In Apostolos-Paul N. Refenes, Yaser Abu-Mostafa, John Moody, and Andreas Weigend (Eds.) Neural Networks in Financial Engineering. Proceedings of the Third International Conference on Neural Networks in the Capital Markets, London, England, 11-13 October, 1995. World Scientific, Singapore, pp. 498-507, 1996.

EXPLORATORY DATA ANALYSIS BY THE SELF-ORGANIZING MAP: STRUCTURES OF WELFARE AND POVERTY IN THE WORLD

S. KASKI and T. KOHONEN
Neural Networks Research Centre, Helsinki University of Technology
Rakentajanaukio 2 C
FIN–02150 Espoo, Finland
E-mail: samuel.kaski@hut.fi

The self-organizing map (SOM) is a method that represents statistical data sets in an ordered fashion, as a natural groundwork on which the distributions of the individual indicators in the set can be displayed and analyzed. As a case study that instructs how to use the SOM to compare states of economic systems, the standard of living of different countries is analyzed using the SOM. Based on a great number (39) of welfare indicators the SOM illustrates rather refined relationships between the countries two-dimensionally. This method is directly applicable to the financial grading of companies, too.

1. Introduction

Prediction of economic time series, which is by far the most intensely studied application of neural networks in the area of economics, can be based on two different philosophies. For instance, a model of the phenomenon which produces the time series can be formed, and the prediction can then be based on the behavior of the model. Alternatively, a general-purpose function approximator, which does not utilize any knowledge of the (economic) nature of the time series, can be used to predict the next sample based on the information contained in the past samples. A linear regression on a few last samples of the time series is frequently used as a general-purpose function approximator.

Many neural networks, most notably the multilayer feedforward network, can be used as effective nonlinear general-purpose function approximators. A network is simply taught using the history of the time series, and after that the network can be used to predict future outcomes.

In more complex prediction tasks, it may be impossible to construct a model because there either does not exist enough explicit knowledge about the factors that affect the time series, or the knowledge would be too costly to acquire. The general-purpose function approximators may also fail to achieve desired accuracy, or may not be considered reliable in relation to eventual extensive costs associated with wrong decisions. In such cases the methods of *exploratory data analysis*[1] can still be used.

By the methods of exploratory data analysis, the structures in an arbitrary data set, describing different aspects of a phenomenon of interest, can be illustrated. In the task of time series prediction the data set should be chosen to describe aspects of the system *state* that defines the time series, e.g., aspects of the stock markets. When the exploratory data analysis methods are then used to represent the state, the representation can be considered as an automatically produced non-parametric model of the system. In addition to serving as an illustrative model, exploratory data analysis may also be directly used in predicting future outcomes. In prediction, knowledge about the current state of the system may be essential since an optimal predictor may need to be a function of the state.

More generally, the exploratory data analysis methods can be used to illustrate the structures in any statistical data set. If the set describes the state of the same system at different times like in the case of time series analysis, the relations of the states will be illustrated. If, on the other hand, the set describes the states of different systems, the systems will become comparable.

In this study we demonstrate how the self-organizing map,[2,3] a neural network model that has been used in exploratory data analysis, is able to describe structures in a macroeconomic system. The map is shown to illustrate the "welfare states" of the countries of the world, when the data set describes different aspects of the standard of living. State transitions can easily be followed on the map. This study is hoped to serve as a recipe on how, using standard procedures, the state of any micro- or macroeconomic system can be presented in an easily understandable form. Only the data set needs to be changed.

2. Exploratory Data Analysis

The understanding and description of a complex entity like the standard of living requires simultaneous consideration of a large collection of statistical indicators describing its different aspects and their relationships.

The central goal in exploratory data analysis is to present a data set in a form that is easily understandable but that at the same time preserves as much essential information of the original data set as possible. The exploratory data analysis methods are general-purpose instruments that illustrate the essential features of a data set, like its clustering structure and the relations between its data items.

One may distinguish two categories of exploratory data analysis tools with somewhat different goals. First, some tools like the Sammon projection[4] *project* the multidimensional data set to, e.g., a two-dimensional plane, while trying to preserve its whole structure (the distances between the data items) as well as possible. The other methods try to find *clusters*[1] of data, whereby in stead of the large data set only a small number of clusters needs to be considered. Cluster analysis is usually based on the so-called ultrametric distance, defined along hierarchical clustering graphs, whereas direct comparison of the similarity of any two data sets may be misleading.

A vast number of different algorithms to perform clustering is available. Choosing suitable algorithms and applying them correctly requires thorough knowledge of both the algorithms and the data set. There must exist enough clustering tendency in the data set in order that the use of clustering algorithms would be sensible at all, and as different clustering algorithms tend to find clusters of different shapes, the suitability of the shapes to describe the data set must be verified.

The projection methods, on the other hand, do not reduce the amount of data to be presented. Although they illustrate the essential features of the data set, the illustration is costly to obtain and may still be difficult to understand if the data set is large.

The self-organizing map algorithm is a unique method in that it combines the goals of the projection and clustering algorithms. It can be used at the same time to visualize the clusters in a data set, and to represent the set on a two-dimensional map in a manner that preserves the nonlinear relations of the data items; nearby items are located close to each other on the map. Moreover, even if no explicit clusters exist in the data set, the self-organizing mapping method reveals "ridges" and "ravines". The former are open zones with irregular shapes and high clustering tendency, whereas the latter separate data sets that have a different statistical nature.

3. The Self-Organizing Map

The self-organizing map (SOM) [2,3] is an adaptive display method particularly suitable for the representation of structured statistical data. Since the algorithm itself has thoroughly been described in the references and an easily applicable software package SOM_PAK [5] is publicly available, we only briefly mention some of the properties of this mapping that are relevant in exploratory data analysis: (1) The mapping represents a data set in an *ordered* form, whereby mutual similarities of data samples will be visualized as geometric relations of the images of the samples on the map. (2) The natural order inherent in the mapping enables the map to be used as a natural *groundwork*, on which the individual statistical indicators can be visualized as gray levels. The naturally ordered groundwork is more easily comprehensible than bare statistical tables. (3) The structures in the data set can automatically be visualized on the map, whereby the degree of clustering is represented by shades of gray.[6–8] (4) The commonplace problem of *missing data* in statistics can be treated elegantly.[9] The

SOM thus combines the above important goals (1), (2), and (3) of exploratory data analysis.

The principle of the self-organizing map usually relates to a two-dimensional regular array of nodes (here a hexagonal array), in which an adaptive parameter vector $\mathbf{m}_i$, a *model vector*, is associated with every node i. By means of a parallel comparison mechanism, the model vector $\mathbf{m}_c$ that matches best with a certain input sample $\mathbf{x}$ is identified (selected). During computation of the map, the nodes communicate information about this selection in the lateral direction of the array in relation to the interaction strength

$$h_{ij}(t) = h(\mathbf{r}_i, \mathbf{r}_j; t),$$

where h_{ij} is usually called the *neighborhood function*, $\mathbf{r}_i$ and $\mathbf{r}_j$ are the geometric locations of the nodes i and j, respectively, in the array, and t is the discrete-time coordinate; the degree of interaction is often time-variable. In the simplest case, $h_{ij} = h(\|\mathbf{r}_i - \mathbf{r}_j\|;\ t)$, and with increasing distance between i and j and time, usually $h_{ij} \rightarrow 0$. When computation begins, the parametric model vectors $\mathbf{m}_i$ can be initialized with rather arbitrary values $\mathbf{m}_i(0)$. The external input $\mathbf{x} = \mathbf{x}(t) \in R^n$ has some probability density function $p(\mathbf{x})$ and affects all the nodes in parallel.

In the present problem, $\mathbf{x}$ represents a country picked up at random for computation. Its components are defined as values of statistical indicators given for this country. In this problem, $p(\mathbf{x})$ is thus only defined at discrete points in the $\mathbf{x}$ space, i.e., at the indicator vectors, but anyway $\mathbf{x}$ is a *stochastic* variable, because the country is *picked up at random* to the sequence of the $\mathbf{x}$.

In the SOM algorithm, at each step t of the random sequence of the given $\mathbf{x}(t)$ values, the values of the $\mathbf{m}_i$ are gradually and adaptively changed in the following self-organizing (discrete-time) process toward their asymptotic values that depend on $p(\mathbf{x})$:

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + h_{ci}(t)[\mathbf{x}(t) - \mathbf{m}_i(t)]. \tag{1}$$

Here $c = c(\mathbf{x}; \mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_M)$ is the index of the parameter vector $\mathbf{m}_i$ that has the closest vectorial value to $\mathbf{x}$:

$$c = \arg\min_i \{\|\mathbf{x} - \mathbf{m}_i\|\}. \tag{2}$$

The norm is Euclidean. With time, the $\mathbf{m}_i$ then tend to be ordered along with the array in a meaningful way. It can be shown that the point density function of the $\mathbf{m}_i$ tends to approximate $p(\mathbf{x})$ or at least some monotone function of $p(\mathbf{x})$; but in addition, the $\mathbf{m}_i$ attain their values *in an orderly fashion* from the domain of $p(\mathbf{x})$.

4. Structures of Welfare and Poverty in the World

In this section we demonstrate within the framework of a case study how the SOM can be used to aid the understanding of a complex data set. The states of any macro- or microeconomic systems like financial states of companies can similarly be described using exactly the same procedures.

Based on a set of statistical indicators describing different aspects of the standard of living, the SOM can be used to present on a *"welfare map"* the welfare "states" of the countries. Also mappings of countries based on socio-economic variable sets have been presented before,[10,11] but with quite small data sets and only to demonstrate the mapping itself and not its interpretation. We intended to search for *types of welfare and poverty* from the description of the standard of living formed by the SOM, types which are manifested as clustered areas (although not necessarily closed clusters) on the structure diagram. The relations between the types and between different countries are expressed as geometrical relationships on the map, and the types can be characterized using the indicator values displayed on the SOM groundwork. The degree of clustering is also represented in this study by shades of gray.

4.1. The Data Set

The statistical indicators used in analyzing the standard of living were picked up from the World Development Report.[12] The data set was chosen to reflect as many aspects of the standard of living as possible, while trying to avoid indicators only remotely connected with it. A total of 39 indicators were chosen to describe factors like health, education, consumption, and social services.[13] The variance of all indicators was scaled to unity. In general, the variances should be scaled to reflect their relative importance; here, however, no differences in the importance could be assumed.

The indicator most often used as a simple measure of the standard of living, the gross national product (GNP) per capita, was excluded because it would have been difficult to scale this central and strong indicator in relation to the others, and because in this manner it became possible to relate our more indirectly reflected results to the GNP per capita.

4.2. Computation of the Maps

The procedures for obtaining good self-organized maps have been described in the documentation of the SOM_PAK [5] program package. The welfare map used in this study consisted of 13 by 9 units arranged on a two-dimensional hexagonal lattice. The map was chosen from ten different candidate maps, using the average weighted quantization error as the criterion.[13]

All of the chosen indicators were not available for all of the countries. This problem of missing data can be treated by the self-organizing map algorithm simply by considering at each learning step only those indicators that are available.[9] If only a small fraction of the indicators is missing, the remaining indicators contain enough information for the organizing process. For some of the countries a substantial number of the indicators was missing, however, and the indicator sets of such countries may disturb the ordering process. We excluded from the computation of the maps countries having 12 or more of the 39 components missing. After the map was computed these countries could, however, be tentatively mapped on the computed SOM for their rough comparison with the others.

The computed map was used to illustrate three aspects of the welfare data set. First, the countries were mapped on the SOM. The geometrical relations of the countries on the map then illustrated the relations between the standard of living of the countries. Second, the clustering tendency in the data set was illustrated by displaying the distances between the "welfare profiles" of neighboring map locations with gray levels.[6–8] The shades were additionally smoothed to make them more clearly discernible. Third, the individual welfare indicators were displayed as gray levels on the map groundwork.

4.3. Results

The countries of the world were ordered by the SOM as shown in Fig. 1. Some aspects of the distribution of the standard of living in the world can already be inferred from the order of the countries on the map. The order seems to somewhat reflect the geographical organization of the countries, although no geographical information was presented when computing the maps. For example, the European countries occupy the upper left corner of the map, and the right edge consists mainly of African countries. Thus, the fact that the standard of living is correlated with the geography is reflected in the order of the map.

Variations in most of the indicators occur mainly in the horizontal direction, as can be seen by displaying the indicators on the SOM groundwork[13] (examples are shown in Fig. 2). The horizontal dimension of the map thus seems to reflect a kind of an overall standard of living, decreasing from the OECD countries on the left to the poorest African countries on the right. The GNP per capita of the countries also varies predominantly along the horizontal direction together with the general tendency, which can be seen if the GNP is visualized on the SOM groundwork (Fig. 3).

The more specific structures of welfare and poverty are visible as light areas surrounded by dark stripes in Fig. 1. There are few clear-cut clusters but several areas (white "hills" and "ridges" surrounded by black "ravines" in the figure) of a high clustering tendency. These areas correspond to fine structures of welfare

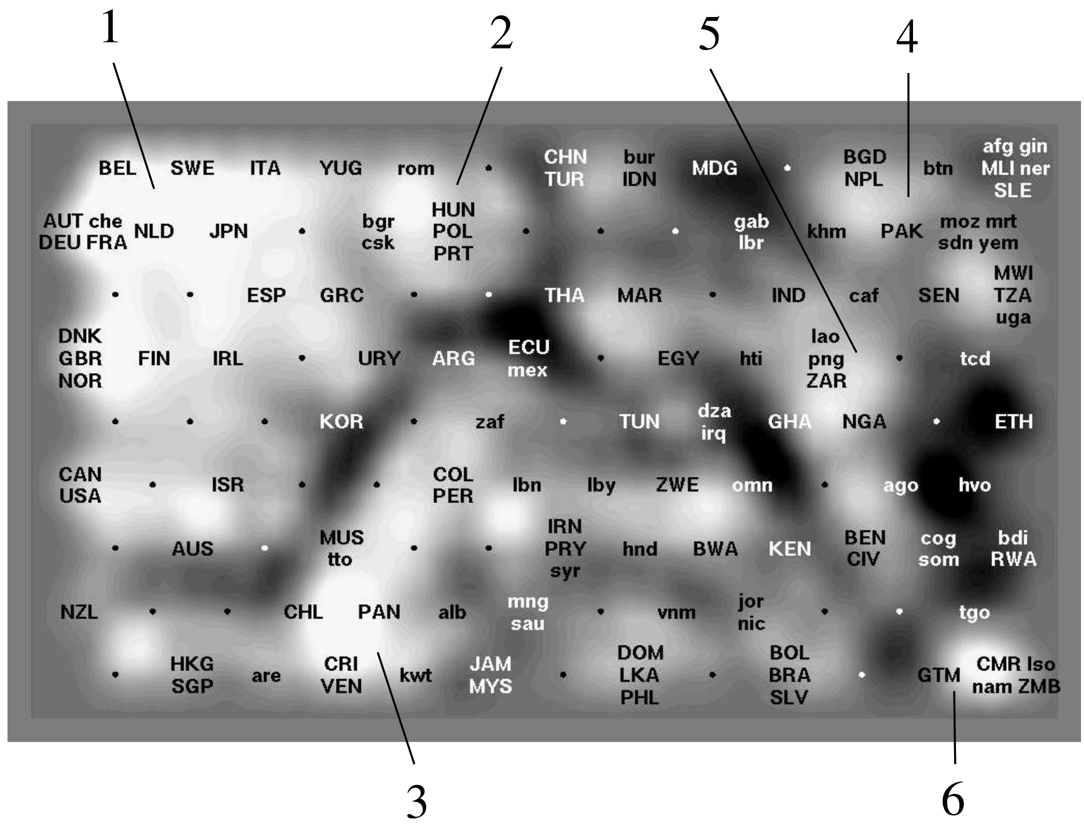

Figure 1: Structured diagram of the data set chosen to describe the standard of living. The order of the abbreviated country names indicates the similarity of the standard of living of the countries, and the shades of gray indicate the degree of clustering. Light areas represent areas of a high degree of clustering and dark areas gaps in the degree of clustering. The countries labeled in lower case were not used in the computation of the map because too many indicators were missing from them. Dots denote map locations which did not correspond to any counties.

and poverty, and the qualities of the fine structures can be characterized by the values of the statistical indicators relating to these areas.

The most clearly distinguishable hill marked with (1) on the left consists mainly of the OECD countries. The countries in this area have the best indicator values (cf. Fig. 2). In the right flank of this area there is a less clearly identifiable hill (2) formed mostly of countries from Eastern Europe. Many countries from Central and South America belong to hill (3), which is clearly separated from hill (1) by a dark “ravine”. The profile of indicator values of the countries in this area is similar to the profiles of the OECD countries, but not as extreme. Near the right edge of the map there are hills consisting mainly of Asian (hill 4) and African (hills 5 and 6) countries. The countries in hill (4) have a welfare

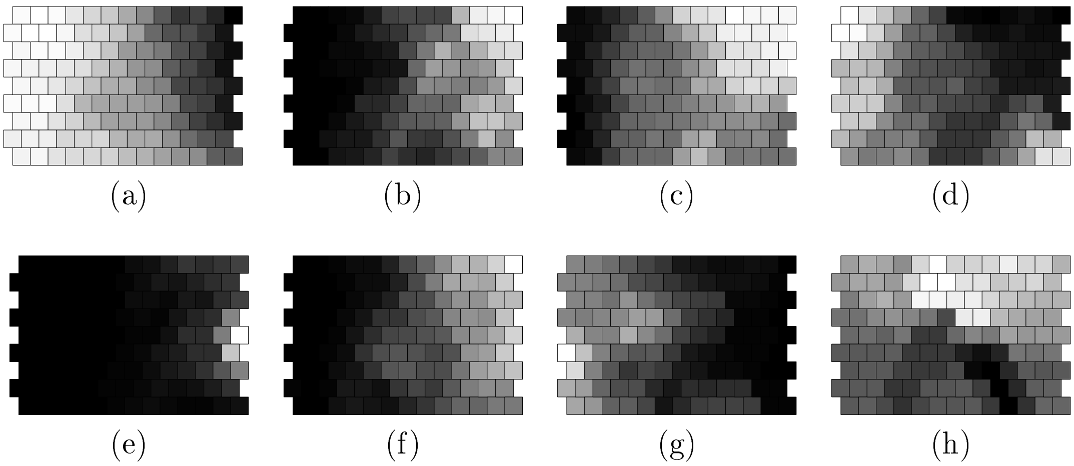

Figure 2: The values of some of the indicators visualized on the SOM groundwork: (a) Life expectancy at birth (years); (b) Adult illiteracy (%); (c) Share of food in household consumption (%); (d) Share of medical care in household consumption (%); (e) Population per physician; (f) Infant mortality rate (per thousand live births); (g) Tertiary education enrollment (% of age group); and (h) Share of the lowest-earning 20 percent in the total household income. In each display, white indicates the largest value and black the smallest, respectively.

profile almost completely opposite to the profile of the OECD countries. Hill (5) is similar to hill (4), although not as extremely opposite to hill (1). In hill (6), school enrollment is higher and illiteracy lower than in hills (4) and (5).

In between the extreme welfare and poverty types in the different ends of the map there are intermediate types consisting of, e.g., Asian and Arabic countries. The specific characteristics of these types can be easily illustrated by referring to indicator values displayed, like in Fig. 2, on the SOM groundwork.

4.4. Discussion

Based on the 39 statistical indicators chosen to describe the standard of living, structures of welfare and poverty were revealed by the self-organizing map. Most of the indicators varied predominantly in the horizontal direction, which thus corresponds to the dimension of "overall welfare." Also the GNP per capita, which was not used as an indicator, was shown to vary along the same dimension. The overall welfare determined by the more qualitative indicators thus also correlates with the GNP per capita.

In addition to the overall welfare dimension, several specific welfare and poverty structures were found by this method. For example, the OECD countries were contained in a large cluster, and the countries of Eastern Europe were mapped to another cluster near to it, however forming a separate cluster.

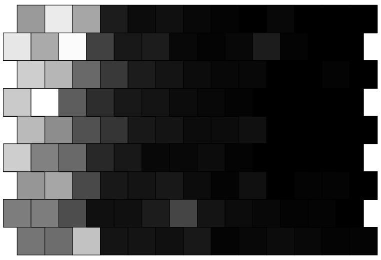

Figure 3: Distribution of the GNP per capita, which was not used in computing the maps, shown over the SOM groundwork. White indicates the largest value in the material and black the smallest, respectively; the map locations that did not represent any countries were set to the average value of the neighbor locations.

5. Conclusions

We presented in this article a case study that demonstrated how the SOM can be used to visualize a large statistical data set. The set consisted of 39 indicators, which described different aspects of the welfare "states" of the countries of the world.

Following exactly the same procedures the SOM could also be used as a decision-support system for analyzing and visualizing sets of statistical indicators in other similar applications. For instance, the method has already been used for the analysis of states of banks.[14] The SOM formed a "solvency map," from which the state of the banks could be inferred at a glance. In time series analysis it is important that the nature of change in the state of the banks can be visualized on the map (e.g., as a slow shift toward the bankrupt region) even if the changes could not be predicted by the traditional methods.

In addition to using the SOM to visualize data sets, the capability of the algorithm to represent states of the system that produces the data set can be used in classifying the states. The SOM has in fact been used as a self-organizing pattern recognizer in many engineering tasks, and also as a part of a system that trades stocks.[15]

References

1. A. K. Jain and R. C. Dubes, *Algorithms for clustering data* (Prentice Hall, Englewood Cliffs, NJ, 1988).
2. T. Kohonen, *Biol. Cybern.* **43** (1982) 59.
3. T. Kohonen, *Self-Organizing Maps* (Springer, Berlin, 1995).
4. J. W. Sammon, Jr., *IEEE Tr. Computers*, **C-18** (1969) 401.
5. T. Kohonen, J. Hynninen, J. Kangas, and J. Laaksonen, *SOM_PAK: The self-organizing map program package* (obtainable via anonymous ftp from

the internet address "cochlea.hut.fi" (130.233.168.48), 1995).

6. J. Iivarinen, T. Kohonen, J. Kangas, and S. Kaski, in *Proc. Conf. on Artificial Intelligence Res. in Finland*, eds. C. Carlsson, T. Järvi, and T. Reponen (Finnish Artificial Intelligence Society, Helsinki, Finland, 1994), p. 122.
7. M. A. Kraaijveld, J. Mao, and A. K. Jain, in *Proc. 11th Int. Conf. Pattern Recognition* (IEEE Comput. Soc. Press, Los Alamitos, CA), p. 41.
8. A. Ultsch and H. P. Siemon, in *Int. Neural Network Conf.* (Kluwer, Dordrecth, 1990), p. 305.
9. T. Samad and S. A. Harp, *Network* **3** (1992) 205.
10. F. Blayo and P. Demartines, *Bull. des Schweizerischen Elektrotechnischen Vereins & des Verbandes Schweizerischer Elektrizitätswerke* **83** (1992) 23.
11. A. Varfis, in *NATO ASI on Statistics and Neural Networks* (1993).
12. World Bank, *World Development Report 1992* (Oxford Univ. Press, New York, NY, 1992).
13. S. Kaski and T. Kohonen, Tech. Rep. A24, (Helsinki Univ. of Technology, Laboratory of Computer and Information Science, Espoo, Finland, 1995).
14. B. Martín-del-Brío and C. Serrano-Cinca, *Neural Comput. & Applic.*, **1** (1993) 193.
15. C. L. Wilson, in *Proc. IEEE Int. Conf. Neural Networks* (IEEE Service Center, Piscataway, NJ, 1994), p. 3651.