Mika Sulkava and Jaakko Hollmén (2003). Finding profiles of forest nutrition by clustering of the Self-Organizing Map. In Proceedings of the 4th Workshop on Self-Organizing Maps (WSOM 2003). Kitakyushu, Japan, 11-14 September 2003, pages 243-248.

Finding Profiles of Forest Nutrition by Clustering of the Self-Organizing Map

Mika Sulkava, Jaakko Hollmén

Helsinki University of Technology, Laboratory of Computer and Information Science, P.O. Box 5400, FIN-02015 HUT, Finland
Tel. +358-(0)9-451 5467, Fax. +358-(0)9-451 3277
e-mail: Mika.Sulkava@hut.fi,Jaakko.Hollmen@hut.fi
Keywords: Self-Organizing Map, clustering, forest nutrition

Abstract— Understanding the nutritional states and profiles of tree species is important for monitoring the well-being of forests. Data from foliar surveys are available, but there is still need to better understand the underlying nutritional mechanisms in trees. In this paper, the nutrient concentrations of pine and spruce needles in Finland between 1987–2000 are analyzed to build nutrition profiles. The profiles are built from the data by clustering of the Self-Organizing Map. The VS algorithm divides the data into base clusters using region growing and forms a hierarchy from the base clusters. The hierarchy tree is pruned and the final clusters are selected from the pruned tree. We were able to divide the measurements into six groups. In each group the growth of the needles and the amounts of the nutrients were different and thus, different groups represented different kinds of growing conditions. With the help of the domain expert, using the results of the clustering method, it was possible to construct a temporal model that characterizes the development of the forests of Finland.

1 Introduction

Living plants are capable of taking up substances from the environment and using them for the synthesis of their cellular components. These nutrients play an important role in the physiological and biochemical processes of forest ecosystems. Because the foliar mineral composition is closely related to the environment, chemical foliar analysis is a useful diagnostic and monitoring tool in environmental and forest research [4].

In this study, clustering of the Self-Organizing Map is used to analyze the relations within the chemical composition of tree foliage. Measurements from pine and spruce forests in Finland between 1987 and 2000 are used as the data set for the method.

Previously, the algorithm has only been tested with artificial data and thus, this was the first opportunity to test the performance of the algorithm with real-world data. The clustering method has been presented in more detail in [9] and some results with the nutrition data in [5, 6].

It is found that clustering of the SOM is a useful tool in forest nutrition analysis. The clustering method is able to effectively represent the structure in the relations of nutrient concentrations.

In Section 2, the forest nutrition data used in the analysis is described. In Section 3, the standard SOM is described and it is followed by a description of the VS algorithm used in this study. The results are shown in Section 4 and the study is summarized in Section 5. Section 6 discusses further work.

2 Forest Nutrition Data

The measurements were made from needle samples collected from the conifer trees in forests of Finland. The measured variables were different nutrient concentrations and the mass of the needles. The measurements and analyses of the data described in this section were mostly carried out by the personnel of the Finnish Forest Research Institute. For details concerning the measurement techniques, see [4].

Three concentration measurements: nitrogen N (mg/g), sulfur S (mg/g), and phosphorus P (mg/g) and the mass of the needles were used in the analysis, as they were thought as important by the domain experts. In addition, some other nutrient concentrations including for example Calcium Ca, magnesium Mg, and potassium K were measured, but not used with the analysis method. The needle mass (NM) was reported as the mass of 1000 needles ($g/1000$). 36 stands throughout Finland were sampled annually for the above-mentioned variables between 1987–2000. There was, however, some missing data. From the 504 measurements (14 years, 36 stands), there were 216 missing values from NM and 137 from each of the nutrient concentrations. Altogether 31% of the measurements were missing. In 16 stands, the main tree species was Norway spruce (*Picea abies*) and in the rest Scots pine (*Pinus sylvestris*).

3 Clustering using the Self-Organizing Map (SOM)

A clustering algorithm based on the Self-Organizing Map (SOM) [3] was chosen for this problem because of its good visualization properties. The SOM preserves neighborhood relations and presents high-dimensional datasets on a 2-dimensional grid. It is therefore an important tool in data mining with its main applications in visualization and clustering.

3.1 The Self-Organizing Map

The Self-Organizing Map consists of a low-dimensional, usually 2-dimensional, regular grid of map units that are connected to adjacent ones by a neighborhood relation [3]. The grid can be effectively used to visualize and explore properties of the data [8]. Each map unit i is represented by a prototype vector, $\mathbf{m}_i = [m_{i1}, \ldots, m_{id}]^T$, where d is input vector dimension.

The prototype vectors define a tessellation of the input space into a set of Voronoi sets

$$V_i = \{\mathbf{x} \mid \|\mathbf{x} - \mathbf{m}_i\| < \|\mathbf{x} - \mathbf{m}_j\| \, \forall j \neq i\}, \qquad (1)$$

where $\mathbf{x}$ are the data vectors and $\| \cdot \|$ is the Euclidean norm. In effect, each data vector belongs to the Voronoi set of the prototype to which it is closest. In other words, SOM quantizes the training data set with a representative set of prototype vectors. The quantization process is regularized by the neighborhood relation such that topology of the data set is preserved.

3.2 The VS clustering algorithm

Clustering algorithms can be divided into two main categories: partitive and hierarchical algorithms. Partitive algorithms divide the data set into non-overlapping partitions whereas hierarchical algorithms construct a hierarchy tree of the clusters. Usually, the structure of the hierarchy is such that all the data belongs to the top level cluster and at the bottom level, each data vector forms a separate cluster.

In this study, the hierarchical VS clustering algorithm, named after its developers [9], is used to divide the data into clusters. It is a two-level approach: first, a SOM is trained and the data is partitioned into a large number of Voronoi sets, each corresponding to one map unit. Subsequently, the map units are clustered. All data vectors in a Voronoi set belong to the same cluster as the corresponding map unit. An advantage over more traditional methods like k-means is that the result can be effectively visualized on the 2-dimensional grid. Another advantage is that by clustering the SOM rather than the data directly, significant gains in the speed of clustering can be obtained [8].

U-matrix is a commonly used tool to cluster the SOM visually [7]. It visualizes distances between each map unit and its neighbors. Unfortunately, when clusters are identified visually, the results by different people are not necessarily the same. Therefore, an automated clustering algorithm that follows the results of the U-matrix was used. The details of the algorithm are presented in [9]. The basic idea of the VS algorithm is as follows:

(1) The data is quantized with SOM and a distance matrix, showing the median distances between neighboring map units is calculated.
(2) The map is divided into a set of base clusters. This is done using region growing with local minima of the distance matrix as seed points.
(3) A cluster hierarchy is constructed from the base clusters using an agglomerative algorithm and a pruning procedure.
(4) The final partitioning with suitable number of clusters is obtained from the hierarchy.

The phases (2)–(4) are described in detail in the following sections.

3.2.1 Region growing

The region-growing starts with setting the local minima of the distance matrix as seed points. These are the map units whose median distance to neighboring units is smaller than the median distance of any of the neighboring units to their neighbors. Next, the unassigned map unit with smallest distance to a cluster is found and assigned to the corresponding cluster. Here, a continuity constraint is used to ensure that the clusters form continuous areas on the map. Only those unassigned map units are considered for merging that are neighbors of map units that belong to a cluster. Assigning the map units is continued until all map units belog to a cluster.

This procedure provides a partitioning of the map into a set of base clusters, number of which is equal to the number of local minima on the distance matrix. A problem is that the distance matrix may have some local minima that are products of random variations in the data rather than real local maxima of the probability density function of the data. Such base clusters can be pruned out of the clustering in a hierarchical fashion.

3.2.2 Cluster hierarchy

In cluster analysis, constructing a cluster hierarchy is often beneficial [2]. Apart from the need for pruning explained above, a cluster hierarchy may represent the true structure of the data better than a single-level partitioning. Some clusters can be considered super-clusters, consisting of several sub-clusters, which allows the data to be investigated at several levels of detail.

An agglomerative clustering algorithm was used to construct the cluster hierarchy from the base clusters. This, however, produces a binary tree which may not be representative of the true structure of the data. The extra intermediate clusters were pruned out as explained in [9].

3.2.3 Final partitioning

A particular partitioning is obtained from the pruned tree by starting from the top with all data in a single cluster, and traversing the tree downwards by always splitting the intermediate node with best clustering validity index. The selection criterion for the number of final clusters was the following: the sub-clusters were combined into a super-cluster if the distance between the means of the sub-clusters was smaller than the standard deviation of the whole data.

4 Results

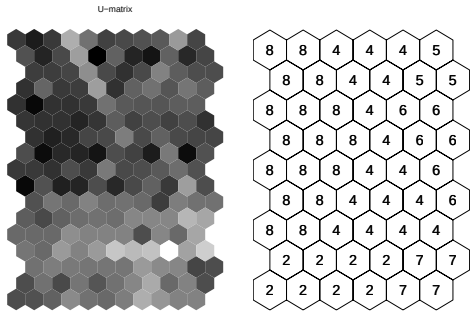
Because there was only a little a priori information about the structure of the data, a clustering method was used to analyze the relationships between the nutrients at the stand level. The N, S and P concentrations (c) in the needles and needle mass of the current year's needles were used as the 4-dimensional data vector in the clustering method. The data vector for measurement stand i and year t was thus

$$\mathbf{x}_{i,t} = (c_{i,t}(N),\ c_{i,t}(S),\ c_{i,t}(P),\ NM_{i,t})^T . \quad (2)$$

These elements were used because of their importance to the growth of the tree and their dynamic temporal behavior. Needle mass was used because it enabled us to interpret the clusters based on nutrient concentration as well as nutrient content.

The data set was clustered using the algorithm explained in Section 3.2. The SOM was initialized linearly, i.e. the map weight vectors were initialized along the two greatest eigenvectors of the covariance matrix of the data. Then, the map was trained with SOM Toolbox[1] using the batch algorithm [3] in two rough training epochs and five fine tuning epochs. The final neighborhood width was 1 in order to ensure good quality of quantization. A map consisting of a regular hexagonal grid with 6×9 map units was used.

The U-matrix of the SOM is shown in Figure 1a and the final clustering result in Figure 1b. As expected, the clustering result is similar to the U-matrix [9]. The hierarchical structure of the clustering can be seen in Figure 2. One cluster (9) was pruned out of the hierarchy and bottom level clusters 1 and 3 were combined into cluster 8 in the final clustering. Clusters 2 and 7 are the most different from all other clusters. These two clusters have the highest mean values of the concentration variables.



(a) U-matrix (b) Clustering

Figure 1: *(a) U-matrix of the SOM. Darker color indicates smaller value. (b) The clustering result. The numbers (2-8) indicate which cluster the map unit belongs to. Some numbers are not present, because those clusters weren't chosen from the hierarchy to the final clustering.*

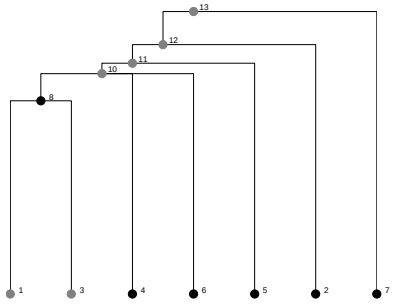


Figure 2: *Cluster hierarchy of the SOM. Black circles are the final clusters.*

The mean values and standard deviations of all the measurements for the clusters are shown in Tables 1 and 2. The component planes of the SOM are shown in Figure 3.

Needle mass is the variable that had the most missing measurements. Despite of that, needle mass contributed quite much to the clustering result. It can be seen that clusters 4–7 have rather low and clusters 2 and 8 rather high mean needle mass values. These two groups of clusters don't overlap each other much with respect to needle mass, i.e. their standard deviations are fairly small. The reason for this is that according to histograms (not shown), needle mass has a clearly bimodal probability density function, whereas all the concentration variables have more or less unimodal probability density functions. The bimodality of needle mass is simply caused by the two different tree species having needles of different size. Now, the two groups of clusters approximately correspond to the two different peaks of the probability density function of needle mass.

[1] `http://www.cis.hut.fi/projects/somtoolbox`

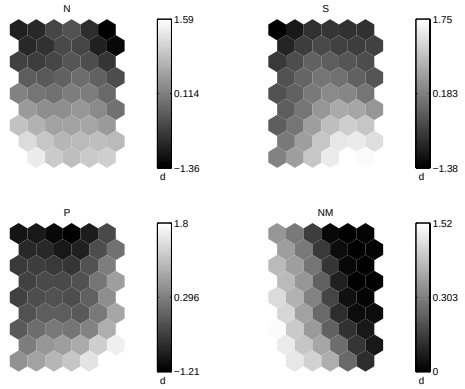


Figure 3: *The component planes of the SOM. The values of the component planes correspond to the values of the normalized data.*

Table 1: *The means and standard deviations of the concentration and needle mass measurements of the clusters for spruce.*

Cluster	2	4	5
N (mg/g)	14.9±1.8	12.2±0.9	10.1±0.5
S (mg/g)	0.95±0.07	0.95±0.12	0.84±0.07
P (mg/g)	1.79±0.12	1.32±0.19	1.50±0.21
NM (g/1000)	-	4.2±0.8	4.4±0.9
Ca (mg/g)	4.84±1.31	4.61±1.26	3.29±0.90
Mg (mg/g)	1.15±0.16	1.16±0.16	1.12±0.08
K (mg/g)	7.12±0.53	6.38±0.86	6.34±0.91
Cluster	6	7	8
N (mg/g)	11.6±0.7	13.6±0.9	12.4±1.6
S (mg/g)	0.91±0.07	1.11±0.10	0.82±0.06
P (mg/g)	1.71±0.17	2.05±0.24	1.36±0.17
NM (g/1000)	4.3±0.8	4.5±0.7	-
Ca (mg/g)	4.98±1.30	5.07±1.16	5.23±1.43
Mg (mg/g)	1.22±0.16	1.28±0.12	1.16±0.19
K (mg/g)	6.42±1.10	6.49±1.10	6.82±0.51

The clusters have slightly different meanings for the tree species. The nutrition profiles, i.e. the mean values and standard deviations of the clusters for the two species are in Tables 1 and 2. For pine, the clusters can be explained as follows [5]. Cluster 5 represents trees with multiple-nutrient deficiency. All the concentrations and needle mass are low. Clusters 4 and 6 represent a sub-optimal nutrient status. Cluster 4 is characterized by a deficiency of P and cluster 6 may have P-excess. Clusters 2 and 7 have high S and P concentrations. Both clusters have excess of these nutrients but only cluster 2 has high needle mass. Cluster 8, which is the most common one, has favorable S and P concentrations but N is probably a limiting factor of the growth.

For Norway spruce, cluster 7 represents forests with an excess of N and S in their needles. Nevertheless, the needles of cluster 7 have the highest needle mass of all six clusters for spruce. Clusters 4, 5 and 6 are characterized by deficiency of respectively Mg, Ca & K, and K. Only a small but constant number of Norway spruce forests was characterized by cluster 5. In spruce forests this is a more dynamic cluster than than in pine forests. Due to the absence of needle mass data, the nutrient status of clusters 2 and 8 could not be characterized. For Norway spruce, clusters 2 and 8 have no observations for needle mass, and consequently values for the nutrient content are not available. Norway spruce needles just started to show these nutrient profiles in 1999 and 2000. In these years the needle mass was not measured.

Table 2: *The means and standard deviations of the concentration and needle mass measurements of the clusters for pine.*

Cluster	2	4	5
N (mg/g)	13.6±0.9	12.5±0.7	9.7±0.7
S (mg/g)	1.05±0.09	0.97±0.08	0.86±0.09
P (mg/g)	1.76±0.16	1.51±0.14	1.58±0.12
NM (g/1000)	12.6±2.8	8.9±1.5	5.2±1.4
Ca (mg/g)	2.38±0.40	2.20±0.40	2.96±1.02
Mg (mg/g)	1.14±0.15	1.06±0.09	1.12±0.15
K (mg/g)	5.61±0.45	5.10±0.56	6.45±0.62
Cluster	6	7	8
N (mg/g)	11.4±1.1	13.6±0.7	11.8±1.0
S (mg/g)	0.94±0.10	1.18±0.10	0.89±0.08
P (mg/g)	1.75±0.17	2.04±0.13	1.43±0.12
NM (g/1000)	8.6±1.7	9.6±1.7	11.1±2.6
Ca (mg/g)	1.80±0.31	2.36±0.30	1.97±0.40
Mg (mg/g)	1.05±0.15	1.24±0.19	1.06±0.13
K (mg/g)	5.58±1.02	5.87±0.45	5.09±0.53

The temporal switching of the cluster of a stand was analyzed to find out if there is any structure in the development of nutrient concentrations. The most common switches were 8–2, 8–4, 2–8, 4–8 and 6–4. When considering two consecutive switches, the most common series were 8–2–8, 2–8–2, 8–4–8, 4–5–4, 4–6–4 and 4–8–2. These results are not very surprising since the most common clusters are 8, 4 and 2 and usually, the shifts happen between the most common clusters.

Because the absolute numbers of switches didn't reveal much information about the data, the conditional probabilities of switching the cluster were calculated. The estimated matrices correspond to transition probability matrices in a first-order Markov model. They are shown for spruce and pine stands respectively in Tables 3 and 4. The high probabilities tell something about the most typical switches. The process in spruce stands seems to have some tendency to converge to cluster 4 and in pine stands to cluster 8.

The typical transitions between clusters can be visualized more effectively by constructing graphs that show the switches between clusters. These graphs are shown in Figure 4 for both species separately. Spruce stands usually belong to clusters 4–7, cluster 4 being the most common one and 6 the second most common.

Table 3: *The temporal cluster switch probability matrix for spruce. The rows show the conditional probabilities of transition from a certain cluster to another.*

Prev.\Curr.	2	4	5	6	7	8
2	0.00	1.00	0.00	0.00	0.00	0.00
4	0.02	0.62	0.09	0.17	0.02	0.09
5	0.00	0.58	0.25	0.08	0.00	0.08
6	0.05	0.17	0.05	0.59	0.10	0.05
7	0.06	0.11	0.00	0.33	0.50	0.00
8	0.00	0.40	0.00	0.00	0.00	0.60

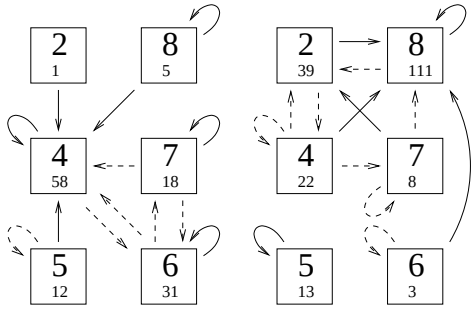
Table 4: *The temporal cluster switch probability matrix for pine. The rows show the conditional probabilities of transition from a certain cluster to another.*

Prev.\Curr.	2	4	5	6	7	8
2	0.41	0.13	0.00	0.00	0.03	0.44
4	0.27	0.14	0.00	0.00	0.14	0.45
5	0.08	0.08	0.77	0.00	0.00	0.08
6	0.00	0.00	0.00	0.00	0.33	0.67
7	0.62	0.00	0.00	0.00	0.12	0.25
8	0.11	0.08	0.00	0.01	0.00	0.80

Also, switches that happen with high probability are 5–4 and 8–4. In pine stands, the process is most of the time in cluster 8. As time goes on, more stands switch to cluster 8 than from cluster 8. Clusters 5, 6 and 7 are less common than 2 and 4. Switches that happen with high probability are 2–8, 4–8, 6–8 and 7–2.

In the years 1987, 1988, 1991 and 1993, the high needle mass cluster 8 was less usual than otherwise. In 1987 and 1988, the high sulfur concentration clusters 2 and 7 were more usual than normally. The reason for high number of low needle mass clusters in 1991 might be that there were a lot of measurements missing from everywhere else but southern Finland, where the low needle mass clusters are normally more probable than elsewhere. This is caused by the fact that there are no spruce stands in northern Finland. What reduces the significance of this result is that in 1991, there were needle mass measurements only from two stands. In 1993, the low needle mass cluster 4 was the most common one. Starting from 1995, the number of stands in cluster 4 has decreased.

The clustering result on a geographical map for each year can be seen in Figure 5. The probability of a stand to be in cluster 8 doesn't seem to be very much connected to the geographical position of the stand. In southern Finland, other clusters are a little more common than cluster 8. Stands in the other clusters are spread more unevenly on the map. A stand in cluster 2 is most likely in northern or south-western Finland. Clusters 4 and 7 can usually be found in southern Finland and cluster 4 also in middle Finland. Cluster 5 exists most often in south-eastern Lapland (northern region in Finland) and cluster 6 in southern and western Finland.



(a) Spruce (b) Pine

Figure 4: *Graphs showing the typical cluster switches. Solid line denotes a probability higher than 0.4 and dashed line a probability between 0.1 and 0.4. The smaller number under the cluster number is the number of years a stand has belonged to that cluster. Typical transitions are shown for spruce (a) and for pine (b).*

5 Summary and Conclusions

In this study, the nutrient concentrations of pine and spruce needles in Finland between the years 1987–2000 were analyzed using clustering of the Self-Organizing Map. The VS clustering algorithm was used in the clustering and its performance with real-world data was tested.

The clustering provided new information about the relations of the nutrients between different years and locations. With the clustering method, it was possible to divide the measurements into six groups. In each group, the growth of the needles and the amounts of the nutrients were different and thus, different groups represented different kinds of growing conditions. Using the result of the clustering method, it was possible to construct a temporal model that characterizes the development of the forests of Finland.

6 Further work

Processing and interpretation of large-scale foliar surveys is hampered by difficulties to describe the structure in the data set and to interpret the observations. The VS clustering method was found to be a promising approach to describe the structure in the data set, however, some further improvements are needed to improve the interpretation of the observations.

First, a model should be constructed that effectively uses both the spatial and temporal dimensions of the data. One possibility to achieve this would be to construct different time series models for different parts of the country. It would probably also be worth trying to use separate models for different tree species and perhaps even different weather conditions. The weather data could be included as a more internal part to the models. In this study, only the current year's needles

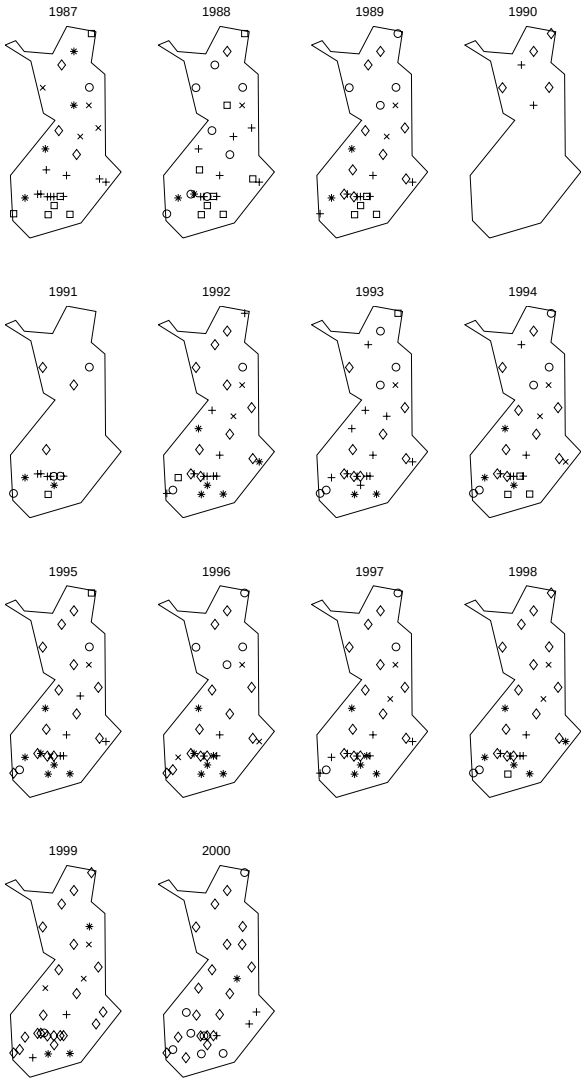


Figure 5: *Clustering of the measurement stands for each year plotted on the map of Finland. "○" = cluster 2, "+" = cluster 4, "×" = cluster 5, "∗" = cluster 6, "□" = cluster 7, "◇" = cluster 8.*

were analyzed. Using also measurements from older needles, more information about the growth and development of the needles could be extracted.

In the future, the clustering model could be enhanced by using probability distributions instead of the crisp clusters. This kind of fuzzy clustering can be obtained by using for example Gaussian mixture models or building the Gaussian distributions on top of the Self-Organizing Map as in [1]. In addition to the SOM, simple and perhaps useful visualizations of the data could be achieved using other projection methods like Sammon's mapping and principal component analysis.

7 Acknowledgements

The data used in this study were colleced under the Finnish National Programme of the International Cooperative Programme on Assessment and Monitoring of Air Pollution Effects on Forests (ICP Forests) of the UN/ECE, and EU Scheme on the Protection of Forests against Atmospheric Pollution. We thank Dr. Sebastiaan Luyssaert at the Finnish Forest Research Institute for fruitful collaboration and many enlightening discussions. We also thank Dr. Juha Vesanto (the V in VS) for collaboration.

References

[1] Esa Alhoniemi, Johan Himberg, and Juha Vesanto. Probabilistic measures for responses of self-organizing map units. In *Proceedings of the International ICSC Congress on Computational Intelligence Methods and Applications (CIMA'99)*, pages 286–290. ICSC Academic Press, 1999.

[2] George Karypis, Eui-Hong (Sam) Han, and Vipin Kumar. Chameleon: Hierarchical clustering using dynamic modeling. *Computer*, 32(8):68–75, August 1999.

[3] Teuvo Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, Berlin, third edition, 2001.

[4] Sebastiaan Luyssaert, Hannu Raitio, Pieter Vervaeke, Jan Mertens, and Noël Lust. Sampling procedure for the foliar analysis of deciduos trees. *Journal of Environmental Monitoring*, 4:858–864, 2002.

[5] Sebastiaan Luyssaert, Mika Sulkava, Hannu Raitio, and Jaakko Hollmén. Is nitrogen deposition altering the chemical composition of Norway spruce and Scots pine needles in Finland? Submitted, 2003.

[6] Mika Sulkava. Identifying spatial and temporal profiles from forest nutrition data. Master's thesis, Helsinki University of Technology, May 2003.

[7] A. Ultsch and H. P. Siemon. Kohonen's Self Organizing Feature Maps for Exploratory Data Analysis. In *Proceedings of International Neural Network Conference (INNC'90)*, pages 305–308, Dordrecht, Netherlands, 1990. Kluwer.

[8] Juha Vesanto and Esa Alhoniemi. Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, 11(3):586–600, May 2000.

[9] Juha Vesanto and Mika Sulkava. Distance matrix based clustering of the self-organizing map. In *Proceedings of the International Conference on Artificial Neural Networks - ICANN 2002*, Lecture Notes in Computer Science, No. 2415, pages 951–956. Springer-Verlag, 2002.