Integrating Environmental, Technological and Financial Data In Forest Industry Analysis

Juha Vesanto[1], Petri Vasara[2], Riina-Riitta Helminen[2] and Olli Simula[1]
e-mail Juha.Vesanto@hut.fi, Petri.Vasara@poyry.fi,
Riina-Riitta.Helminen@poyry.fi, Olli.Simula@hut.fi

[1] Helsinki University of Technology, P.O.Box 2200, FIN-02015 HUT, Finland
[2] Jaakko Pöyry Consulting, Jaakonkatu 3, FIN-01620 VANTAA, Finland

Abstract. The Self-Organizing Map (SOM) is a powerful neural network method for the analysis and visualisation of high-dimensional data. In this paper, the SOM algorithm is applied to the analysis of the technology of world paper and pulp industry. It is seen that the method can be used on environmental, technological and financial data to produce a comprehensive view of the industry as a whole.

1 Introduction

The emergence of environmental issues as a significant dimension in forest industry decision-making has led to a situation of uncertainty and questions not yet answered. Many environmental factors are unknown and, simultaneously, these very same factors should be integrated with their technological and financial counterparts into everyday problem solving.

This situation is paralleled in the problem which the joint Jaakko Pöyry Consulting/HUT project ENTIRE, set out to solve: mapping environmental, technological and financial data and combining them. Each component is a problem in its own right; the integration phase creates new aspects of the problem.

Both the high dimensionality of the data, the need for visualisation and our own earlier experience in forest industry process monitoring clearly pointed towards the use of the Self-Organising Map (SOM) [2] as the analysis tool. Indeed, what we are doing here can equally well be seen as a form of process monitoring. The basis for the project was, in addition to the problem and the method, the data at our disposal. The financial databanks of Jaakko Pöyry Consulting go back into the seventies and serve as the basis e.g. for the Jaakko Pöyry Company Handbook. The technological databanks, separated into the paper machine and the pulp mill sections, cover most of the world to a very high level of detail.

The environmental databanks are the newest and most problematic. Given that in most of the world environmental data are not available to the public, the distribution and reliability of available data is very uneven. This must in general be addressed by going deep where possible and extrapolating from available example mills where necessary.

Each of the areas to be mapped (environmental, technological, financial data) is worthy of a separate look, but it was felt that presenting the results should start with the most secure component (technology) followed by the most interesting (the integration).

2 Analysis Method Based on the Self-Organizing Map

The Self-Organizing Map (SOM), developed by professor Kohonen, is one of the most popular neural network models [2]. The SOM algorithm is based on unsupervised learning, which means that the training is entirely data-driven and that little or no *a priori* information about the input data is required. The SOM can, thus, be used to detect features inherent to the problem.

A SOM consists of neurons which are organized on a regular, typically 2-dimensional grid. Each neuron of the SOM is represented by an n-dimensional weight vector, where n is equal to the dimension of the input vectors. The number of neurons may vary from a few dozen upto several thousand. The SOM algorithm performs a topology preserving mapping from the high-dimensional input space to the map units so that the relative distances between data points are preserved. Data points lying near each other in the input space are mapped to nearby map units of the SOM. The SOM can, thus, serve as a clustering tool of high-dimensional data. The SOM also has a capability to generalize, i.e. the network can interpolate between previously encountered inputs.

The SOM forms an elastic net that during learning folds onto the "cloud" formed by the input data. The net tends to approximate the probability density function of the input data. This means that the weight vectors tend to drift there where the data is dense, while there are only a few weight vectors where data is sparsely located.

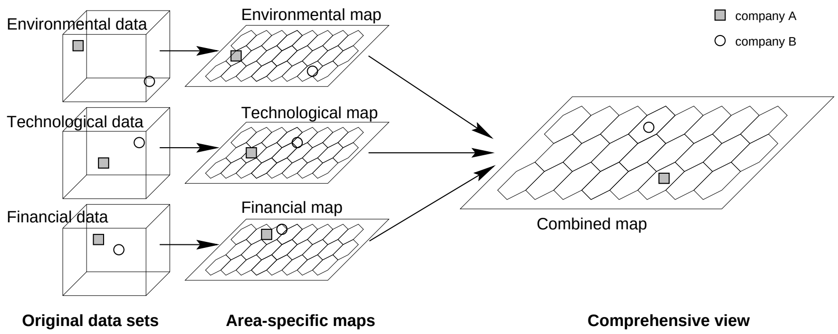


Fig. 1. The hierarchical data fusion method. In comparative study the differences and similarities between companies (A/B) can be seen from the maps.

The SOM has been found to be a powerful tool in data mining [1]. Instead of a single SOM, however, our analysis method stretches further by using a hierarchical combination of SOMs. The architecture of the method is depicted in figure 1.

The available data is first separated into initial data pools. The exact number of pools depends on the application and the goals of the analysis. One SOM is then trained for each of the data pools. These low level maps give the analyzer information about the relations of different data vectors within that specific area. By combining the lower level maps a comprehensive view of the whole data set is obtained. The combined SOM together with the lower level maps can be used to get information about the relations of data vectors in each area separately and in the big picture.

3 Case study

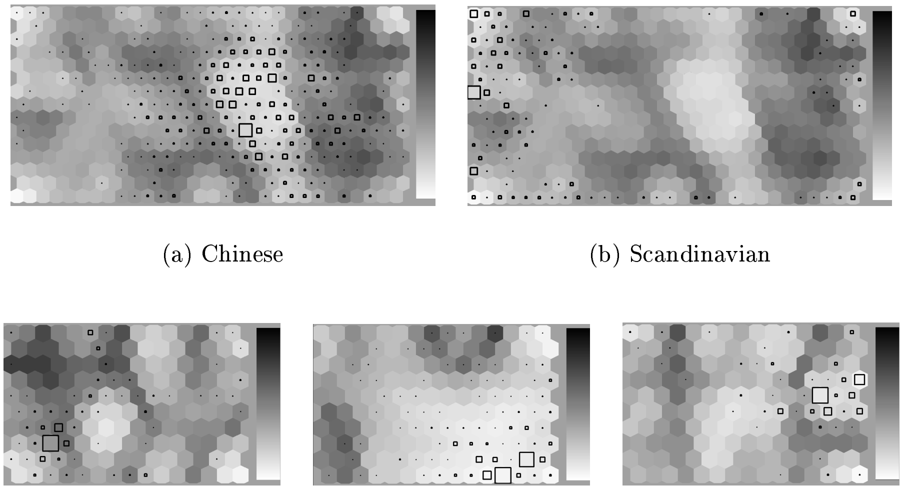
In the case study, the technology of paper mills all over the world was studied. The data was divided into three separate sets: information of the mill itself, its paper machines and its pulp production. One SOM (sized 15x10 neurons) was trained for each of the three sets, and a fourth SOM (30x15 neurons) was built using the concatenated BMU coordinates of the input data on the three low level maps. The same setup can be used to combine environmental, technological and financial data to form a comprehensive whole as depicted in figure 1.

Figures 2a and 2b show the unified distance matrix (u-matrix) [3] of the combined map, and the distribution of Chinese and Scandinavian paper mills on the map. The two mill sets are easily separable, although there was no geographic information present in the data. It can also be seen that in the area, where the Chinese mills are, the values of the u-matrix are very low. That is, the variation between weight vectors in that area is low, which means that most of the Chinese mills resemble each other.

Figures 2c, 2d and 2e show the distribution of Chinese paper mills on the three low level maps. Also from these figures it is apparent that most of the Chinese mills are centered on a single area of the map. Taking a closer look at the weight vectors in these areas, it can be seen that typical Chinese paper mill has small capacity, a large number of paper machines and it most probably produces printing/writing paper. The paper machines are small, slow and the paper weight is low. The pulp is produced chemically.

4 Conclusion

Given the nature of the problem, we are only in the initial phases. The amount and quality of the data can and must be improved. The data vectors should be trimmed down, and the consequent identification of key separators will have value in itself. The time dimension must be dealt with better, and the trajectories of a mill or company in the environmental, technological and financial spheres may give clues to forecasting the unit's success: an analogy, again, to process



(a) Chinese (b) Scandinavian

(c) Mills (d) Paper machines (e) Pulp production

Fig. 2. Chinese (a) and Scandinavian (b) paper mills on the u-matrix of the combined map. The lighter the color, the smaller the value of the u-matrix. The black squares represent the classified mills. The bigger the square, the more mills in that map unit. The second row shows Chinese paper mills on the three low level maps: mills (c), paper machines (d) and pulp production (e).

monitoring. The link to eco-efficiency (value added per environmental impact) is also worthy of exploration. However, even in its present stage ENTIRE offers scope for analysis and interpretation of the dimensions and their connections. It is possible that no hitherto unknown relationships will be discovered, but even the confirmation of thumb rules and expert intuition is of value.

5 Acknowledgments

This work has been carried out in the technology program "Adaptive and Intelligent Systems Applications". The work has been financed by Jaakko Pöyry Consulting and the Technology Development Center of Finland (TEKES).

References

1. S. Kaski. *Data Exploration Using Self-Organizing Maps.* Doctor's Thesis, Helsinki University of Technology, 1997.
2. Teuvo Kohonen. *Self-Organizing Maps.* Springer, Berlin, Heidelberg, 1995.
3. A. Ultsch and H.P. Siemon. Kohonen's self organizing feature maps for exploratory data analysis. In *Proc. INNC'90, Int. Neural Network Conf.*, pages 305–308, 1990.