Are they Really Neighbor? A Statistical Analysis of the SOM Algorithm Output

Eric de Bodt[1], Marie Cottrell[2], Michel Verleysen[3]

[1] Université Catholique de Louvain, IAG-FIN, 1 pl. des Doyens,
B-1348 Louvain-la-Neuve, Belgium
and
Université Lille 2, ESA, Place Deliot, BP 381,
F-59020 Lille, France

[2] Université Paris I, SAMOS-MATISSE, UMR CNRS 8595
90 rue de Tolbiac,
F-75634 Paris Cedex 13, France

[3] Université Catholique de Louvain, DICE, 3, place du Levant,
B-1348 Louvain-la-Neuve, Belgium

Abstract. One of the attractive features of Self-Organizing Maps (SOM) is the so-called "topological preservation property": observations that are close to each other in the input space (at least locally) remain close to each other in the SOM. In this work, we propose the use of a bootstrap scheme to construct a statistical significance test of the observed proximity among individuals in the SOM.

1. Introduction

The SOM algorithm was introduced by Kohonen in 1981 and has been the focus of a sizeable amount of attention in the scientific community since then. Numerous applications have been proposed (see Kohonen [1995] for a representative list of them) and the theoretical properties have been carefully studied (see Cottrell, Fort & Pagès [1998] for a review of the established results up to now). Henceforth, we will consider here that the SOM algorithm is familiar to the reader.

One of the most attractive features of SOM (in particular for applications in the field of data analysis) is the so-called "topological preservation property": after organization through the training algorithm, observations that are close to each other in the input space (at least locally) belong to units that are neighbors (or are actually within the same unit).

In real applications, the SOM algorithm is used on a finite data set, which can be seen as a sample from some unknown distribution. As the Kohonen algorithm is a stochastic process, the results cannot be exactly the same from one run to another one, and an important question that arises about the resulting map is: "Is it reliable?» For example, a question that has not received a lot of attention up to now is the statistical significance of the observed neighborhood in the SOM obtained after learning. Having observed that two individuals from the analyzed sample belong to neighbor units, what is the probability that they are actual neighbors in the population? In other words, what is the sampling distribution of the observed proximity and is it possible to propose a statistical test to assess their significance?

We propose the use of the bootstrap approach to evaluate the reliability of the map on both points of view of *quantification* (evaluated by the sum of squares intra-classes, cf. eq. 1) and of *neighborhood significance* (evaluated by the stability of the observed proximity on the map, cf eq. 2).

We will first recall the central ideas of the bootstrap, as introduced by Efron [1979] and address specific difficulties encountered when applying the bootstrap for the SOM maps (for references on bootstrap procedures and their applications see e.g. Efron, Tibshirani [1986, 1993], Freedman [1981, 1984], LePage, Billard [1992], Noreen [1989], …).

2. A Bootstrap Procedure adapted to the SOM Algorithm

The main idea of the bootstrap, introduced by Efron in 1979, is to use the so-called "plug-in principle". Let F be a probability distribution depending on an unknown parameter $\boldsymbol{q}$. Let $\boldsymbol{x}=x_1, x_2,..., x_n$ be the observed sample of data and $\hat{\boldsymbol{q}} = T(\boldsymbol{x})$ an estimate of $\boldsymbol{q}$. The bootstrap consists of using artificial samples (called *bootstrapped samples*) with the same empirical distribution as the initial data set in order to guess the distribution of $\hat{\boldsymbol{q}}$. Each *bootstrapped sample* consists in n uniform drawings with repetitions from the initial sample. If $\boldsymbol{x}^*$ is a bootstrapped sample, $T(\boldsymbol{x}^*)$ will be a bootstrap replicate of $\hat{\boldsymbol{q}}$.

Our approach will consider bootstrap samples of the data in order to build the sampling distribution of different features of the map. The two features that we analyze here are the quality of the quantification and the neighborhood relation.

2.1. Quantification

The quality of the *quantification* is evaluated by the sum of all the distances between the observations and their winning code vector (the weight vector of the closest unit, which is the representative vector of the class they belong to). This sum is called *distortion* in the quantification theory, and *intra-classes sum of squares* by the statisticians. It can be expressed by:

$$SSIntra = \sum_{i=1}^{U} \sum_{x_j \in C_i} d^2(x_j, G_i) \qquad \text{eq.1}$$

where U is the number of classes (or units), C_i is the i-th class, G_i is the code vector of class C_i, and d is the classical Euclidean distance in the data space.

Note that the decreasing function associated with the SOM algorithm for a constant size of neighborhood and finite data set is *the sum of squares intra-classes extended to the neighbor classes*, (see Ritter *et al* [1992]). But actually, in the last part of the iterations, no neighbor is considered. So at the end, the SOM algorithm is equivalent to Simple Competitive Learning and exactly minimizes the *SSIntra* value.

The bootstrapped samples will allow us to build the sampling distribution of *SSIntra*.

2.2. Neighborhood relations

The stability of the neighborhood relations is simply evaluated by the number of cases where, during the bootstrap process, two observations are neighbor or not neighbor. The stability of neighborhood therefore has to be evaluated for a couple of observations and, classically, we have to define the radius of neighborhood at which the proximity is taken into account (see eq. 2). For any pair of data x_i and x_j,

$$STAB_{i,j}(r) = \frac{\sum_{b=1}^{B} NEIGH_{i,j}^{b}(r)}{B} \qquad \text{eq.2}$$

where $NEIGH^{b}_{i,j}(r)$ is an indicator function that returns 1 if the observations x_i and x_j are neighbor at the radius r for the bootstrap sample b, and B is the total number of bootstrapped samples. A perfect stability would lead $STAB_{i,j}$ to always be 0 (never neighbor) or 1 (always neighbor). We can study the significance of the statistics $STAB_{ij}(r)$, by comparing it to the value it would have if the observations fell in the same class (or in two classes distant of less than r) in a completely random way.

Let U be the total number of classes and v the size of the considered neighborhood. The size v of the neighborhood can be computed from the radius r by $v = (2r + 1)$ for a one-dimensional SOM map (a string); and $v = (2r + 1)^2$ for a two-dimensional SOM map (a grid). For a fixed pair of observations x_i and x_j, with random drawings, the probability of neighboring would be v/U. If we define a Bernoulli random variable with probability of success v/U, (where success means: "x_i and x_j are neighbor"), the number Y of successes on B trials is distributed as a Binomial distribution, with parameters B and v/U. So, it is possible to build a test of the hypothesis H_0 "x_i and x_j are only randomly neighbor" against the hypothesis H_1 "the fact that whether x_i and x_j are neighbor or not is meaningful".

If B is large enough (i.e. greater than 50), the binomial random variable can be approximated by a Gaussian variable and, for example, with a test level of 5%, we conclude to H_1 if Y is less than $B\frac{v}{U} - 1.96\sqrt{B\frac{v}{U}\left(1-\frac{v}{U}\right)}$, or greater than $B\frac{v}{U} + 1.96\sqrt{B\frac{v}{U}\left(1-\frac{v}{U}\right)}$. Note that B depends on the pair (x_i, x_j), since the samples have to contain it. *This gives a level of significance to the presence/absence of the neighborhood relations.*

2.3 Specific problems

The application of the bootstrap procedure to the SOM algorithm raises two specific problems:
-the minimized function has a sizeable amount of local minima. Part of the variability of the estimated statistics (*SSIntra*, $STAB_{i,j}$) can be due to this convergence problem. As in Zapranis and Refenes [1999], we will analyze the impact of the "convergence difficulty" on the stability of the estimations.

-to evaluate $NEIGH^{b}_{i,j}(r)$, it is necessary that x_i and x_j are present in the bootstrap sample b, which is in no way guaranteed. To solve this problem, we use the same approach as in Efron and Tibshirani [1993]: the $STAB_{i,j}(r)$ is evaluated only on the bootstrap samples that contain observations x_i and x_j.

The proposed bootstrap procedure is summarized in Figure 1. The terminology we will use to present our results is the following:

- if no re-sampling is done (in order to analyze the variability of the results due only to convergence problems), we will talk of Monte-Carlo (**MC**) simulation,
- if re-sampling is done, we will talk of Bootstrap (**B**) simulation,
- if, for each bootstrap iteration, the SOM Map is initialized at random (in the input data space), we will talk of Common Monte Carlo (**CMC**) or Common Bootstrap (**CB**) (depending on the activation of re-sampling or not),
- if, for each bootstrap iteration, the SOM Map is initialized with the weight vectors obtained after the convergence of the initial learning, we will talk of Local Monte Carlo (**LMC**) or Local Bootstrap (**LB**),
- if we do the same computations as in the previous point, but we add a small random perturbation to the weight vectors, we will talk of Local Perturbed Monte Carlo (**LPMC**) or Local Perturbed Bootstrap (**LPB**).

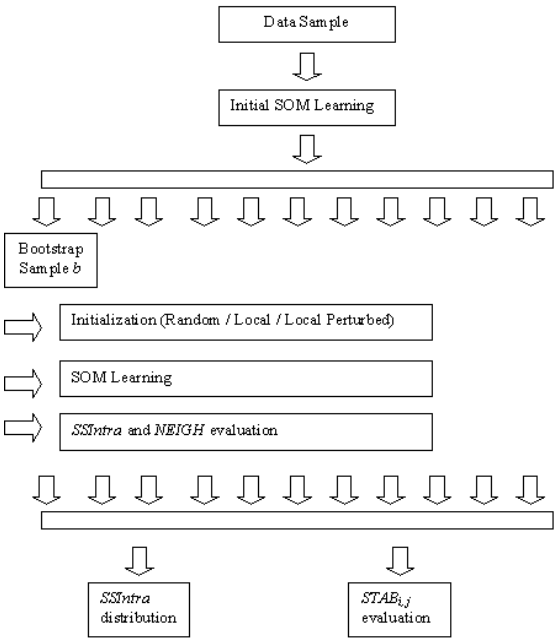

Figure 1: Bootstrap procedure for the SOM algorithm

3. Examples

3.1. Simulated data set and SOM Map

The results that we present and analyze here have been obtained on three simulated data sets, each one representing a specific situation. We will call them Gauss_1, Gauss_2 and Gauss_3. In each case, they are two-dimensional data sets, obtained by random drawing in an uncorrelated Gaussian distribution. They are respectively represented in figures 2, 3, and 4. In the first data there is only one cluster of observations. The second one contains three clusters of equal variance and some overlap. The third one is also composed of three clusters, but of

different variance and no overlap. Each data set has 500 observations. For data sets Gauss_2 and Gauss_ 3, observations 1-166, 167-333 and 334-500 are in the same cluster.

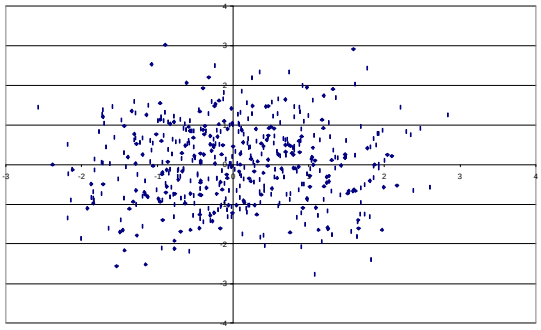

Figure 2: Gauss_1 data set

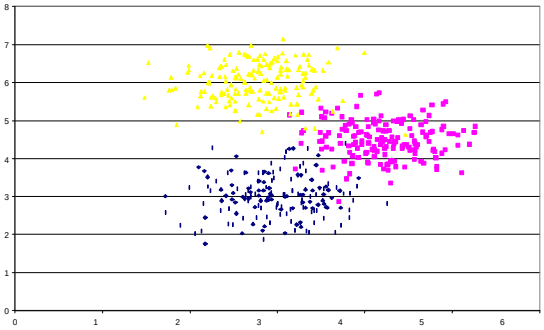

Figure 3: Gauss_2 data set

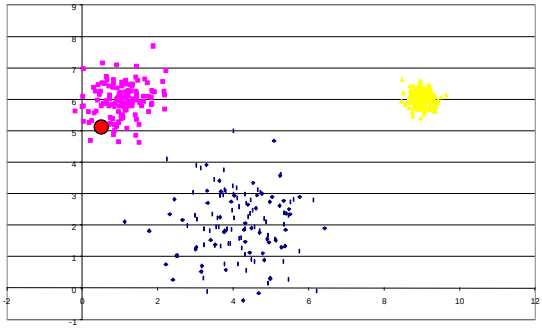

Figure 4: Gauss_3 data set

3.2 Variability of *SSIntra* due to convergence of the algorithm

The first point we present, with attendant results, is the variability of *SSIntra* due to convergence of the SOM algorithm. The point here is to see if the existence of local minima can introduce variability in the estimation of *SSIntra*. For the sake of conciseness, the results presented here are limited to a one-dimensional SOM map (or string), composed of either 3 or 6 units. Classically, the neighborhood and the learning rate are decreasing during the learning.

Table 1 summarizes the coefficients of variation (CV)[1], for the distribution of *SSIntra* obtained by CMC (no re-sampling and random initialization at each iteration), Table 2, the CV obtained by LMC (no re-sampling, fixed initialization at each iteration), and Table 3, the CV obtained by LPMC (no re-sampling, small random perturbation of the fixed initialization). Each result presented here has been established with 5000 independent samples[2].

The comparison shows quite clearly that the mode of initialization has no influence on the stability of the *SSIntra* estimation. By switching from CMC to LMC (or PLMC), i.e. by fixing the initialization of the weight vectors, the obtained coefficients of variation are almost the same. This result is very different from those obtained by Zapranis and Refenes [1999] when applying bootstrap to MLP and emphasize the great robustness of the SOM algorithm. The most interesting result that appears in tables 1-3 is the important impact of the number of units on the CV in Gaussian_3 cases. As can be seen in figures 2, 3 and 4, Gauss_3 is the only case with well-separated asymmetric clusters. It is clear that the "natural" number of units should be 3 and, in some sense, a SOM map with 6 units is over parameterized. The unstability of *SSIntra* seems, at first sight, to indicate the wrong choice of number of units. It is this point in particular that we will explore in the next section of this paper.

	3 units	6 units
Gauss_1	0.052	0.045
Gauss_2	0.051	0.046
Gauss_3	0.076	**0.101**

Table 1: Coefficients of variation of *SSIntra* for Common Monte-Carlo (CMC)

	3 units	6 units
Gauss_1	0.053	0.044
Gauss_2	0.049	0.045
Gauss_3	0.064	**0.103**

Table 2: Coefficients of variation of *SSIntra* for Local Monte-Carlo (LMC)

	3 units	6 units
Gauss_1	0.052	0.045
Gauss_2	0.051	0.046
Gauss_3	0.067	**0.101**

Table 3: Coefficients of variation of *SSIntra* for Local Perturbed Monte-Carlo (LPMC)

3.3 Assessing the right number of units in a SOM Map

Table 4 shows the CV of *SSIntra* obtained from the three simulated data sets presented in section 3.1. The results have been obtained using 100 bootstrap samples. They confirm those highlighted in the previous section. For Gauss_1, where there is only one natural cluster, the CV of *SSIntra* exhibits oscillations around 0.45. For Gauss_3, as expected, the addition of a fourth unit generates a large increase in the CV. The result seems to be surprising for the

[1] The coefficient of variation CV is equal to 100 σ/μ, where σ is the standard deviation and μ is the mean value.

[2] Such a large number of samples, in practice, is not necessary (100 being enough); but, we wish to be certain of the numerical stability of the results.

Gauss_2 data set, where there is no increase of the CV of *SSIntra* when adding a fourth unit. The explanation lies in the strictly symmetrical form of the three clusters and in their overlapping positions (the instability of the location of the fourth unit does not change the level of *SSIntra* obtained from one bootstrap sample to another bootstrap sample).

Nb of units	Gauss_1	Gauss_2	Gauss_3
1	0.052	0.043	0.055
2	0.045	0.060	0.089
3	0.059	0.054	0.065
4	0.055	0.049	**0.144**
5	0.044	0.066	0.152
6	0.051	0.047	0.120
9	0.054	0.047	0.109
12	0.037	0.049	0.092
15	0.040	0.040	0.080

Table 4 : Coefficients of variation of *SSIntra* obtained after Local Bootstrap

We observe the same on a real data set called POP[3]. Each country is described by 6 ratios: annual population growth, mortality rate, analphabetism rate, population proportion in high school, GDP per head and GDP growth rate. As shown in figure 5, for the POP data set, the increase of the CV of *SSIntra* is situated near the addition of the seventh or eighth unit.

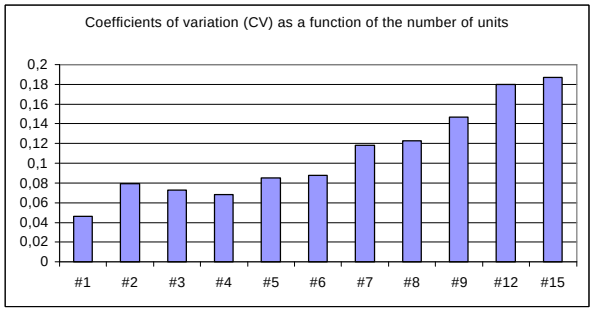

Figure 5: Evolution of the CVs of *SSIntra* for the POP data set when increasing the number of units of the one dimensional SOM Map

3.4 A statistical test of the proximity relations in the SOM Map

In this section, we present results concerning the stability of the neighborhood relations that appears in the SOM maps.

[3] This actual data (extracted from official public statistics for 1984) was used in Blayo, F. & Demartines, P. (1991): *Data Analysis: How to Compare Kohonen Neural Networks to Other Techniques?* in *Proceedings of IWANN'91*, Ed. A.Prieto, Lecture Notes in Computer Science, Springer-Verlag, 469-476. They are available at http://panoramix.univ-paris1.fr/SAMOS/

Pair of obs.	Gauss_2 3 units	Pair of obs.	Gauss_3 3 units
137/43 C1/C1	1	137/43 C1/C1	0
137/255 C1/C2	0	137/255 C1/C2	1
137/437 C1/C3	0	137/437 C1/C3	0
137/70 C1/C1	1	137/70 C1/C1	0
137/278 C1/C2	0	137/278 C1/C2	0
43/255 C1/C2	0	43/255 C1/C2	0
43/437 C1/C3	0	43/437 C1/C3	0
43/70 C1/C1	1	43/70 C1/C1	1
43/378 C1/C1	0	43/378 C1/C1	0
255/437 C2/C3	0	255/437 C2/C3	0
255/70 C2/C1	0	255/70 C2/C1	0
255/378 C2/C3	0	255/378 C2/C3	0

Table 5: Frequencies of neighborhood obtained by Local Bootstrap

Table 5 shows the results concerning $STAB_{i,j}$. We took a 3-units one-dimensional Kohonen map. The two pairs of columns concern the neighborhood with radius r=0, (i.e. the observations are considered as neighbors only if they belong to the same class). In the columns "Pair of obs", the cluster ownership are mentioned (e.g. the first pair of observations in Gauss_2 data set is 137/43; C1/C1 means that observations 137 and 43 are both members of cluster 1). All estimations have been computed with 100 bootstrap samples. The main results are as follows:

- For the Gauss_2 data set, we obtain strictly what was expected: if two observations are in the same cluster, the probability they belong to the same unit is 1 (and vice-versa).
- For the Gauss_3 data set, the conclusions are the same as those obtained for the Gauss_2 data set, except for observation 137, which is wrongly associated with some observations of the second cluster. In figure 4, we mark this observation with a big point. As we can see, it is located in the second cluster (while issued from the first one). This corresponds to an error of classification due to its location and the results obtained by bootstrap are fully coherent.

3.5 Real data sets

We use two real data sets. One is the POP data set previously used in section 3.3. The other one is the POP_93 data set, which contains 95 countries described by the same ratios, measured in 1993 (instead of 1984 for the POP data set.

Table 6 shows the results for the POP data set, for a 6-units one-dimensional SOM map and a neighborhood radius r of 0 (column 2) and of 1 (column 3). The first column mentions the country names.

Pair of countries	POP(r=0) 6 units	POP (r=1) 6 units
Turkey/Upper Volta	0.04*	0.65*
Turkey/Cuba	0*	0.22*
Turkey/Sweden	0*	0.05*
Turkey/France	0*	0*
Turkey/Greece	0*	0.25*
Upper Volta/Cuba	0*	0*
Upper Volta / Sweden	0*	0*
Upper Volta / France	0*	0*
Sweden/France	1*	1*
Cuba / Sweden	0.02*	0.81*
Cuba / France	0.02*	0.78*
Cuba / Greece	0.69*	0.97*

*significant at 1%

Table 6: Frequencies of neighborhood obtained by Local Bootstrap

Table 7 shows the results for the POP_93 data set, for a 7×7-units two-dimensional SOM map, and a neighborhood radius of 2.

Pair of Countries	POP_93 (r=2) 49 units
Greece/France	0.18*
Australia-France	0.82*
Greece/Belgium	0.21*
Turkey/France	0.02*
Singapore/USA	O.49
Sweden/Japan	0.73*
Greece/Malta	1*
Canada/France	0.84*
Sweden/France	0.97*
USA/Zimbabwe	0*
USA/Finland	0.85*
USA/Australia	0.68*

*significant at 1%

The levels of significance have been calculated from a Binomial distribution with p=1/6, 3/6, 25/49, respectively according to the values of r and U (see section 2.2).

For the POP data set, as well as for the POP_93 data sets, the observed similarities between the countries agree with the economic situation (in 1984, or in 1996), as far as we know. It is necessary to study the map in a more detailed way to fully interpret the results, but it is out of the scope of this paper. However, it is evident that France is completely different from Upper-Volta (presently Burkina-Faso), and that France and Sweden are very similar with respect to the considered variables. The same remarks hold for all the observed pairs.

4. Conclusion

These are preliminary results, but are nonetheless very promising. We intend to pursue these tracks by:

-systematically studying how to determine the correct number of units using the coefficients of variation of the *SSIntra* for the bootstrapped samples, according to the number of units;

-analyzing the stability of the neighborhoods according to the number of units more deeply (as we saw, the stability disappears when the number of units is over-dimensioned);

-applying these methods to numerous real data and applying, in this context, well-known numerical optimizations to the Monte-Carlo procedure.

We think that this kind of work can supply the innumerable users of the SOM maps with a new tool that can make them increasingly confident in the power and effectiveness of the Kohonen algorithm.

References

[1] Cottrell M., Fort J.C. & Pagès, *Theoretical Aspects of the SOM Algorithm*, Neurocomputing, 21, 1998, p. 119-138.

[2] Efron B., *Bootstrap Methods : Another Look at the Jackknife, The 1977 Rietz Lecture*, The Annals of Statistics, vol. 7, n°1, 1979, p. 1-26

[3] Efron B. & Tibshirani R., *Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy*, Statistical Science, vol. 1, n°1, 1986, p. 54-77

[4] Efron B. & Tibshirani R., *An Introduction to the Bootstrap*, Chapman and Hall, 1993

[5] Freedman D.A., *Bootstrapping Regression Models*, Annals of Statistics, vol. 9, 1981, p. 1218-1228

[6] Freedman D.A., *On Bootstrapping Two-Stage Least-Squares Estimates in Stationary Linear Models*, vol. 12, n°3, 1984, p. 827-842

[7] Kohonen T., *Self-Organizing Maps*, Springer, Berlin, 1995.

[8] LePage R. & Billard L., *Exploring the Limits of Bootstrap*, Wiley, 1992

[9] Noreen, E.W., *Computer Intensive Methods for Testing Hypotheses - An Introduction*, Wiley, 1989

[10] Ritter H., Martinetz T. and Schulten K., *Neural Computation and Self-Organizing Maps*, Addison-Wesley, Reading, 1992.

[11] Zapranis A. & Refenes A.P., *Principles of Neural Model Identification, Selection and Adequacy*, Springer, 1999.

Appendix 1

The POP Data Set (1984)

Country	ANCRX	TXMORT	TXANAL	SCOL2	PIBH	CRXPIB	ID
South Africa	2,9	89	50	19	2680	-2,9	1
Algeria	2,9	114	58,5	47,9	2266	0,1	2
Saudi Arabia	4,2	111	75,4	39,7	10827	-10,8	3
Argentina	1,2	44	5,3	69,5	2264	2	4
Australia	1,3	10,4	0	86	9938	-1,2	5
Bahrain	3,8	57	20,9	76,3	8960	-10,1	6
Brazil	2,2	75	23,9	62,3	1853	-3,9	7
Cameroon	2,4	106	55,1	44,5	939	6,5	8
Canada	1	10	0,9	93	9857	3	9
Chile	1,7	42	7,7	85,2	1853	-0,5	10
China	1,4	71	31	44	231	10	11
South Korea	1,6	33	8,3	82,1	1716	9,3	12
Cuba	0,7	16,8	8,9	78,7	2046	5,2	13
Egypt	2,7	74	58,1	45,8	626	6	14
Spain	0,9	9,6	6,8	88	5316	2,3	15
United States	1	11,2	0,8	91	11732	3,3	16
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
West Germany	-0,2	11,4	0,5	89	5103	4,2	42
East Germany	-0,1	12	0,7	87	12176	1	43
United Kingdom	-0,1	10,1	0,8	83	8655	3,5	44
Senegal	2,6	152	77,5	19,2	430	2,3	45
Sweden	0,1	7	0,6	85	13920	1,8	46
Switzerland	0,6	8	0,9	88	15522	-0,1	47
Syria	3,8	60	46,3	50,7	1717	5,8	48
Turkey	2,1	119	31,2	42	1491	3	49
USSR	0,9	28,8	0,8	96	4562	4	50
Venezuela	3	40	19	57,7	3823	-2	51
Vietnam	2,3	97	13	59,5	220	5,2	52
Yugoslavia	0,9	31	13,2	83	2067	-1,3	53

Where: ANCRX is the annual population growth, TXMORT is the mortality rate, TXANAL is the analphabetism rate, SCOL2 is the population proportion in high school, PIBH is the GDP per head and CRXPIB is the GDP growth rate.

From: Blayo, F. & Demartines, P. (1991): *Data Analysis: How to Compare Kohonen Neural Networks to Other Techniques?* in *Proceedings of IWANN'91*, Ed. A.Prieto, Lecture Notes in Computer Science, Springer-Verlag, 469-476.