Martinez & Martinez, 2005). Consequently, soil quality databases are good subjects for data mining – the search of clusters, patterns and structures – that could be used to construct hypotheses or develop knowledge, specially unsupervised learning algorithms such as self-organizing maps (SOM, Kohonen, 2001).

Self-organizing maps perform a nonlinear mapping of the data set onto a two-dimensional grid through an unsupervised learning algorithm, allowing exploration of relationships and patterns between parameters, as well as structures in a single parameter (e.g. approximating the probability distribution of the parameter). Therefore, the SOM method is a suitable technique for clustering, visualization and extraction of information from multi dimensional data (Vesanto, 2002; Penn, 2005; Herbst & Casper, 2008). The data set used for training the map could be tailored to include or combine relevant parameters, for example combining physical parameters with biological indicators (Vesanto, 2002; Astel *et al.*, 2007).

The SOM method has been successfully applied in classification of sediment samples (Alvarez-Guerra *et al.*, 2008, 40 samples, 13 parameters), sediment chemistry (Astel *et al.*, 2007), soil classification using data from near-infrared spectroscopy (Fidencio *et al.*, 2001), soil group classification (Tissari *et al.*, 2007), sole-carbon source utilization profiles (Leflaive *et al.*, 2005), soil biological and chemical quality (Mele & Crowley, 2008; Dhar & Cherkassky, 2011), soil bacterial community comparisons (Wu *et al.*, 2008), water quality and hydrological records (Kalteh *et al.*, 2008) and ecological sciences (Chon, 2011). Astel *et al.* (2007) pointed out that the SOM method has a stronger 'resolving power' for classification than PCA in analysing large data sets (ca. 15 000 samples, 14 parameters and 23 sampling locations), combining chemical indicators such as pH, dissolved oxygen, biological oxygen demand, chloride and nitrate, among others. Merdun (2011) explored a soil database (19 variables) to analyse relationships among soil, chemical and hydraulic soil properties. Our approach is similar to Merdun (2011) and Astel *et al.* (2007) on the methodological aspects, but differs in the specific methods used to cluster the maps and the assessment of the quality of the clustering process.

Taking advantage of the visualization and pattern discovery features of self-organizing maps, we applied this technique to a soil quality database composed of 1200 samples and 17 variables. The aim of this research was twofold: first, to apply SOM as an exploratory technique to an unlabelled SQDB (time and location for each sample are unknown) and to extract knowledge from the data set, that looks for associations, patterns and groups, and second, to bring attention to some specific issues regarding the application of SOM, such as clustering the map. This exploratory phase aims to get insight from the data rather than construct a rigid, unique model of the data ('making sense of the data'; Vesanto, 2002). Discussion regarding specific practices to improve soil quality or the level of certain parameters is outside the scope of this work.

Data and methods

Soil data set. The SQDB used in this study is a set of 1240 unlabelled samples from the valley and piedmont (*precordillera*) within the Central Valley of Chile, covering an approximate area of 7500 km^2 (Figure 1). The predominant soils are Andisols – soils derived from volcanic ash – Typic Melanoxerand, Humic haploxerand and Typic haploxerand, having large values of organic matter (7–12%), small bulk densities (0.6–1.0 g/cm^3), large water-holding capacity and phosphate fixation (Zagal *et al.*, 2002; Sandoval *et al.*, 2007, 2008).

Analyses were carried out for pH (−), organic matter (OM) (%), nitrate nitrogen (NO_3-N) (mg/kg), manganese (Mn) (mg/kg), phosphorus (P) (mg/kg), potassium (K) (mg/kg), calcium (Ca) (cmol/kg), zinc (Zn) (mg/kg), magnesium (Mg) (cmol/kg), sodium (Na) (cmol/kg), bases (cmol/kg), copper (Cu) (mg/kg), exchangeable aluminium (Ex Al) (cmol/kg), potassium saturation (K Sat) (%), calcium saturation (Ca Sat) (%), boron (B) (mg/kg), magnesium saturation (Mg Sat) (%), sulphate–sulphur (SO_4^--S) (mg/kg) and iron (Fe) (mg/kg). Samples were taken between the years 2000 and 2008 from the topsoil. Parameters were part of periodical surveys for soil fertility status at farm scale carried out by a fertilizers manufacturing company. Under a confidentiality agreement, it is not possible to disclose the exact location of sampling points.

Self-organizing map: how does it work?. Self-organizing maps (SOM) are a set of neurons (the basic working units) forming a two-dimensional array (Figure 2a). The SOM is composed of two fully connected layers: the input layer and

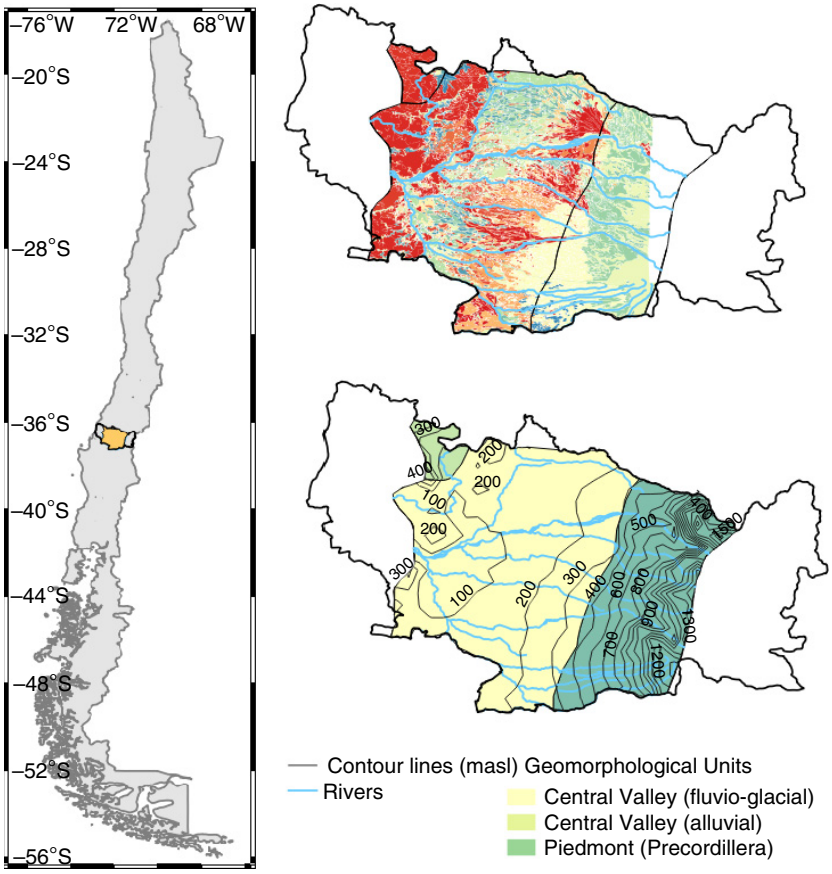

Figure 1 Map of the study site. The upper map shows the variability of soil series, and the bottom map shows the geomorphological units.

the output layer. Each neuron from the output layer has a double representation: a reference vector that has as many components as the number of input variables and its position in the grid. The set of reference vectors is called a codebook. A SOM applies an unsupervised learning algorithm that preserves data topology, that is data are organized on the grid in such a way that observations that are close together in the high-dimensional space have close positions in the map (Kaski, 1997; Kohonen, 2001; Vesanto, 2002; Herbst & Casper, 2008). Therefore, the codebook approximates the space spanned by the original data set (input data or input matrix) as well as its probability distribution (Park *et al.*, 2003; Herbst & Casper, 2008). The map – the projection of the codebook on a grid – allows identifying groups of observations with similar characteristics (clusters) by taking into analysis all parameters simultaneously.

The training algorithm considers an input matrix $\mathbf{T}_{s\times p}$, formed by s samples (rows) and p parameters (columns). The input matrix could be the SQDB itself or a subsample of the SQDB by removing samples and/or parameters. The codebook $\mathbf{P}_{N\times p}$ has N neurons where each row is the reference vector of the i-th neuron. To train the map, a sample is randomly taken from $\mathbf{T}_{s\times p}$. Then, the best-matching unit (BMU) is determined, where the BMU is the neuron whose Euclidean distance between its reference vector and the sample is minimal (the closest location on the p-dimensional space). After, the codebook is updated by moving the BMU and its neighbours to a location closer to the input vector according a neighbourhood function. The algorithm is applied to all samples and repeated until a stop-rule is achieved. In a practical sense, a SOM is a partitioning algorithm with an adaptive behaviour (Murtagh & Hernández-Pajares, 1995).

There are two canonical tools to visualize the SOM. The unified distance matrix (**U**-matrix) calculates the distances of reference vectors to each of its immediate neighbours in the grid, as well as the median distance. It displays the distance structures, using a colour scale in a two-dimensional array of neurons, maintaining the topology and allowing identification of the clusters, boundaries and representative neurons (Ultsch & Siemon, 1990; Vesanto, 2002; : Peeters *et al.*, 2007). Figure 3 shows different options to visualize data using SOM. We used a synthetic data set composed of four cluster and 10 variables (863 samples). Clusters are seen as those map units that have smaller distances (cold colours), and borders between clusters have larger distances (warm colours) (Martinez & Martinez, 2005). The distance matrix is

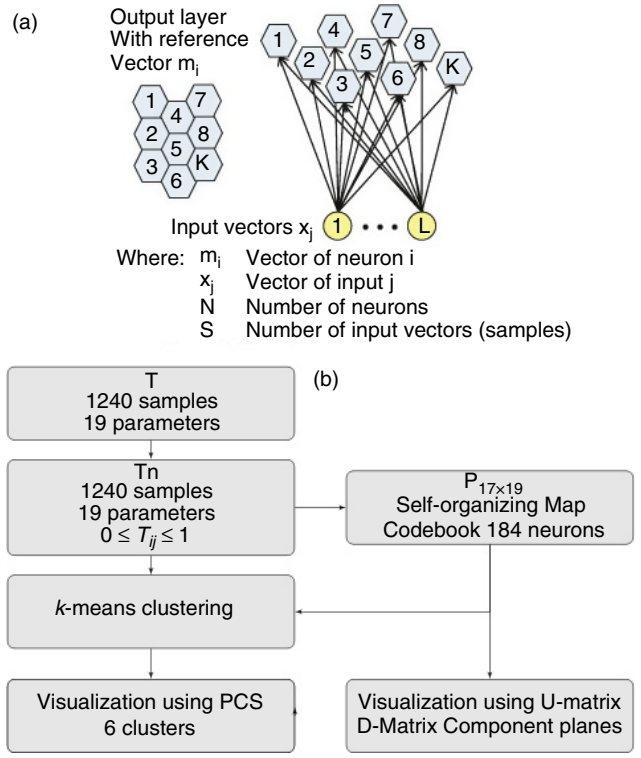

Figure 2 (a) Workflow and (b) structure of a self-organizing map (from Wallner *et al*., 2013. Used with permission of Elsevier).

based on the **U**-matrix, having the same number of neurons of the map by only retaining median distances. For the synthetic data set, the **U**-matrix and distance matrix (Figures 3a,b) show clear borders and compact cluster (similar colouring within the cluster). Figure 3c shows the clustered map after applying the *k*-means partitioning method. It is worth noting that the clustered map (Figure 3c) displays the same pattern as the Distance matrix (Figure 3b). Thus, the visualization of distance-based matrices is a useful tool for visual inspection of data groups. The second approach is the component planes (Figure 3d). They represent the distribution of a parameter in the input data set and the contribution of parameters to cluster structures of the trained SOM (Park *et al*., 2003). By visual inspection, parameters with similar distribution (association or correlation) can be identified (Vesanto, 2002). For the synthetic data set, there is no relationship among the last three variables.

The map size (number of neurons and the ratio between the numbers of neurons in the vertical and horizontal axes) is important to detect differences and patterns in the input data set. In the case of too few neurons (compressing the input data), patterns or structures are hidden and too many neurons (extrapolating the input data) do not add information, despite displaying smoother maps.

Procedure. This work focuses on the exploratory phase: analysis of the nature, reliability and complexity of the data set, and extracting knowledge. A general sketch of the workflow applied in this work is shown in Figure 2b.

We used the SOM Toolbox 2.0 for MATLAB. Using the parameter values of each of the 1240 samples, the size of the input matrix is $\mathbf{T}_{1240\times17}$. Given that the training algorithm uses Euclidean distances, data must be normalized before training to avoid distortion in the results. We normalized each column of **T** in [0,1] using the linear transformation:

$$x' = \frac{x_i - \min(\mathbf{x})}{\max(\mathbf{x}) - \min(\mathbf{x})} \tag{1}$$

where x' is the new normalized value, x_i is the original variable value, min(**x**) is the minimum value of the original variable – a column of $\mathbf{T}_{s\times p}$, and max(**x**) is the maximum original value. This procedure was applied to each column of **T**.

The number of neurons was set using Vesanto's rule (Vesanto, 2002), which define the optimal number of neurons as $5\sqrt{s}$, where s is the number of samples as described elsewhere (Vesanto, 2002; Céréghino & Park, 2009).

To assist in the identification of clusters, we applied two partitioning methods: *k*-means and fuzzy *k*-means clustering, both of which have been described elsewhere in the literature. For further details on clustering the SOM, please refer to Vesanto & Alhoniemi (2000) and Wu & Chow (2004). The *k*-means algorithm is a nonhierarchical clustering technique that optimizes some criterion to partition the observations into a specified or predetermined number of groups (Martinez & Martinez, 2005). The *k*-means method assigns each data point to the cluster where the distance between the data points to the cluster centroid is smallest, that is it minimizes the variance within each cluster. Fuzzy *k*-means assign a membership value of each sample to all clusters. Thus, each sample partially belongs to any cluster, to some degree, depending on the membership value.

For *k*-means and fuzzy *k*-means, the optimal number of cluster must be defined using some efficiency criteria, such as the Davies–Bouldin (DB) index (Davies and Bouldin, 1979):

$$\mathrm{DB} = \frac{1}{n}\sum_{i=1,i\neq j}^{n} \max\left(\frac{S_i + S_j}{d(C_i, C_j)}\right) \tag{2}$$

where n is the number of clusters, S_i and S_j are the average distances of all points in clusters i and j to their cluster centres (within cluster scatter), and $\mathrm{d}(C_i,C_j)$ is the distance between cluster centres C_i and C_j. Small values of this index correspond to clusters that are compact, that is low variance within cluster, and whose centres are far away from each other (Park *et al*., 2003; Razavi & Coulibaly, 2013) (Figure 4).

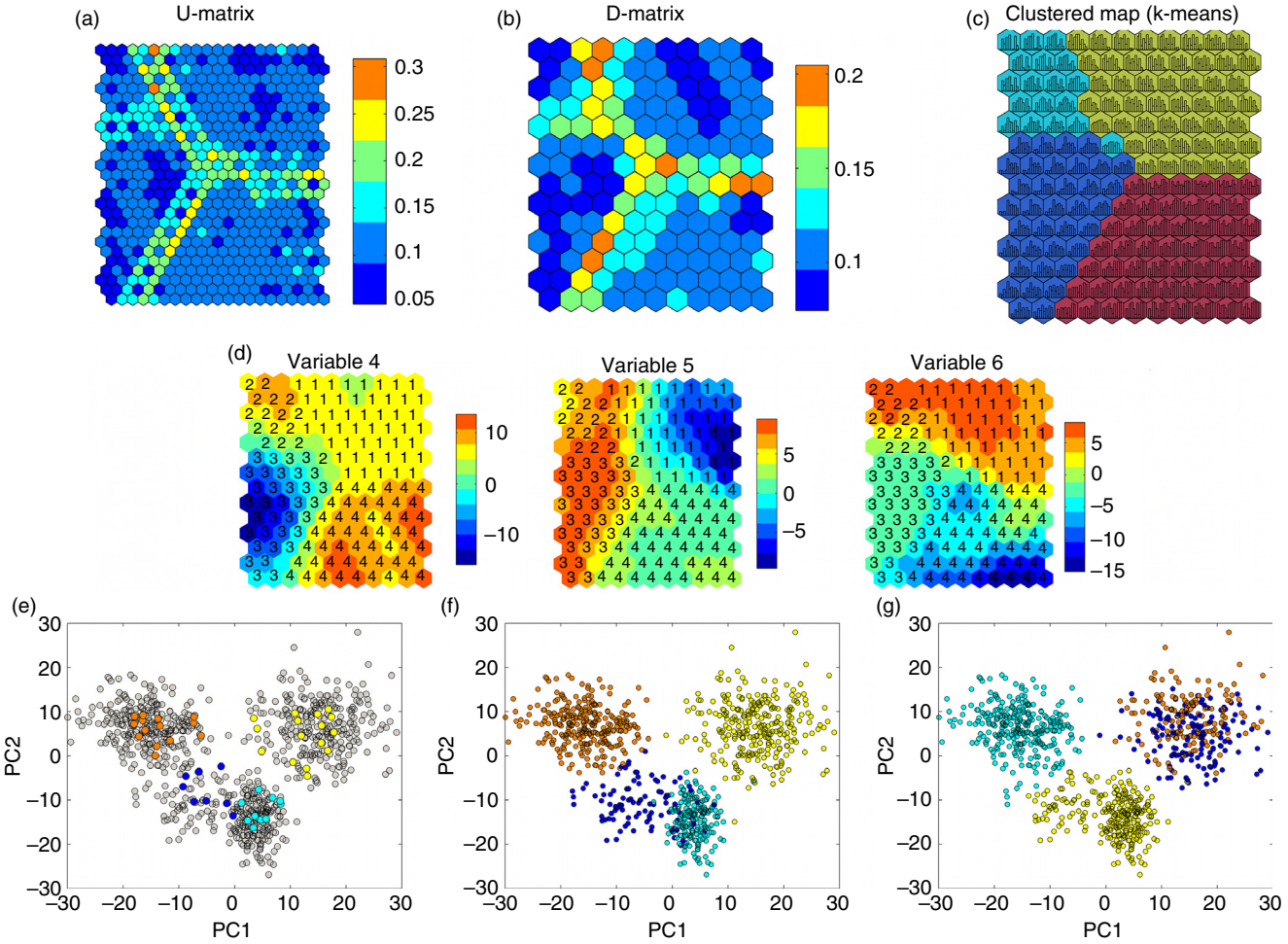

Figure 3 (a) **U**-matrix, (b) **D**-matrix, (c) clustered map and (d) component planes for a synthetic data set of 1200 samples, 10 variables and four clusters. Data set obtained from http://personalpages.manchester.ac.uk/mbs/Julia.Handl/generators.html; (e) projection of the map's neuron in (c) in to the two-first principal components; (f) map-based clustering of the synthetic data set; and (g) *k*-means clustering of the synthetic data set.

We used PCA for visualization tasks (Webster, 2001). The main purpose of PCA is to reduce the dimensionality of a data set (Martinez & Martinez, 2005), while retaining as much of the variation in the original data set as possible. PCA transforms the data to a new set of variables – the principal components – that are a linear combination of the original variables.

To define the PCAs, the original data set of s observations and p parameters is converted to a covariance **C** matrix or correlation matrix **R** following (Webster, 2001):

$$\mathbf{C} = c_{ij} = \frac{1}{n}\mathbf{X}^T\mathbf{X} \tag{3}$$

$$\mathbf{R} = r_{ij} = c_{ij}/\sqrt{c_{ii}c_{jj}} \tag{3}$$

where **X** is the matrix containing the centred data (mean subtracted), and s is the number of samples.

The eigenvectors of **R** (or **C**) form a p-by-p matrix, with each column containing coefficients for one principal component. Thus, the larger they are in absolute value, the stronger is their influence (Webster, 2001). The principal components (eigenvalues) are the variances along the new axes, and they are ordered from largest to smallest with the proportions of the total variance for which they account (Webster, 2001).

The new variates or scores are obtained by multiplying the data set X by the eigenvectors. Scores resulting from the covariance matrix will be dominated by parameters with the largest variances. This is often the case when the variables are of different types or units (Martinez & Martinez, 2005). Thus, if the data are composed of available P with a variance of 26.5 (mg/L) and pH with a variance of 0.25, then the P will swamp pH, hiding the effect of the latter (Webster, 2001). As working with matrix **R** gives them equal weight, the correlation matrix should be used for PCA when the variances along the original dimensions are very different (Webster, 2001; Martinez & Martinez, 2005). Using correlation matrices gives more sensitive analysis and allows

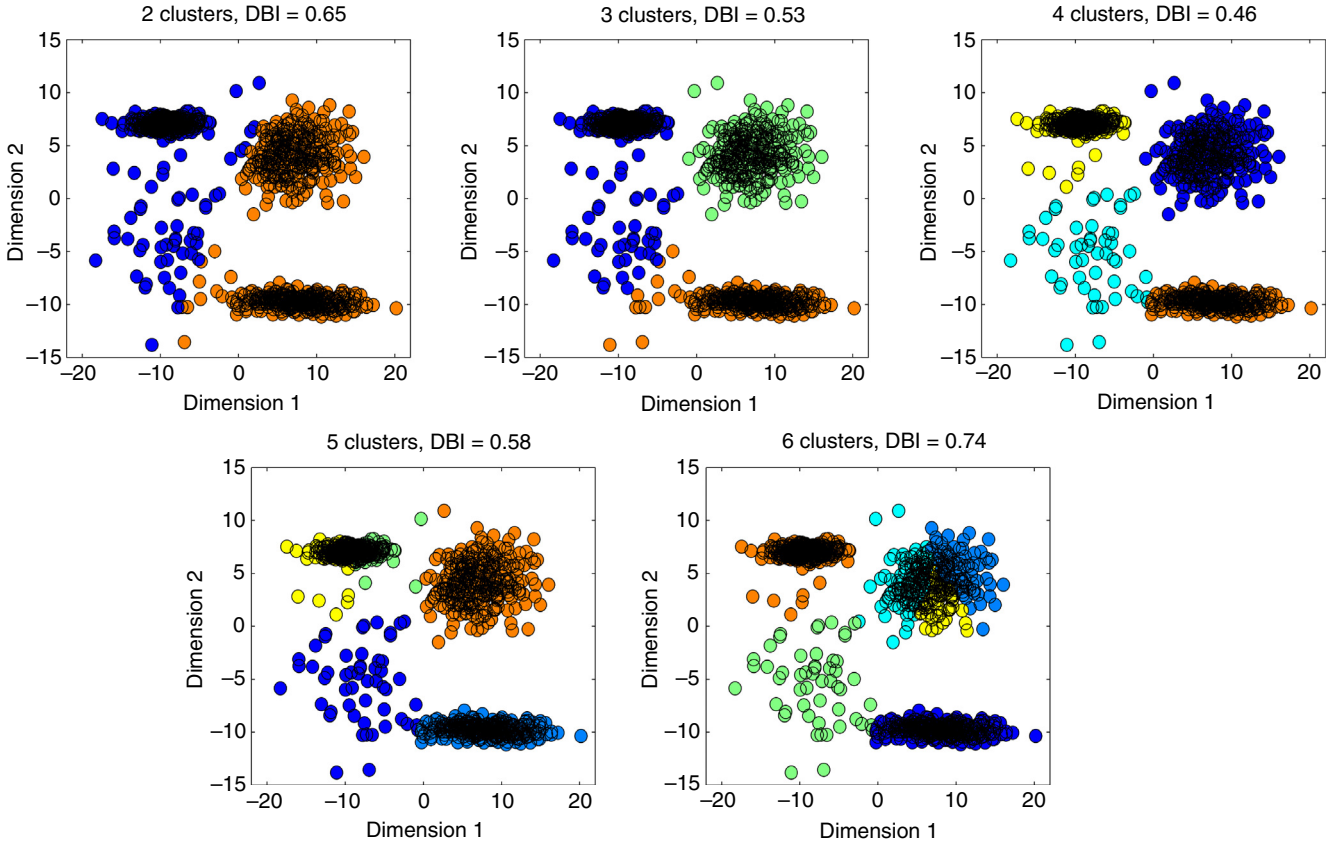

Figure 4 The Davies-Bouldin Index for different number of cluster. Data used are a synthetic set of 800 samples, two variables and four clusters.

comparison of the results of PCA among different analyses. This is the main reason for carrying out normalization procedures to the SQDB.

This rigid rotation of the data to new orthogonal axes allows visualization of highly dimensional data into two- or three-dimensional scatter plots (Webster, 2001). Figures 3e–g shows different projection of samples and neurons onto the first two principal components, showing compact clusters from the synthetic data set.

The number of clusters that minimizes Davis–Bouldin index is taken as the optimal number of clusters, and it allows assessment and comparison between partitioning methods (Figure 4). To better identify patterns in the codebook, we applied the *k*-means algorithm setting $k = 6$ as case study (results for other values of k not shown).

Results and discussion

Exploring the SQDB

Figure 5 shows three tools to visualize high-dimensional data through SOM: **U**-matrix, distance matrix and component planes. In the upper part of the map (**U**-matrix), the neurons are closer to each other (cold colours), but the **U**-matrix does not show clear borders (lines of neurons with warm colours) between potential clusters. Moreover, it shows two regions (middle left and middle right) where neurons are rather separated as warm colours appear. The latter implies that those neurons are associated with samples substantially different than other samples in the data set. The same was observed in the distance matrix because the distance matrix is based on the **U**-matrix. As the number of neurons in the map (11 columns × 17 rows = 187 units) is less than the number of samples, most of the neurons in the map are the BMU (hits) of more than one sample in the database. The 24 upper left-corner neurons account for 25% of the samples. Moreover, across the map, there are neurons with <2 hits, which could be considered as interpolating neurons, smoothing the space spanned by the codebook. Figure 5 shows the clustered map after using *k*-means using a colour code. Thus, highly similar neurons will belong to the same cluster. In this case, the **U**-matrix barely resembles the clusters.

The component planes (lower panels in Figure 5) display the structure of the neurons for each parameter. In each component plane, the clusters are defined by numbering each neuron, making it possible to visualize how each parameter contributes to the grouping. It is possible to identify three patterns: (i) isolated subsets with smooth transitions: pH,

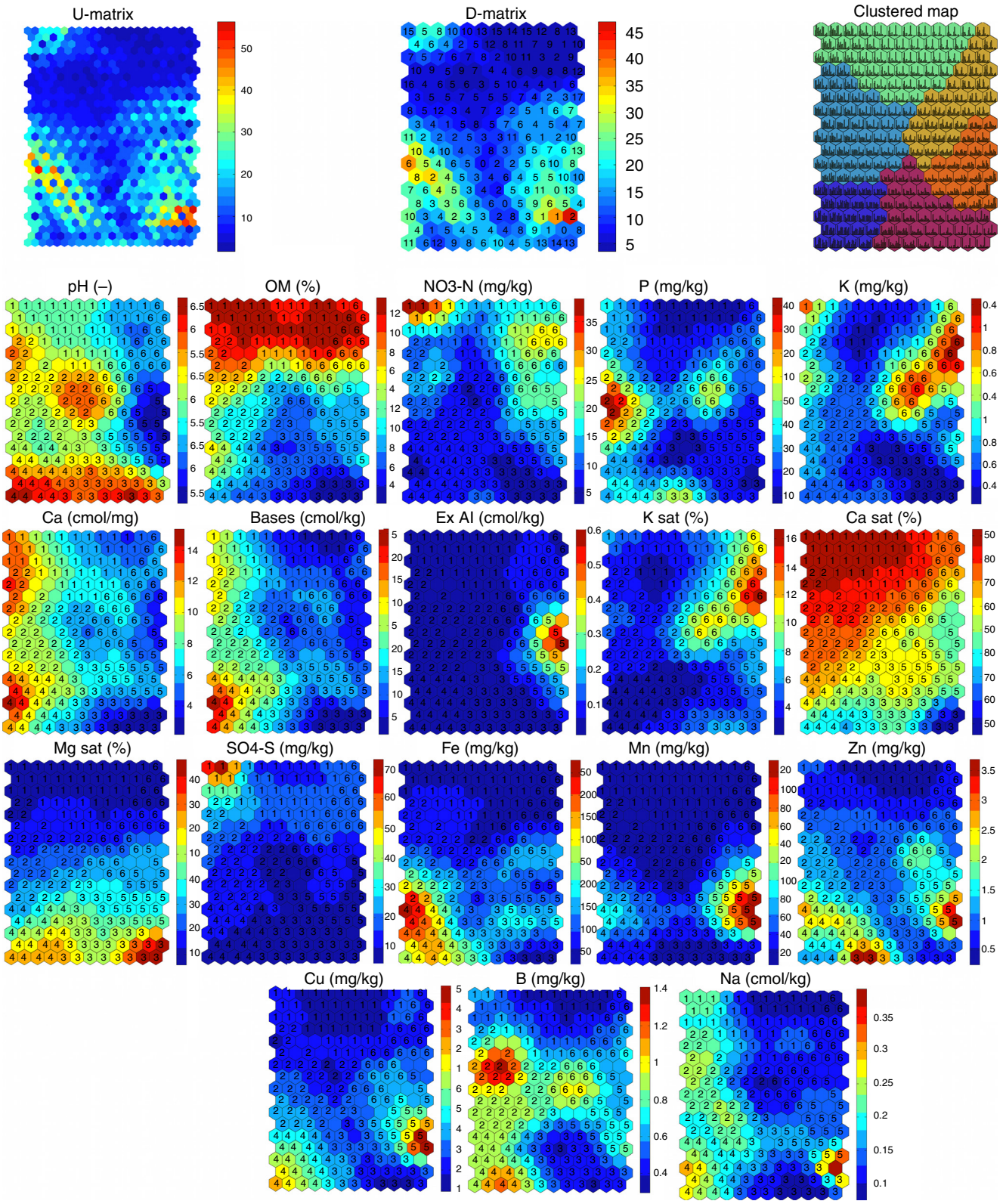

Figure 5 Top panels: **U**-matrix and distance matrix showing distance between each neuron and its neighbours. Cold colours indicate close location and warm colours indicate larger distance between neurons. Numbers in the neurons are the number of hits. Colours were assigned after clustering the self-organizing map (SOM) by *k*-means. Lower panels: component planes for each parameter. The numbers indicate the number of the cluster. Colour bars show nonscaled values. Large (small) values are associated with warm (cold) colours.

Ca, Na, bases, Mg Sat. This pattern presents small groups of neurons with large (or small) surrounded by similar neurons. The smooth transitions from large to small values show that there are not clear border between clusters (see Na in Figure 5), and values are part of a wide range, such as for pH; (ii) large-small values. This pattern appears for OM and Ca Sat as there is a balanced proportion of large and small values; and (iii) almost homogeneous small values with an isolated subset of large values for NO_3-N, Mn, Mg, P, K, Zn, Cu, Ex Al, K Sat, B, S, Fe. This pattern shows homogeneous values but a subset of samples with higher values. A clear example of this pattern is Ex Al, which shows that 95% of samples falls in the range of 0–0.3 cmol/kg, while the remaining 5% falls in the range from 0.3 to 0.6 cmol/kg. On the other hand, the component plane for Ca shows a more uniform distribution of the number of samples falling in each bin.

Even though there exists widely known relationships and dependencies among parameters, one advantage of the component planes is that it is possible to visualize similar patterns implying association between parameters: patterns for Ca and bases are very similar (Pearson's correlation $r = 0.9$ for input data) and Ca Sat and Mg Sat are opposite (Pearson's correlation $r = -0.78$ for input data).

However, not all parameters match another, as shown by pH, meaning that for this data set, no important associations with other investigated parameters are observed. For example, large K and K Sat values characterize cluster 6, and cluster 1 is characterized by low pH values and high values for Mn, Cu and Ex Al.

Prior knowledge of the database is normally available (e.g. sampling area, information use), so the use of SOM can give new insights about the database. In our study, some management-related information could be extracted. For instance, cluster 1 in the component plane for pH shows values below 5.5 and large values for Mn, Cu and Ex Al. Previous studies carried out by the local agricultural research station (Undurraga *et al.*, 2004) showed that pH values below 5.5 constrain nutrient uptake, suggesting the need for both management and monitoring practices. pH can be corrected by applying agricultural lime 30–60 days before seeding. Lime application rates are determined by soil tests. In the case of organic matter, soils derived from volcanic ashes present mean values of about 8% of OM (Zagal *et al.*, 2002). However, the corresponding component plane shows a considerable extent of cold colours (warmer colour are related to larger content) over the map, indicating that ca. 50% of the soil samples have levels below 8%. Samples with OM < 8% might belong to soil series presenting natural small organic matter content or could be indicative of loss of organic matter due to intensive management practices. At the foothills of the Andes Mountains, a loss of organic matter has been reported due to intensive tillage and stubble burning

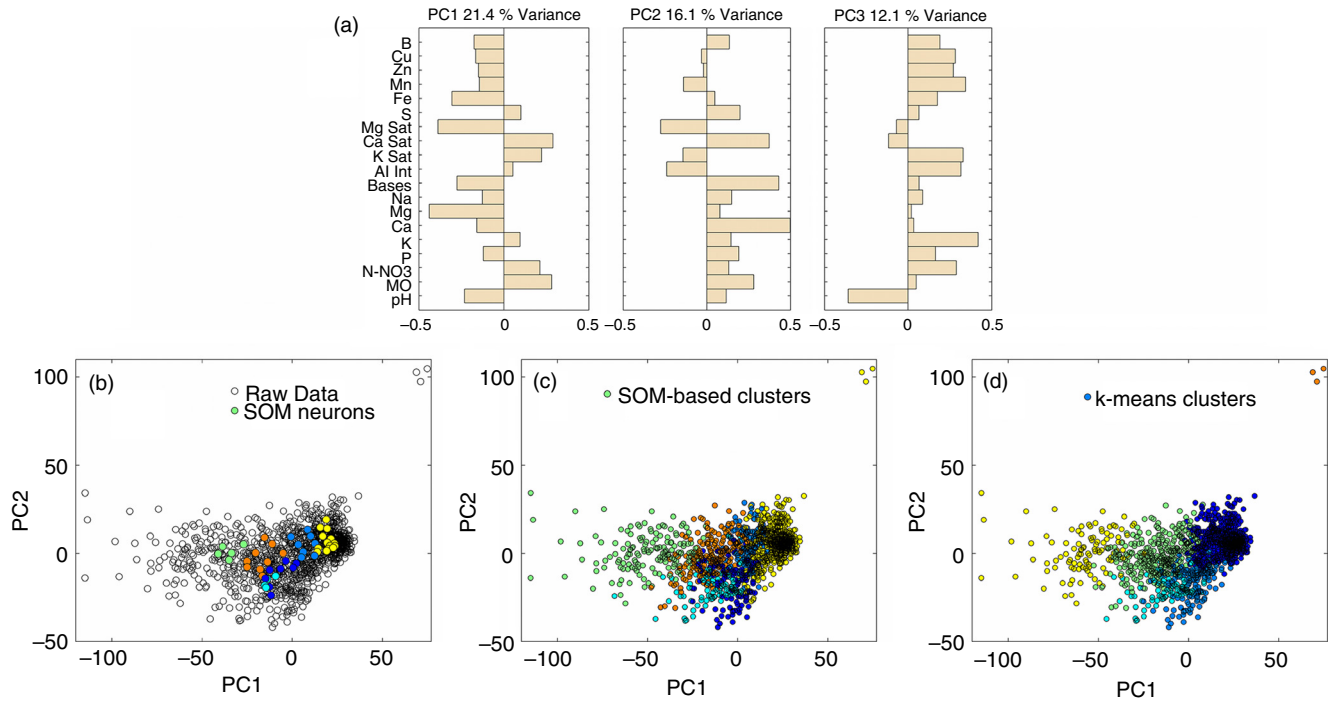

Figure 6 (a) Eigenvalues for the first three principal components used in the projection of input data, neurons and clustered data. Scatter plots for the first two principal components for (b) partition of the database after clustering the map; (c) direct application of the *k*-means algorithm to the input data set; and (d) input matrix and the codebook. Different clusters have different colours.

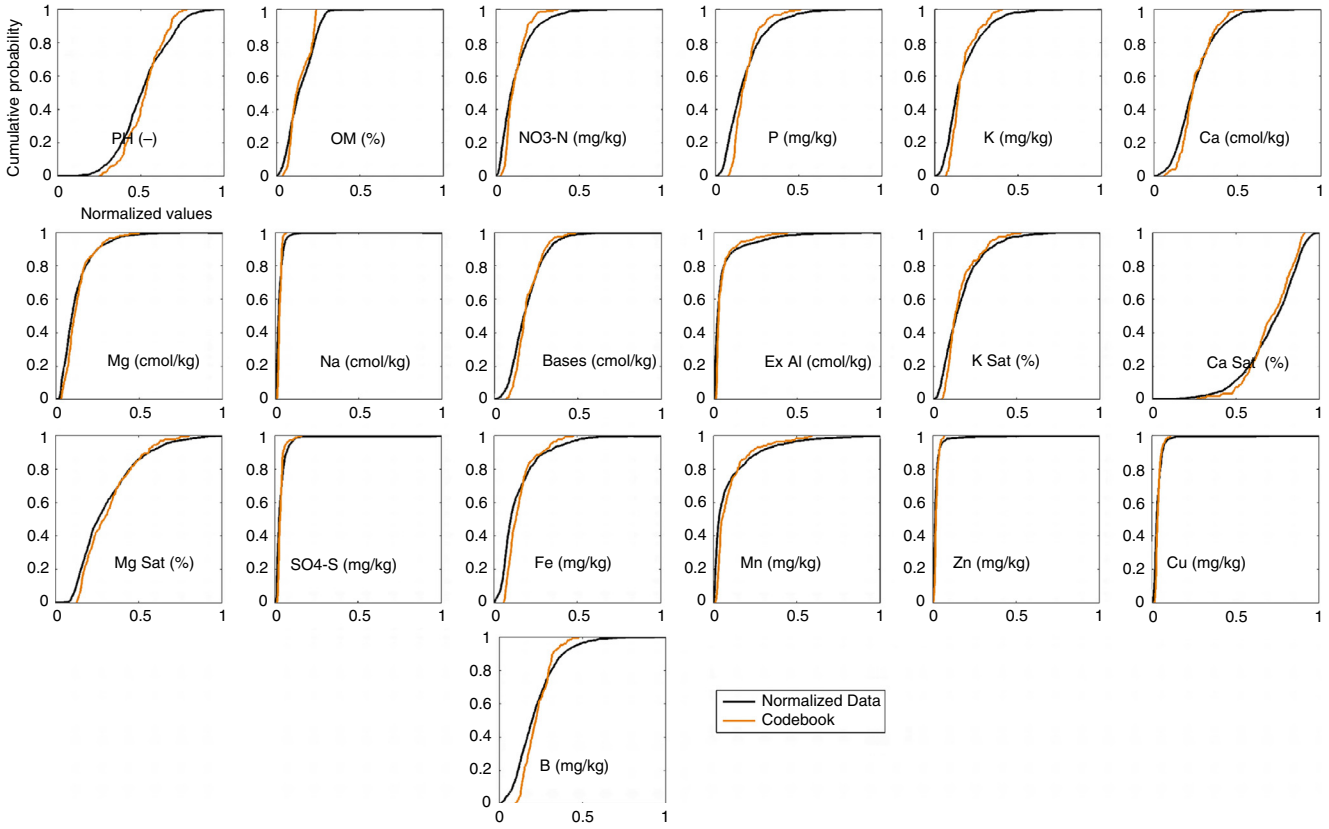

Figure 7 Cumulative probability for each parameter in both the input matrix and the codebook. Parameters are pH (–), organic matter (OM) (%), nitrate nitrogen (NO_3-N) (mg/kg), manganese (Mn) (mg/kg), phosphorus (P) (mg/kg), potassium (K) (mg/kg), calcium (Ca) (cmol/kg), zinc (Zn) (mg/kg), magnesium (Mg) (cmol/kg), sodium (Na) (cmol/kg), bases (cmol/kg), copper (Cu) (mg/kg), exchangeable aluminium (Ex Al) (cmol/kg), potassium saturation (K Sat) (%), calcium saturation (Ca Sat) (%), boron (B) (mg/kg), magnesium saturation (Mg Sat) (%), sulphate–sulphur (SO_4^--S) (mg/kg) and iron (Fe) (mg/kg).

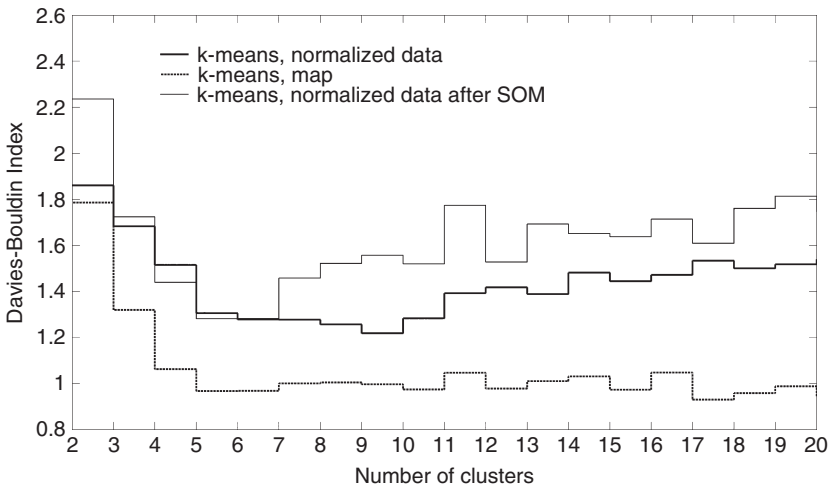

Figure 8 Comparison of DBI for different number of clusters. For clarity, values are shown as continuous between clusters.

(Rodríguez *et al.*, 2000; Sandoval *et al.*, 2008). As noted earlier, cluster 6 presents medium values for NO_3-N and P, but very large values for K compared against the rest of the map. High levels of K suggest the necessity to tailor fertilizing practices to avoid accumulation. In addition, it is possible to observe that some values of micronutrient

concentration (e.g. Mg and Zn) also point to a potential problem if fertilizing practices are not corrected.

SOM performance

For visualizing the data set, we used the first two PCs accounting for 37% of variance to produce two-dimensional scatter plots (Figure 6a). Even though the retained variance is low, our aim was to visualize a multidimensional data set without doing any statistical inference from PCA. Figure 6b shows the projection of input data set and neuron's reference vectors onto the first two PCs. As there are more neurons where the data set is denser, the SOM is approximating the frequency distribution. However, it is worth noting that the SOM is approximating the frequency distribution of the SQDB as a whole, but not approximating the probability distribution of a population. In one single database, one can have samples from several different populations. That said, the large values for Mn and Ex Al form a different group, but they could not be *a priori* considered as 'outliers' regarding the frequency distribution. Figure 7 shows good agreement when comparing the cumulative probability (CP) for each parameter in both the data set and the codebook. The codebook's CP, however, does not span the whole range [0,1], meaning that some very dissimilar samples are associated with 'mean' neurons.

After dividing the codebook by k-means, samples in the input matrix were grouped according to their BMU cluster, that is if neuron N belongs to cluster X, all samples having N as BMU were assigned to cluster X (the two-level approach; Vesanto & Alhoniemi, 2000). A comparison of Figure 6c (k-means clustering of input data) and Figure 6d (data clustering by applying k-means to the SOM) does not show clear qualitative or quantitative differences. Figure 8 shows DBI index for increasing values of k. The one-step approach – direct application of k-means algorithm to the SQDB – shows a small improvement in the DBI values after 10 clusters, while the two-step approach – grouping the SQDB based on map clustering – shows a minima close to $k = 5$ cluster, depending upon whether the method used to cluster the codebook is k-means or fuzzy k-means. This fact might be considered as indicative that the codebook effectively captured the topological features of the SQDB. The two-step approach generates a reduction in the computational cost (especially for larger data sets) and noise reduction in the clustering process, making the clustering less sensitive to variations in the input data (Vesanto & Alhoniemi, 2000).

Conclusions

The SOM method has potential practical advantages over other exploratory techniques. Its ability to visualize multidimensional data sets helps in the exploratory phase to gain understanding of the data and can also be used cooperatively with other techniques. Once the SQDB has been mined, the SOM could be used as a classification tool. The SOM also demonstrates potential to identify the quality state of the soil and define parameters to be improved.

The application of SOM, the two-step approach, improves the quality of the clustering compared against the direct application of methods such as k-means to a SQDB. Also, the computational load is less.

To summarize, a trained user in this methodology can use it to draw management strategies for soil management by inspecting patterns on data, association between parameters, and outliers based on the integration of different parameters in a single vector. The input matrix can be tailored to include relevant information for specific problems, allowing the design to fit-to-purpose indices.

Acknowledgements

This research was benefited from FONDECYT Grant No 11090032 and CONICYT/FONDAP-15130015. We would like to thank Professor Michael Goss for his guidance and suggestions to improve the manuscript. Thanks to Roto Quezada for supporting our research.

References

Alvarez-Guerra, M., González-Piñuela, C., Andrés, A., Galán, B. & Viguri, J. 2008. Assessment of Self-organizing Map artificial neural networks for the classification of sediment quality. *Environment International*, **34**, 782–790.

Astel, A., Tsakovski, S., Barbieri, P. & Simeonov, S. 2007. Comparison of self-organizing maps classification approach with cluster and principal components análisis for large environmental data sets. *Water Research*, **41**, 4566–4578.

Céréghino, R. & Park, Y. 2009. Review of the Self-organizing Map (SOM) approach in water resources: commentary. *Environmental Modelling & Software*, **24**, 945–947.

Chon, T.-S. 2011. Self-organizing maps applied to ecological sciences. *Ecological Informatics*, **6**, 50–61.

Davies, D & Bouldin, D. 1979. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1, **224**, 227.

Dhar, S. & Cherkassky, V. 2011. Application of SOM to analysis of Minnesota soil survey data. *Proceedings of the 2011 International Joint Conference on Neural Networks (IJCNN)*, 633–639. doi: 10.1109/IJCNN.2011.6033280.

Fidencio, P.H., Ruisánchez, I. & Poppi, R.J. 2001. Application of Artificial Neural Networks to the classification of soils from São Paulo state using near-infrared spectroscopy. *Analyst*, **126**, 2194–2200.

Herbst, M. & Casper, M.C. 2008. Towards model evaluation and identification using Self-organizing Maps. *Hydrology and Earth System Sciences*, **12**, 657–667.

Kalteh, A.M., Hjorth, P. & Berndtsson, R. 2008. Review of the self-organizing map (SOM) approach in water resources: analysis, modelling and application. *Environmental Modelling & Software*, **23**, 835–845.

Kaski, S. 1997. Data exploration using Self-organizing Maps. PhD thesis, Helsinki University of Technology, Finland.

Kohonen, T. 2001. *Self-organizing maps*, 3rd edn, pp. 105–175. Information Sciences Berlin, Heidelberg, New York.

Leflaive, J., Céréghino, R., Dager, M., Lacroix, G. & Ten-Hage, L. 2005. Assessment of Self-organizing Map to analyze sole-carbon source utilization profiles. *Journal of Microbiological Methods*, **62**, 89–102.

Martinez, W. & Martinez, A. 2005. *Exploratory data analysis with MATLAB*. CRC Press, Boca Raton, FL.

Mele, P. & Crowley, D. 2008. Application of self-organizing maps for assessing soil biological quality. *Agriculture, Ecosystems & Environment*, **126**, 139–152.

Merdun, H. 2011. Self-organizing map artificial neural network application in multidimensional soil data analysis. *Neural Computing & Applications*, **20**, 1295–1303.

Murtagh, F. & Hernández-Pajares, M. 1995. The Kohonen self-organizing map method: an assessment. *Journal of Classification*, **12**, 165–190.

Park, Y., Céréghino, R., Compin, A. & Lek, S. 2003. Applications of artificial neural networks for patterning and predicting aquatic insect species richness in running waters. *Ecological Modeling*, **160**, 265–280.

Peeters, L., Bacao, R., Lobo, V. & Dassargues, A. 2007. Exploratory data analysis and clustering of multivariate spatial hydrogeological data by means of GEO3DSOM, a variant of Kohonen's Self-organizing Map. *Hydrology & Earth System Sciences*, **11**, 1309–1321.

Penn, B. 2005. Using self-organizing maps to visualize high-dimensional data. *Computers & Geosciences*, **31**, 531–544.

Razavi, T. & Coulibaly, P. 2013. Classification of Ontario watersheds based on physical attributes and streamflow series. *Journal of Hydrology*, **493**, 81–94.

Rodríguez, N., Ruz, E., Valenzuela, A. & Belmar, C. 2000. Efecto del sistema de laboreo en las pérdidas de suelo por erosión en la rotación trigo-avena y praderas en la precordillera andina de la región centro sur. *Agricultura Técnica*, **60**, 259–269.

Sandoval, M.A., Stolpe, N.B., Zagal, E.M. & Mardones, M. 2007. The effect of crop-pasture rotacions on the C, N and S contents of soil aggregates and structural stability in a volcanic soil of south-central Chile. *Acta Agriculturae Scandinavica Section B-Soil and Plant Science*, **57**, 255–262.

Sandoval, M., Stolpe, N., Zagal, E., Mardones, M. & Celis, J. 2008. No-tillage organic carbon contribution and effects on an Andisol structure from the Chilean Andean foothills. *Agrociencia*, **42**, 139–149.

Tissari, S., Lersi, J. & Kolchmainen, M. 2007. Classification of Soil Groups using weights-of-evidence-method and RBFLN-neural nets. *Natural Resources Research*, **16**, 159–169.

Ultsch, A. & Siemon, H. 1990. Kohonen's self organizing feature maps for exploratory data analysis. In: *International Neural Network Conference* Publishers, pp. 305–308.

Undurraga, P., Rodríguez, N., Yoshikawa, S. & Claret, M. 2004. Antecedentes generales de los suelos del secano interior y fertilidad de suelos de la comuna de Ninhue. In: *Manejo y practicas conservacionistas del suelo para un desarrollo sustentable del secano* (eds S. Riquelme, C. Perez & Y. Shigehiko). INIA Bulletin 124, pp. 1–37. Instituto Nacional de Investigaciones Agropecuarias, Chillán, Chile.

Vesanto, J. 2002. Data exploration process based on the Self-organizing Maps. PhD thesis, Helsinki University of Technology, Finland.

Vesanto, J. & Alhoniemi, E. 2000. Clustering of the Self-organizing Map. *IEEE Transactions on Neural Networks*, **11**, 586–600.

Villamil, M., Miguez, F. & Bollero, G. 2008. Multivariate analysis and visualization of soil quality data for No-Till systems. *Journal of Environmental Quality*, **37**, 2063.

Wallner, M., Haberlandt, U. & Dietrich, J. 2013. A one-step similarity approach for the regionalization of hydrological model parameters based on Self-Organizing Maps. *Journal of Hydrology*, **494**, 59–71.

Webster, R. 2001. Statistics to support soil research and their presentation. *European Journal of Soil Science*, **52**, 331–340.

Wu, S. & Chow, T. 2004. Clustering of the self-organizing map using a clustering validity index based on inter-cluster and intra-cluster density. *Pattern Recognition*, **37**, 175–188.

Wu, T., Chellemi, D., Graham, J., Martin, K. & Rosskopf, E. 2008. Comparison of soil bacterial communities under diverse agricultural land management and crop production practices. *Microbial Ecology*, **55**, 293–310.

Zagal, E., Rodríguez, N., Vidal, I. & Quezada, L. 2002. Actividad microbiana en un suelo de origen volcánico bajo distinto manejo agronómico. *Agricultura Técnica*, **62**, 297–309.