Int. J. Mach. Learn. & Cyber. (2012) 3:233–245
DOI 10.1007/s13042-011-0057-5

ORIGINAL ARTICLE

Visual tracking of the millennium development goals with a fuzzified self-organizing neural network

Peter Sarlin

Received: 27 July 2011 / Accepted: 27 October 2011 / Published online: 15 November 2011

Abstract This paper uses the self-organizing map (SOM), a neural network-based projection and clustering technique, for monitoring the millennium development goals (MDGs). The eight MDGs represent commitments to reduce poverty and hunger, and to tackle ill-health, gender inequality, lack of education, lack of access to clean water and environmental degradation by 2015. This paper presents a SOM model for cross sectional and temporal visual benchmarking of countries and pairs the map with a geospatial dimension by mapping the clustering onto a geographic map. The temporal monitoring is facilitated by fuzzifying the second-level clustering with membership degrees. By creating an MDG index, and associating the SOM model with it, the model enables cross sectional and temporal analysis of the overall MDG progress of countries or regions. Further, the SOM model enables analysis of country-specific as well as regional performance according to a user-specified level of aggregation. The result of this paper is an MDG map for visual tracking and monitoring of the progress of MDG indicators.

Keywords Self-organizing maps · Millennium development goals · Projection · Clustering · Geospatial visualization

1 Introduction

During the largest-ever gathering of world leaders in September 2000 in New York, the Millennium Declaration was adopted. The Declaration, endorsed by all 189 United Nations member countries, was later translated into eight millennium development goals (MDGs). The MDGs represent commitments to reduce poverty and hunger, and to tackle ill-health, gender inequality, lack of education, lack of access to clean water and environmental degradation by 2015. The eight goals are defined as follows.

Goal 1 Eradicate extreme poverty and hunger
Goal 2 Achieve universal primary education
Goal 3 Promote gender equality and empower women
Goal 4 Reduce child mortality
Goal 5 Improve maternal health
Goal 6 Combat HIV/AIDS, malaria and other diseases
Goal 7 Ensure environmental sustainability
Goal 8 Develop a global partnership for development

The goals of the Millennium Declaration are tracked using 21 concrete, numerical benchmark targets. For example, halve the proportion of people whose income is less than $1/day and reduce by three quarters the maternal mortality ratio. The progress of these 21 targets is measured using 60 quantifiable statistical indicators, such as proportion of population below $1/day and maternal mortality ratio. With only a few years to go before the deadline, tracking the progress towards the goals is indeed of central importance. However, the tracking has mainly concerned measuring the current state or linear and log-linear projections into the future (see Sahn and Stifel [24] and UNECOSOC [35], for example). Using these types of methods, thorough and simultaneous comparisons of all the MDG dimensions over time and across countries is impossible. A composite index, such as the Human Development Index (HDI) [34], could be constructed of the MDGs for comparing the general progress over time and across countries. Lately, further developed poverty

P. Sarlin (✉)
Department of Information Technologies,
Turku Centre for Computer Science,
Åbo Akademi University, Joukahaisenkatu 3–5,
20520 Turku, Finland
e-mail: psarlin@abo.fi

indices have been created; the 'improved' HDI [18] and the Multidimensional Poverty Index (MPI) [1]. However, Ravallion [20] shows that mashup indices of development come with several disadvantages, not least the ambiguity in deriving the index and the impossibility of simultaneous in-depth analysis of the individual elements.[1] For objective and precise assessment of countries in comparison with each other, such as in the case of the MDGs, one has to utilize multidimensional datasets and methods that account for the complexity of the problem.

By examining raw statistical tables, we can analyze one measure for a handful of countries at a given point in time. Likewise, two- and three-dimensional visualization tools of ordinary spreadsheet programs have limited capabilities for higher dimensions. The analysis becomes even more cumbersome when the temporal dimension is included. To date, visual monitoring of the MDGs has to some extent been neglected. United Nations has co-developed three visualization tools for tracking the progress of the MDG indicators: MDG Monitor, MDG Mapper and DevInfo. While MDG Monitor and MDG Mapper enable visualization of one indicator at a time on a geographic heat map, the DevInfo enables a number of different visual representations of global to regional distributions of each indicator. They do not, however, enable features of exploratory data analysis (EDA) [31], such as projection of multidimensional data onto a two-dimensional plane and illustration of the structures in a dataset. When attempting analysis of multidimensional data, such as statistical indicators for monitoring the MDGs, methods of EDA are feasible. EDA attempts to describe different aspects of the phenomena of interest in an easily understandable form by illustrating the structures in a dataset, but at the same time preserving information of the original data. The self-organizing map (SOM) [13, 14] is a neural network-based EDA tool. While most neural networks utilize supervised learning for classification tasks, such as Boehme et al. [2] and Tong and Mintram [30], the SOM is an unsupervised two-layered neural network that elegantly combines the goals of projection and clustering techniques. The visual representation of the SOM is not only created by mimicking the functioning of the human brain, as neural networks in general do, but also by enabling utilization of the pattern recognition capabilities of the human brain when interpreting the visualizations.

Though being a common method in economic and financial analysis, such as Resta [21], Eklund et al. [7] and Sarlin and Peltonen [29], the SOM has not been used frequently for tracking development, especially not the MDGs. Using the SOM, Kaski and Kohonen [11] represent welfare states of countries on a "welfare map", while Naq and Mitra [16] find patterns in the recent socioeconomic development and poverty on a SOM. Collan et al. [3] use social and economic indicators for transition countries, and two European benchmarks, to show differences in the state of transition on a SOM. Lately, Tyler and Gopal [32] have examined patterns among Sub-Saharan African countries using PCA and the SOM. They conclude that the techniques are highly effective to compress multidimensional qualitative and quantitative data. However, these studies do none of the following: implement the SOM as a general, global and visual monitoring system; target specifically indicators of the MDGs; pair the SOM with a fuzzified second-level clustering; pair the SOM with geospatial data; or pair the SOM with a development index. The model should be as global as possible, since the state of Sub-Saharan African countries in relation to the rest of the world is of central importance. More specifically, this paper attempts a comprehensive study of MDG indicators and has its main focus on the visualization capability of the SOM. The main aim of the paper is to create a SOM-based tool for visual monitoring of a global dataset consisting of MDG indicators. This type of analysis could, of course, be applied on other types of high dimensional panel datasets. The dataset includes 14 MDG indicators and 232 countries, and spans over 1990–2008. Firstly, the SOM will be used for two types of analyses of the MDG indicators: for visual benchmarking of countries and for visual analysis of the evolution of a country. Likewise, a combination of the above analyses enables visual benchmarking over time. Following Sarlin and Eklund [27], the SOM representation is enhanced by a fuzzified clustering. It is an aid for analyzing the temporal dimension, as the cluster centers express the representative MDG states for the countries, while the fluctuations of the fuzzified membership degrees represent their variations over time. It also enhances the representation of the crispness of data by not only showing its best-matching unit on the SOM grid, but by also expressing the resemblance to all clusters. Further, a fuzzification of the SOM units enables analyzing the distance structure on the SOM as well as its topological ordering. We also create an MDG index which associates a mashup state to each data point. This is beneficial, and sometimes even necessary, for assessing MDG directions on the map. These analyses are not only performed on samples of the panel dataset, cross-sectional and time-series data, but also on population weighted aggregates for geographic regions, such as Sub-Saharan Africa and the world. Secondly, we project for each country the color code of its corresponding cluster on a geographic map. This enables visual analysis of multidimensional MDGs on a

[1] The HDI, for example, has been criticized for the way its component indices are derived by the raw data (see Noorbakhsh [17]) and the additivity of the aggregation method (see Sagar and Najam [23]).

geographic map. The visual explorations in this paper illustrate the usefulness of the SOM for visual tracking of MDG indicators.

The paper is organized as follows. The first part introduces the methodology used for tracking the progress of MDGs. First, the SOM and its fuzzified clustering is introduced, and then the indicators are discussed. The second part explains the construction of the SOM model and defines the clusters on the map. The third part shows visual analyses using the SOM model and a geospatial mapping. Finally, the fourth part concludes by presenting the key findings and recommendations for future research.

2 Methodology

There exist two distinguishable categories of multivariate EDA methods: multidimensional scaling (MDS) and clustering methods. MDS methods, and its many variants [5], attempt to preserve the whole structure of the dataset, while project multidimensional data to an easily interpretable format, such as a two-dimensional plane. The MDS methods do not, however, reduce the amount of presented data. The clustering techniques, on the other hand, attempt to find mean profiles in the data. Thus, they reduce the amount of data by enabling analysis of a small number of clusters instead of the whole dataset. The vast variety of clustering algorithms requires thorough knowledge of both the dataset and the algorithms, since there must exist enough clustering tendency for clustering to be feasible and different algorithms are suitable for datasets of different shape.

By combining the objectives of both categories for EDA, the SOM performs simultaneously a projection and a clustering. On the one hand, the SOM projection differs from other projection techniques, such as MDS, by focusing on preserving the neighborhood relations, instead of trying to preserve the distances between data. Further, rather than projecting data into a continuous space, such as MDS methods, the SOM uses a grid of neurons onto which data are projected. Thus, it enables visualization of multidimensional data on a two-dimensional plane consisting of neurons with preserved neighborhood relations. On the other hand, the two-level clustering of the SOM, i.e., separation of data into neurons and neurons into clusters as proposed by Vesanto and Alhoniemi [36], differs from other statistical clustering techniques. They assert that the SOM differs from other clustering methods in terms of computational cost and by being more robust on data that are non-normally distributed. Further, user satisfaction with the SOM and its information products has, by end-users within the domain of financial analysis, been evaluated as superior than currently used methods (e.g. [7]). This indicates possible achievement of positive results in tracking MDGs.

2.1 Self-organizing maps

The SOM is a non-linear and non-parametric artificial neural network utilizing an unsupervised, competitive learning method developed by Kohonen [13]. The SOM may be thought of as a projection maintaining the neighborhood relations in the data [10] or as a spatially constrained form of k-means clustering [22]. The network of neurons consists solely of two layers: the input and the output layer. The number of neurons in the input layer equals the dimensions in the data, while the output layer is a two-dimensional topological grid. Each neuron on the topological grid has its own reference vector. During the course of training, while each neuron learns to attract data with similar characteristics, all neighboring neurons learn, with diminishing weight, to attract similar data. Hence, the map is ordered as to the characteristics of the dataset. The neurons of the map can further be divided into clusters of similar neurons, enabling a simultaneous two-level clustering. The SOM algorithm is described here briefly—for details, see Kohonen [14].

For its superior visual features, the software Viscovery SOMine 5.0 is used in this paper.[2] It employs the batch training algorithm; a slightly different version of the basic SOM algorithm. Instead of processing the data vectors sequentially, the batch algorithm differs by processing all the data vectors simultaneously. The most important advantage of the batch algorithm is the reduction of computational cost. The training process starts with an initialization of the reference vectors. Instead of random initialization, the reference vectors are set in the same direction as the principal components of the input data using principal component analysis (PCA). This reduces the computational cost and enables reproduction of the results. The training algorithm has two steps: (1) finding the best-matching units (BMUs) and (2) adjusting the reference vectors. The second step includes both the adjustment of the reference vectors of the BMUs and the reference vectors in a specified neighborhood of the BMUs. The steps are repeated for a specified number of iterations.

In the first step, the algorithm compares, using the Euclidean distance, each input data vector x with the network's reference vectors m_i to find the best match m_c,

$$\|x - m_c\| = \min_i \|x - m_i\| \quad (1)$$

such that the distance between the input data vector x and the winning reference vector m_c is less than or equal the distance between input x and any other reference vector m_i. During the first step, all the input vectors are presented to the map.

[2] For a thorough discussion of the software, see Deboeck [6].

In the second step, each reference vector m_i is adjusted using the equation for the batch algorithm:

$$m_i(t+1) = \frac{\sum_{j=1}^{N} h_{ic(j)}(t) x_j}{\sum_{j=1}^{N} h_{ic(j)}(t)}, \tag{2}$$

where $h_{ic(j)}(t)$ is a weight that represents the value of the neighborhood function defined for the node i in the BMU $c(j)$ at time t. The index j indicates the input data vectors that belong to the neuron c, and N is the number of these vectors. The function $h_{ic(j)} \in (0, 1]$ is defined as a Gaussian function:

$$h_{ic(j)} = \exp\left(-\frac{\|r_c - r_i\|^2}{2\sigma^2(t)}\right), \tag{3}$$

where r_c and r_i are two-dimensional coordinates of the reference vectors m_c and m_i, respectively, and the radius of the neighborhood $\sigma(t)$ is a monotonically decreasing function of time t. Furthermore, the second-level clustering is done by combining the local ordering information with agglomerative hierarchical clustering, i.e. iteratively agglomerating the closest clusters based on some similarity measure. The following Ward's [41] criterion is used as a basis for measuring the distance between two candidate clusters:

$$D_{kl} = \frac{n_k n_l}{n_k + n_l} \cdot \|c_k - c_l\|^2, \tag{4}$$

where n_k and n_l represent the cardinalities and c_k and c_l the cluster centers of clusters k and l, respectively. However, the algorithm limits the merging of clusters to topologically neighboring clusters by defining the distance between non-adjacent clusters as infinitely large. We use a heuristic cluster indicator for choosing the number of clusters. The indicator starts with each node being its own cluster and computes the Ward distance D_{kl} for all possible numbers of clusters (1,2,…,M). Then the indicator is the ratio of two consecutive distances normalized by an exponential function. Thus, let d represent the number of clusters, then the indicator finds a d that results in a high $D_{kl}(d)$ (distance from d to $d - 1$ clusters) and a low $D_{kl}(d + 1)$, since the next merge ($d - 1$ clusters) would cause higher variance within the clusters.

In addition to dividing the nodes into crisp clusters, as is commonly done, the SOM has been fuzzified using a distance-based approach in Sarlin and Eklund [27] and using fuzzy C-means clustering in Sarlin and Eklund [28]. We follow the approach in Sarlin and Eklund [27] by presenting a fuzzy membership degree using Euclidean distances between data points and the centroids of the clusters. For the cluster centroids, any crisp clustering method, as appropriate for the task in question, is applicable. As we also implement the fuzzification on the nodes, it can as well be used for assessing the topological ordering of the grid. The crisp clustering is fuzzified by computing the inverse distance between data point x_j (or each reference vector m_i) and each cluster center c_k (where $k = 1,2,\ldots,C$):

$$u_{jk} = \frac{1}{\left\|x_j - c_k\right\|^{\frac{2}{\mu-1}}}, \tag{5}$$

where $\mu \in (1, \infty)$ is the fuzzy exponent (i.e., exponential weight) which controls the extent of overlapping between the clusters. However, we normalize the similarity matrix u_{jk} to the following matrix of cluster membership degrees for each node:

$$v_{jk} = \frac{u_{jk}}{\sum_{k=1}^{C} u_{jk}}, \tag{6}$$

to fulfill the probabilistic constraint $\sum_{k=1}^{C} v_{jk} = 1$. The extent of overlapping between the clusters is set by the fuzzy exponent μ. When $\mu \to 1$, the fuzzy clustering converges to a crisp clustering, while when $\mu \to \infty$ the cluster centers tend towards the center of the dataset. $\mu = 2$ and $\mu = 3$ can be seen as benchmarks, since they give squared and simple Euclidean distances. The crisp clusters on the map are shown by contours, while each nodes' cluster membership degrees are visualized on own feature planes for each cluster.

A so-called tension value is specified as a measure of the radius of the neighborhood $\sigma(t) \in [0, 2]$. The tension controls the size of the neighborhood around the BMU that is influenced during the training process. A rule of thumb is that large neighborhood radii, i.e., tension values, result in stiff maps that stress topological ordering at the cost of quantization accuracy [37]. The rest of the parameters in SOMine are the following: map size (the number of neurons), map format (the ratio of x and y dimensions), and the length of training (training cycles). The quality of the map is measured in terms of quantization error and distortion [37]. The average quantization error represents the fitting of the neural map to the data measured by an average of the distances between all input vectors x_j and their corresponding best matching reference vectors m_c, i.e.,

$$qe = \frac{1}{N}\sum_{j=1}^{N} \left\|x_j - m_{c(j)}\right\|. \tag{7}$$

The normalized distortion measures the fit of the map with respect to both the shape of the data distribution and the radius of the neighborhood, and is computed as follows.

$$DM = \frac{1}{N} \frac{\sum_{j=1}^{N}\sum_{i=1}^{M} h_{ic(j)} \left\|x_j - m_i\right\|^2}{\sum_{j=1}^{N}\sum_{i=1}^{M} h_{ic(j)} \Big/ M} \tag{8}$$

where M is the number of reference vectors.

The output of the SOM algorithm may be presented and utilized in different ways. For the purpose of this analysis,

the results are visualized using a two-dimensional grid. As the reference vectors of this two-dimensional grid represent a multidimensional space, we can visualize each element of the vector on an own grid, namely a feature plane. The feature planes represent graphically, for each corresponding variable, the distribution of the variable values on the two-dimensional map. As the feature planes are different views of the same map, one unique point represents the same node on all planes. Thereby, the characteristics of the SOM model can be identified by studying the underlying feature planes. The values on the feature planes may be represented by different color schemes (e.g. colored, grayscale and black and white). In this paper, the feature planes are produced in color, where low to high values are represented by cold to warm scales. The color scales are shown below each corresponding feature plane.

2.2 Choice of countries and indicators

The set of countries is chosen to be as large and complete as possible, where the term country refers, as appropriate, to territories or areas since no opinion is expressed on the legal status of any region. A thorough set of countries enables exploration of possibly divergent countries. Since enabling analysis of countries in Sub-Saharan Africa—African countries without a Mediterranean coastline—is the main target of this paper, it is important to include the other states of MDGs to enable temporal analysis into various new states. Thus, although the main focus is on Africa, the map should be interpreted as a global 'MDG map', enabling analysis of the state in any country. In fact, if such data are available, sub-national, provincial or district level data may also be projected onto the map. However, the dataset used in this study includes 232 countries for the period from 1990–2008.

The set of indicators is chosen to measure the progress towards the MDGs. Prennushi et al. [19] divides indicators into two groups: final and intermediate. Final indicators measure effects of an intervention on the well-being of individuals, while intermediate indicators measure factors that determine, or contribute to the process of achieving, a final indicator. Since the objective of this model is to represent the stage of progress towards the MDGs, final indicators are chosen for analysis. Thus, this model does not concern the fact that final indicators are the result of several factors. Moreover, although it would be computation-wise feasible to include all 60 indicators mentioned in the Millennium Declaration, this study uses fewer variables to measure the goals. This incorporates a high share of the same information and enables a deeper visual analysis and presentation within the constraints of a journal paper. The choice of the analyzed indicators is based on the utilized subset in MDG Monitor of United Nations. The dataset includes thereby 15 indicators for tracking the progress of all eight MDGs.

The dataset was collected from the UNSD (United Nations Statistics Division) database (http://mdgs.un.org/unsd/mdg/). Further information on the transformations and the precise definitions of variables can be found in the Handbook for Monitoring MDGs [33]. To sum up, the dataset used in this study includes annual data of 15 indicators for 232 countries for the period from 1990–2008, enabling up-to-date classification of the wide variety of countries in the world. The indicators, their relation to the MDGs, and their summary statistics are shown in Table 1. Further, an MDG index is created by combining the indicators to a composite variable. First, indicators 2, 6, 9, 10 and 13 (as per the first column of Table 1) were multiplied by -1, so that an increase in the variable indicates an achievement of the particular MDG. Then, for standardizing the contribution of the variables, each indicator is normalized by its own range. Finally, the index is defined as the average of the normalized indicators for each data point x_j.

As discussed by Coudouel et al. [4], poverty comparisons between groups and over time face obstacles in transformations. Poverty comparisons across countries are difficult, since both absolute and relative prices of goods and services differ across countries. That is, although differences in purchasing power parity (PPP) are considered, cross-country comparisons rely on the assumption that the measures are homogeneous across countries. Likewise, poverty comparisons over time require measures that reflect differences over time in the cost of living across regions. This can be done by, for example, converting data from different regions into real measures by deflating the indicators in space and time. However, in practice the data are never fully comparable over time or across regions. Moreover, since data are partly provided by national agencies, there might exist erroneousness. We do not, however, see any of these discrepancies as an obstacle for visual monitoring of development indicators using the SOM—in fact, it motivates visual, rough monitoring using geometrical neighborhood relationships, instead of precise modeling.

3 The SOM model

This section describes the training of the SOM model and its clusters.

3.1 Training the SOM model

The data have been normalized for equal weighting of the indicators. They have been standardized by variance and

Table 1 Summary statistics

No.	Indicator	MDG	Abbreviation	OBS	Mean	Std.Dev	Min	Max	Scaling
1	Population below $1/day, (%)	1	1a. Poverty	294	17.40	20.93	2.00	88.50	Variance
2	Underweight children under 5 years, (%)	1	1b. Poverty	292	17.53	12.76	0.60	48.70	Variance
3	Primary education enrolment ratio	2	2. Education	1204	88.35	15.02	25.90	100.00	Variance
4	Seats held by women in parliament, (%)	3	3a. Women	1535	14.07	9.83	0.00	48.80	Variance
5	Gender Parity Index in primary level enrolment	3	3b. Women	1417	0.95	0.08	0.00	1.30	Range
6	Child mortality rate per 1,000 births	4	4a. Child mortality	730	58.75	62.06	3.00	305.00	Variance
7	Children immunized against measles, (%)	4	4b. Child mortality	1706	83.50	15.89	20.00	99.00	Variance
8	Maternal mortality ratio per 100,000 births	5	5. Maternal health	169	322.37	421.09	1.00	2100.00	Variance
9	Tuberculosis rate per 100,000 population	6	6a. Disease	1708	165.97	221.88	0.00	1400.00	Variance
10	People (15–49 years old) living with HIV, (%)	6	6b. Disease	251	2.14	4.73	0.10	26.50	Variance
11	Proportion of land area covered by forest, (%)	7	7a. Environment	526	30.90	23.65	0.00	98.90	Variance
12	Access to improved drinking water sources, (%)	7	7b. Environment	680	83.29	18.92	17.00	100.00	Variance
13	Metric tons of CO_2 emissions per capita	7	7c. Environment	1556	4.69	6.14	0.00	64.17	Range
14	Official development assistance (ODA) to GNI, (%)	8	8a. Development	290	0.43	0.26	0.10	1.17	Variance
15	Internet users per 100 population	8	8b. Development	1700	14.59	20.80	0.00	90.56	Variance

range as shown in the last column of Table 1. If the range of an indicator is smaller than 8 times its standard deviation, it is scaled by variance; otherwise, the indicator is scaled by range. All data rows are not, however, used for training due to missing values. The SOM overcomes a small share of missing values by solely considering the indicators that are available, giving enough information for the organization process [25]. However, a high share of missing values may disturb the organization of the map [11]. Thus, row vectors with 7 or more missing values have been excluded from the computation of the map. The excluded data vectors can, however, be tentatively projected onto the map for rough visual analysis.

The constructed map is trained using data from 1990–2010. The dataset consists of 1,708 row vectors with a dimensionality of 15—one dimension for each explanatory variable. During the course of the experiment, hundreds of maps were trained using different parameter values (tension, cycles of training, number of clusters and map format). As noted by Kohonen [14], the selection of the parameters is not crucial if the map size is less than a few hundred nodes. As there, of course, still are some differences, the choice of the map was based upon its quantization error, distortion measure, topological ordering, visual cluster structure and interpretability. Quantization error and the distortion measure are those presented in Sect. 2.1, while the topological ordering is evaluated following Kaski et al. [12]. We project the reference vectors into two- and three-dimensional spaces using Sammon's mapping [26] and assess the map in terms of twistedness and adjacent non-neighbors in the Euclidean space. The visual cluster structure is assessed using a U-matrix, on which separable second-level clusters are revealed, and interpretability is a subjective measure of the visualization defined by the analyst. During the experimental stage of training, we vary radius of the neighbourhood σ, number of nodes M and map format (ratio of X and Y dimensions). While the number of second-level clusters is set using the cluster indicator, we keep the map ratio as 100:75. This rectangular map format approximates the ratio between the two principal components in data, which Kohonen [14] proposes as a selection criterion for achieving a stable orientation in the data space.

After an extensive training process, a neural network with 15 neurons in the input layer and 85 output neurons ordered on a map of the size 10×9 was chosen. Since the main purpose of this study is visualization, not classification, a comparatively large map is preferred over a small map [6]. The map was trained with a tension of 1.5 (where $\sigma(t) \in (0, 2])$ for 4 epochs, leading to an average quantization error of 0.51 and a distortion measure of 0.72. A high tension is chosen, since for visualization purposes precise topological ordering is preferred over high quantization accuracy. The two-dimensional topological grid is shown in Fig. 1.

3.2 Defining the clusters

The map is further cut into layers of column vectors—development indicators in this study. These feature planes are shown in Fig. 2. To further distinguish the nodes of the map, it is clustered using hierarchical clustering in

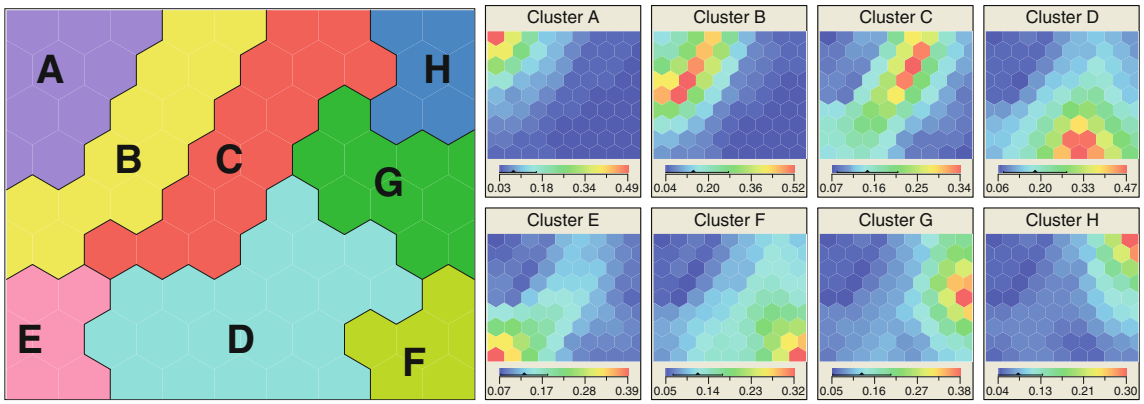

Fig. 1 The SOM grid (*left*) and membership degrees of each node to each cluster (*right*) (color figure online)

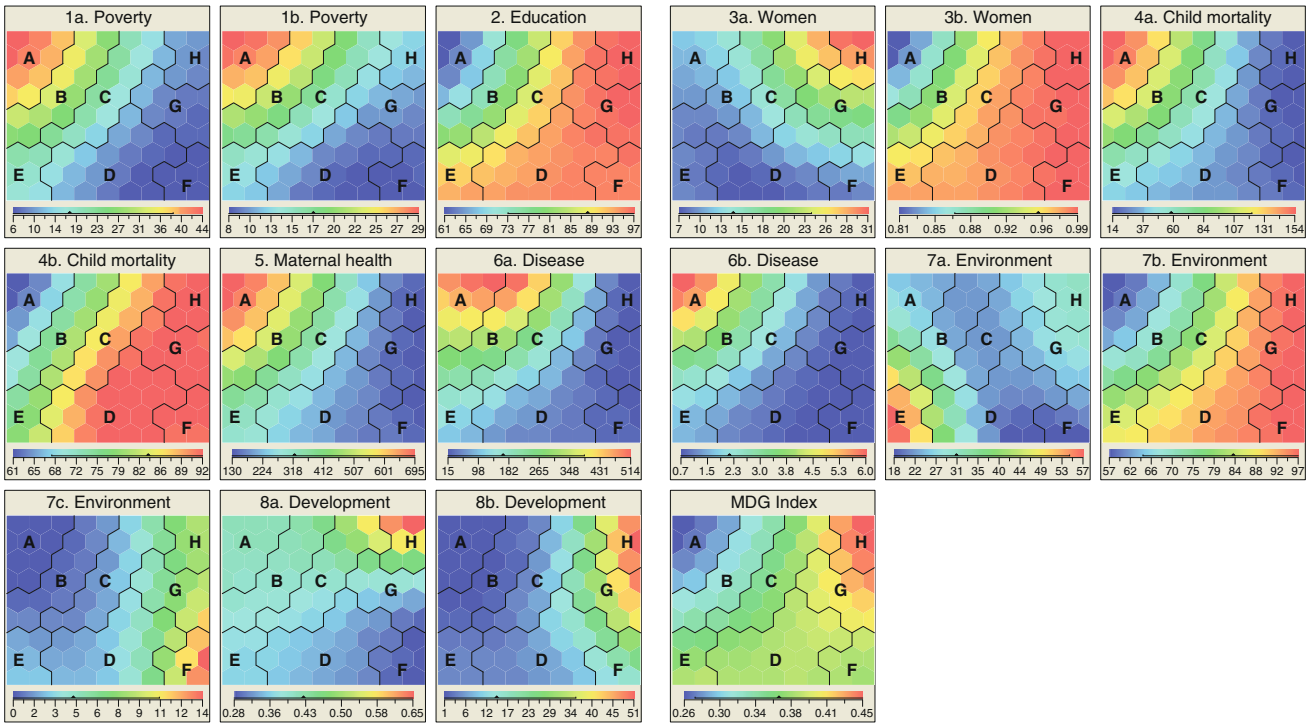

Fig. 2 The feature planes of the SOM grid (color figure online)

conjunction with local ordering information. The clustering is performed with respect to all the explanatory variables and the cluster indicator described in Sect. 2.1 is used for choosing the number of clusters to be eight. The clustering of the two-dimensional map is shown on the left of Fig. 1, while the fuzzification of the clustering is shown on the right. The fuzzification shows the membership of each node to each cluster. The fuzzifier μ was tested for values between 1 and 5. A μ value of 2.0 provided an adequate fuzzification of the map. Without completely eliminating cluster borders, it introduces a fuzziness degree large enough to show relationships between clusters. To facilitate the understanding of the map, below follows a description of the clusters. A more detailed description of the clusters is given in Table 2.

Clusters A, B, C and D represent, in a descending order, the least developed economies, where cluster A is characterized by the worst conditions. Cluster E resembles mainly cluster C and D, but differs from the rest of the clusters by having a high proportion of land area covered by forest. On the other hand, clusters F, G and H represent, in an ascending order, the most developed countries. Cluster F is characterized by high CO_2 emissions, while cluster H is characterized by a high proportion of women in

Table 2 Cluster descriptions

Indicator	Cluster A		Cluster B		Cluster C		Cluster D		Cluster E		Cluster F		Cluster G		Cluster H	
	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)
1a. poverty	57.4	H–VH	32.1	A–H	17.7	L–A	6.3	VL–L	13.5	L	2.3	VL	3.1	VL	2.4	VL
1b. Poverty	31.5	H–VH	24.8	A–H	16.1	L–A	5.8	VL–L	10.4	L	5.0	VL	5.9	VL–L	3.9	L
2. Education	60.7	VL–L	79.1	A	91.6	H–VH	94.5	H–VH	94.4	H	94.3	VH	96.4	VH	97.3	VH
3a. Women	11.4	VL–L	13.0	L–A	18.1	L–H	9.1	VL–L	10.0	VL	8.2	VL–L	20.6	A	32.8	H–VH
3b. Women	0.8	VL–L	0.9	A	1.0	H	1.0	H–VH	1.0	H–VH	1.0	VH	1.0	VH	1.0	VH
4a. Childmortality	161.3	VH	100.5	A	51.7	L	28.4	VL–L	37.7	H	13.5	VL	9.7	VL	5.9	VL
4b. Childmortality	61.8	VL–L	74.9	L–A	87.8	H	91.9	H–VH	77.2	A	91.4	VH	91.9	VH	92.8	VH
5. Maternalhealth	998.0	H–VH	543.0	A–H	237.0	L–A	105.0	VL–L	249.0	L–A	38.0	VL	17.0	VL	11.0	VL
6a. Disease	462.0	H–VH	386.0	A–VH	200.0	L–A	52.0	VL–L	119.0	L	29.0	VL	20.0	VL	8.0	VL
6b. Disease	7.8	H–VH	2.0	A	1.1	L	0.4	VL–L	0.4	L	0.5	VL	0.4	VL	0.2	VL
7a. Environment	27.5	L	32.6	L–A	15.3	L	19.7	VL–A	66.4	H–VH	16.1	VL	34.4	L	34.9	L–A
7b. Environment	55.8	VL	68.8	L–A	83.2	A–H	92.6	H–VH	87.7	A–H	98.6	VH	98.5	H–VH	99.4	VH
7c. Environment	0.4	VL	1.0	VL	2.1	L	3.3	L–A	3.4	L	16.4	A–VH	8.4	A–H	9.5	A
8a. Development	0.0	L	0.0	L	0.6	L–A	0.3	VL–L	0.3	L	0.2	VL–L	0.3	L–A	0.8	A–VH
8b. Development	1.0	VL	2.0	VL	5.8	L–A	8.9	L–A	6.2	VL	15.4	L–A	43.4	A–VH	51.8	H–VH
MDG index	0.3	VL	0.3	L	0.4	L–A	0.4	A–H	0.4	A	0.4	A–H	0.4	H–VH	0.5	VH

VL very low, *L* low, *A* average, *H* high, and *VH* very high

(a) is the cluster center per indicator and cluster

(b) is a measure of the range of an indicator in each cluster

parliament, official development assistance (ODA) and internet users. Obviously, the MDG index shows the lowest values for the least developed and highest for the most developed nations. An interesting finding is, however, that from the MDG perspective the states are approximately equally good in clusters C, D, E and F, indicating that extreme environmental factors counterbalance other welfare indicators.

4 Visual tracking of the MDGs

This section presents visual analyses using the above created MDG map. The map is used for visual benchmarking of countries on a given point in time, for projection of time-series data onto the map and for combining these two tasks by conducting benchmarking over time. Further, we also show a geospatial plot of the cluster color codes at one point in time. Data points are mapped onto the grid using Eq. 1, i.e. to their BMU, and consecutive time-series data are linked with lines, i.e. trajectories.

4.1 Visual benchmarking of countries' development level

In Fig. 3, a cross section for a representative subset of the countries in year 2005 are projected onto the MDG map. Year 2005 is chosen since its missing-value percentage is the lowest. Out of all 232 countries, a sample of countries representing the heterogeneity of the map is shown. This illustration facilitates simultaneously the interpretation of the characteristics of each cluster. This is enhanced by showing decennial global MDG development by a population weighted aggregate of all countries. Below follows a description of the countries in each cluster.

Most of the countries in cluster A are from Sub-Saharan Africa. Exceptions are Afghanistan and Haiti, for example.

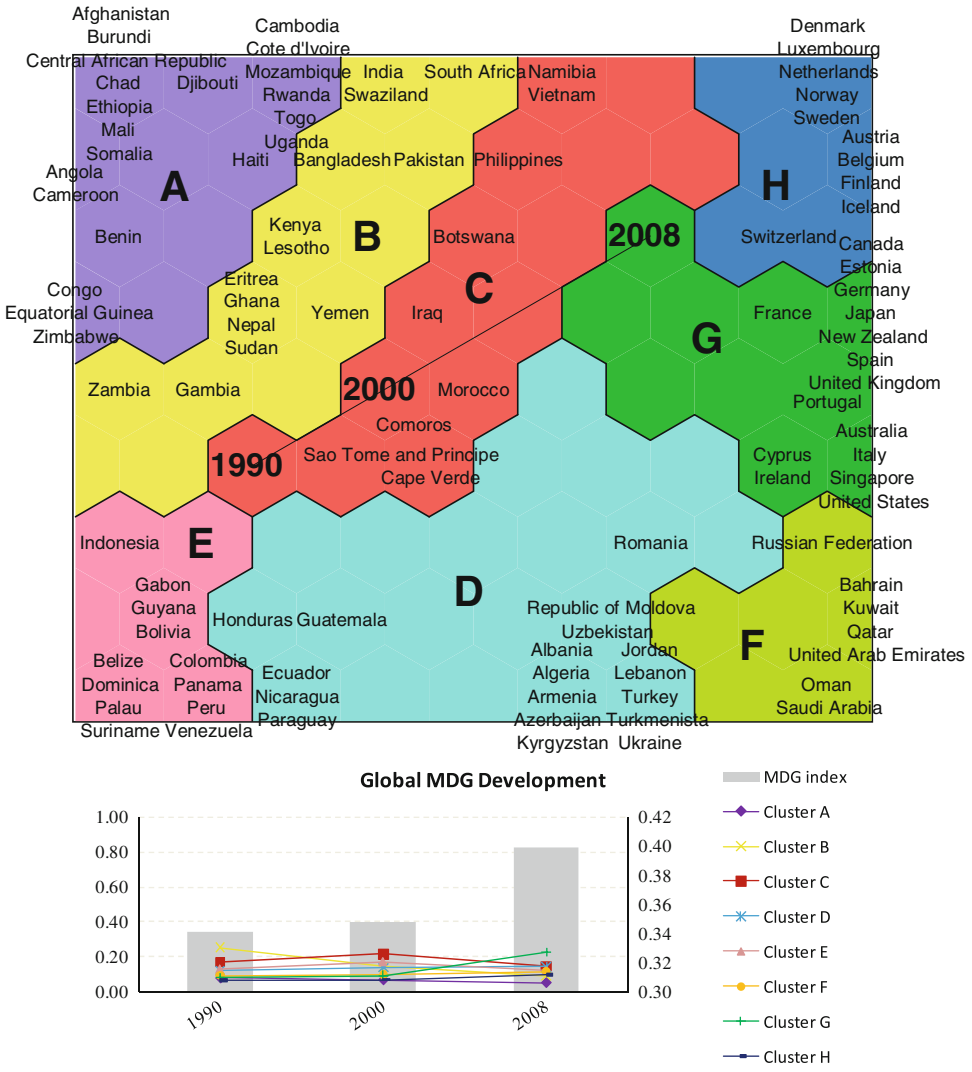

Fig. 3 A SOM grid with data for a descriptive subset of countries in 2005. The *line graphs' scale* corresponds to the left axis and the color coding follows that of the map, while the *bar charts* represent the MDG index value of the node that each data point represents and corresponds to the right axis (color figure online)

Although most of the countries in cluster B are likewise from Sub-Saharan Africa, more countries from other parts of the world, such as India, Pakistan, Bangladesh, Nepal and Yemen, are projected into the cluster. Cluster C still includes countries from Sub-Saharan Africa, but has a higher share of countries from the rest of the world, such as Vietnam, the Philippines, Iraq, Morocco and Cape Verde. Cluster D consists of two groups of countries: a group from South and Central America and a group mainly from Eastern Europe and Western Asia. The former group resembles the countries in cluster E, while the latter group resembles the countries in cluster F. The main difference of cluster E to the neighboring clusters is the high proportion of land area covered by forest, which is thus the principal characteristic of the countries in this cluster. Most of the countries in cluster E are South and Central American, but also a few African and Asian countries are projected into the cluster. Most of the countries in cluster F are Western Asian. Compared to the neighboring clusters, high CO_2 emissions is the main distinguishing factor. The countries projected into cluster G are mainly European, but include also the most developed countries from other continents, such as Japan, Singapore, New Zealand, Australia, Canada and the US. Similarly as cluster D, this cluster is divided into two groups of countries: a group that resembles cluster F, i.e., has high CO_2 emissions, and a group that resembles cluster H. The countries in cluster H are mainly European, such as the Nordic countries, with high measures for the indicators of gender equality and global partnership for development. The figure shows that MDG development of the world has been progressing towards a better state. Since the aggregate is an average of all data, it is commonly projected into the middle of the map. Thus, the variation over time is small, but the changes in the indicators, e.g. MDG index, are still valid. The line graph in Fig. 3 depicts the averaging nature of the data by showing simultaneous memberships to several clusters, but still shows that the overall MDG index has increased over time.

The visual representation of the SOM can be enhanced by combining it with geospatial data [12]. For each country, the color code of the corresponding cluster is projected onto a geographic map. This enhancement enables simultaneous assessment of both social and geographic space. The geographic map can be used for assessing the overall state of the MDG conditions and whether development is regionally dependent. In Fig. 4, the cluster color codes for all countries in 2005 are projected onto a geographic map. The map shows a slight regional dependence of the MDG indicators: cluster A represents mainly Sub-Saharan African countries, cluster E equatorial countries and cluster H Northern European countries. The rest of the clusters represent geospatially heterogeneous groups of countries.

In Fig. 5, the Sub-Saharan African countries are projected onto the map on two points in time. This enables in-depth dynamic analysis of less developed countries' progress. We further enhance the representation by computing a population weighted aggregate of the Sub-Saharan African countries. The left map shows a projection of the Sub-Saharan African countries in 1990 and the right map in 2008. The labels '1990', '2000' and '2008' represent the population weighted aggregate of Sub-Saharan African countries. In 1990, the countries are concentrated in the upper-left corner of the map. Similarly, the countries are still in 2000 projected into the upper-left corner (this projection is omitted from the paper). The Sub-Saharan African countries do, however, show progress during the first half of the MDG period (2000–2008). The right map showing data for 2008 illustrates that a major part of the countries have progressed towards the right part of the MDG map. This transition from cluster A into cluster B

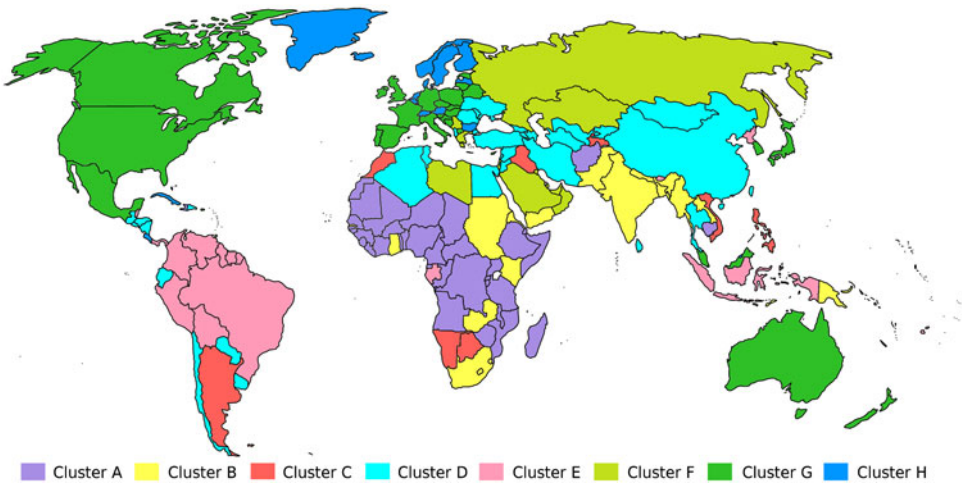

Fig. 4 A geographic map with the cluster color code (see Fig. 1) for all countries in 2005 (color figure online)

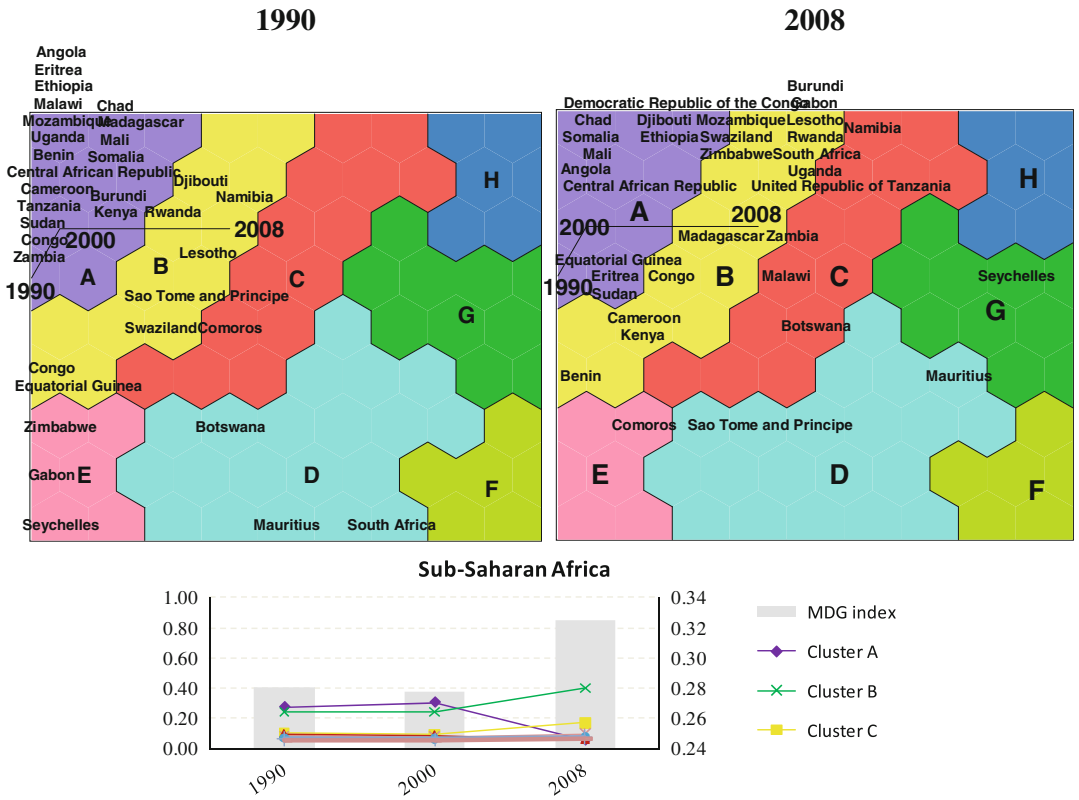

Fig. 5 Maps with data for Sub-Saharan Africa in 1990 (*left*) and 2008 (*right*). The *line graphs' scale* corresponds to the left axis and the color coding follows that of the map, but legends are only shown for clusters with memberships above 0.10. The *bar charts* represent the MDG index value of the node that each data point represents (color figure online)

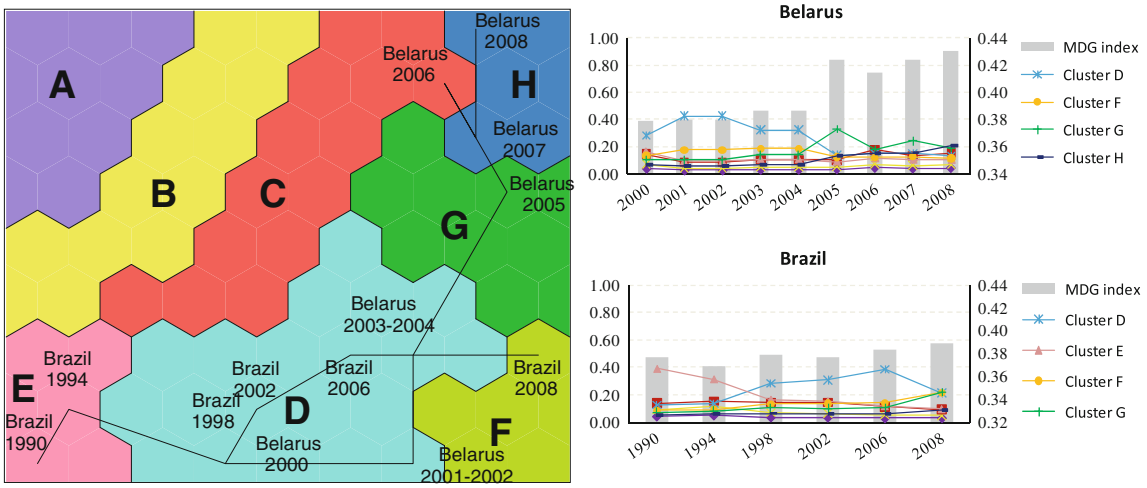

Fig. 6 Maps with quadrennial data for Brazil (*left*) and annual data for Belarus (*right*). See notes for Fig. 5 (color figure online)

and C indicates progress in a desired direction. This is confirmed by the temporal evolution of the aggregate. The line graph shows that Sub-Saharan African countries membership to cluster A actually increased during the period 1990–2000, but from 2000–2008 they have moved to cluster B by showing a strong membership to that cluster and the second strongest to cluster C, not cluster A. The map does not, however, quantify whether the countries

have met the concrete, numerical targets of each MDG. Since the MDG map is used for assessing the progress both over time and across countries, the targets, which are individual for each country, cannot be assessed. If the progress of solely one country is monitored, such as in Fig. 6, then a 'target line', indicating the needed progress for achieving the goal, can be added to each feature plane. However, due to lack of space, the 15 feature planes are not reproduced for the country-specific time-series analyses below.

4.2 Visualizing the evolution of a country

Time-series data can also be projected onto the MDG map. By showing two examples, Fig. 6 illustrates the projection of data series onto the map. The map shows annual evolution from 2000 to 2008 in Belarus and quadrennial evolution from 1990 to 2008 in Brazil. The data for Belarus shows rapid development by moving in only 8 years from the average cluster D into the best in class cluster H. The temporal variations between clusters and increase of the MDG index are shown by the line graph. The development in Brazil, on the other hand, has involved a movement from the average, or below average, cluster E into cluster F, an advanced economy cluster. Interestingly, the line graph depicts that although moving towards a better state in terms of several MDG subdimensions, the high-polluting state does actually not reflect significant improvements in terms of the MDG index.

5 Conclusions

This paper lays out a methodology for visual monitoring of MDGs. With only a few years to go before the deadline of the MDGs, tracking the progress towards the goals is of central importance. In this paper, the SOM is used for tracking the progress towards the multidimensional MDGs. The visual analyses include benchmarking countries based on the MDGs, monitoring the evolution of countries' MDG indicators over time and assessing the geospatial dimension of the multidimensional MDGs. This can be done on any aggregated level varying from country-specific analysis to world aggregates. The fuzzification of the clustering enables a visual representation of the temporal evolution, where the cluster centers express the representative MDG states for the countries and the fluctuations of the membership degrees represent their variations over time. The created MDG index associates a mashup state for each data point, which is necessary for assessing directions on the map. The experiments in this paper indicate that the SOM is an appropriate tool for rough, visual analyses of complex data.

The visual representation in this paper suggests that the United Nations should consider publishing a SOM model, an 'MDG map', for visual monitoring of the progress. In future research, the SOM could be employed for clustering and visualizing the other category of indicators, the intermediate. In contrast to this paper, the model would give indications on what is happening with well-being and its determinants, enabling up-to-date corrective actions. A simultaneous analysis of the intermediate and final indicators would enable examining the relations and impacts between both categories. The focus of future work should also be on enhancing the performance of the learning system by, for instance, maximizing the fuzzy entropy or combining multiple reducts [38–40] and on improving the second-level clustering by testing some newer techniques (e.g. [8, 9, 15]).

Acknowledgments I acknowledge Barbro Back and Tomas Eklund for helpful comments and suggestions.

References

1. Alkire S, Santos ME (2010) Acute multidimensional poverty: a new index for developing countries. Oxford Poverty and Human Development Initiative, Working Paper 38, University of Oxford
2. Boehme O, Hardoon D, Manevitz L (2011) Classifying cognitive states of brain activity via one-class neural networks with feature selection by genetic algorithms. Int J Mach Learn Cybern 2(3):125–134
3. Collan M, Eklund T, Back B (2007) Using the self-organizing map to visualize and explore socio-economic development. EBS Rev 22(1):6–15
4. Coudouel A, Hentschel JS, Wodon QD (2002) Poverty measurement and analysis. In: Klugman J (ed) A sourcebook for poverty reduction strategies. The International Bank for Reconstruction and Development/The World Bank, Washington, pp 29–69
5. Cox T, Cox M (2001) Multidimensional scaling. Chapman & Hall/CRC, Boca Raton
6. Deboeck G (1998) Best practices in data mining using self-organizing maps. In: Deboeck G, Kohonen T (eds) Visual explorations in finance with self-organizing maps. Springer, Berlin, pp 201–229
7. Eklund T, Back B, Vanharanta H, Visa A (2008) Evaluating a SOM-based financial benchmarking tool. J Emerg Technol Acc 5(1):109–127
8. Graaff AJ, Engelbrecht AP (2011) Clustering data in stationary environments with a local network neighborhood artificial immune system. Int J Mach Learn Cybern. doi:10.1007/s13042-011-0041-0
9. Guo G, Chen S, Chen L (2011) Soft subspace clustering with an improved feature weight self-adjustment mechanism. Int J Mach Learn Cybern. doi:10.1007/s13042-011-0038-8
10. Kaski S (1997) Data exploration using self-organizing maps. Acta Polytechnica Scandinavica, Mathematics, Computing and Management in Engineering Series No. 82., Espoo
11. Kaski S, Kohonen T (1996) Exploratory data analysis by the self-organizing map: structures of welfare and poverty in the world. In: Proceedings of the 3rd International Conference on Neural

Networks in the Capital Markets. World Scientific, London, pp 498–507
12. Kaski S, Venna J, Kohonen T (2000) Coloring that reveals cluster structures in multivariate data. Aust J Intell Inf Process Syst 6:82–88
13. Kohonen T (1982) Self-organized formation of topologically correct feature maps. Biol Cybern 66:59–69
14. Kohonen T (2001) Self-organizing maps, 3rd edn. Springer, Berlin
15. Liang J, Song W (2011) Clustering based on steiner points. Int J Mach Learn Cybern. doi:10.1007/s13042-011-0047-7
16. Naq AK, Mitra A (2002) Identifying patterns of socio-economic development using self-organizing maps. J Soc Econ Dev 4(1):55–88
17. Noorbakhsh FA (1998) A modified human development index. World Dev 26:517–528
18. Prados de la Escosura L (2010) Improving human development: a long-run view. CEPR discussion Paper 7982
19. Prennushi G, Rubio G, Subbarao K (2002) Monitoring and evaluation. In: Klugman J (ed) A sourcebook for poverty reduction strategies. The International Bank for Reconstruction and Development/The World Bank, Washington, pp 105–130
20. Ravallion M (2010) Mashup indices of development. Policy Research Working Paper 5432, World Bank
21. Resta M (2009) Early warning systems: an approach via self organizing maps with applications to emergent markets. In: Proceedings of the 2009 conference on New Directions in Neural Networks: 18th Italian Workshop on Neural Networks. IOS Press, The Netherlands
22. Ripley BD (1996) Pattern recognition and neural networks. Cambridge University Press, Cambridge
23. Sagar AD, Najam A (1998) The human development index: a critical review. Ecol Econ 25:249–264
24. Sahn DE, Stifel DC (2003) Progress toward the millennium development goals in Africa. World Dev 31(1):23–52
25. Samad T, Harp SA (1992) Self-organization with partial data. Netw Comput Neural Syst 3:205–212
26. Sammon JW (1969) A non-linear mapping for data structure analysis. IEEE Tran Comput 18(5):401–409
27. Sarlin P, Eklund T (2011a) Financial performance analysis of European banks using a fuzzified self-organizing map. In: Proceedings of the 15th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES2011), Springer, Kaiserslautern, September 12–14, 2011, pp 185–194
28. Sarlin P, Eklund T (2011b) Fuzzy clustering of the self-organizing map: some applications on financial time series. In: Proceedings of the 8th International Workshop on Self-Organizing Maps (wSOM'11), Springer, Helsinki, June 13–15, 2011, pp 40–50
29. Sarlin P, Peltonen T (2011) Mapping the state of financial stability. ECB Working Papers No. 1382
30. Tong DL, Mintram R (2010) Genetic algorithm-neural network (GANN): a study of neural network activation functions and depth of genetic algorithm search applied to feature selection. Int J Mach Learn Cybern 1:75–87
31. Tukey JW (1977) Exploratory data analysis. Addison-Wesley, Reading
32. Tyler Z, Gopal S (2010) Sub-Saharan Africa at a crossroads—a quantitative analysis of regional development. The Pardee Papers, No. 10, May 2010
33. UNDG (2003) Indicators for Monitoring the millennium development goals: definitions, rationale, concepts and methods. United Nations Development Group, New York. Available at: http://unstats.un.org/unsd/mdg/Resources/Attach/Indicators/HandbookEnglish.pdf. Accessed 5 December 2010
34. UNDP (1993) Human development report. Oxford University Press, New York, also published in various other years
35. UNECOSOC (2010) Assessing progress in Africa towards the millennium development goals report. E/ECA/COE/29/15 and AU/CAMEF/EXP/15(V). Available online: http://www.un.org/regionalcommissions/MDGs/eca_assessingprogress10.pdf. Accessed 10 December 2010
36. Vesanto J, Alhoniemi E (2000) Clustering of the self-organizing map. IEEE Trans Neural Netw 11(3):586–600
37. Vesanto J, Sulkava M, Hollmén J (2003) On the decomposition of the self-organizing map distortion measure. In Proceedings of the Workshop on Self-Organizing Maps (wSOM'03), Springer, Hibikino, September 11–14, 2003, pp 11–16
38. Wang XZ, Dong CR (2009) Improving generalization of fuzzy if-then rules by maximizing fuzzy entropy. IEEE Trans Fuzzy Syst 17(3):556–567
39. Wang XZ, Zhai JH, Lu SX (2008) Induction of multiple fuzzy decision trees based on rough set technique. Inf Sci 178(16):3188–3202
40. Wang XZ, Dong CR, Fan TG (2007) Training T-S norm neural networks to refine weights for fuzzy if-then rules. Neurocomputing 70(13–15):2581–2587
41. Ward J (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58:236–244