Bootstrapping Self-Organizing Maps to Assess the Statistical Significance of Local Proximity

Eric de Bodt[1], Marie Cottrell[2]

[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

Abstract. One of the attractive feature 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. While computer intensive at this stage, this work represents a first step in the exploration of the sampling distribution of proximities in the framework of the SOM algorithm.

1. Introduction

The SOM algorithm has been introduced by Kohonen in 1981 and has been the focus of lot of attention of 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). We will consider here that the SOM algorithm is familiar to the reader.

One of the most attractive feature 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 neighbor (or to the same unit). A question that does not have received up-to-now a lot of attention 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, which is the probability that they are really neighbor in the population? In other words, which is the sampling distribution of the observed proximity and is it possible to propose a statistical test to assess their significance?

To answer this question, we will first recall the central ideas of the bootstrap, as introduced by Efron [1979] and insist on the specific difficulties that we have to solve when applying the bootstrap in the field of neural-networks. We will clearly define the concept of proximity, propose a bootstrap procedure adapted to the SOM algorithm and introduce a Binomial test to assess the statistical significance of observed neighborhoods. Before concluding, we present the application of our propositions to three simulated data sets and to a real data base.

2. Bootstrap Procedures and their Applications to Neural-Networks

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

There is (at least) two ways to implement the plug-in principle .[1]:

- the first one, which is called *parametric sampling bootstrap*, supposes that the asymptotic distribution of the parameter estimate is known, which limits its field of application. Using the observed sample of data, the parameter of this distribution is estimated (for example, by maximizing a log-likelihood function). Then, the sampling distribution of any statistic derived from the model[2] is evaluated by a re-sampling scheme based on the estimated asymptotic distribution of the parameter estimate. This method allows to derive the sampling distribution of complex statistics, for which it would be impossible to use the analytical approach (see Efron, Tibshirani, [1993] for several examples).
- The second one, which is called *non-parametric sampling bootstrap*, is directly built on the empirical distribution of the observed sample of data. Using it as the (best) approximation of the population distribution, the following sampling scheme is used to evaluate the sampling distribution of any statistic calculated on the data set (it can be in particular the parameter estimate): draw at random a great number of samples from the original data set with replacement (each bootstrap sample is composed by the same number of observations as the original data set), for each bootstrap data set, evaluate the statistic of interest and use the obtained estimations to build its sampling distribution.

[1] Efron B., Tibshirani R, 1993, p. 35.
[2] The term model is used in its largest meaning.

Numerous propositions have been done in the literature to improve the original propositions of Efron in several directions : improving the Monte-Carlo sampling (see for example LePage and Billard [1992], Noreen [1989]), adapting the approach to the estimation of expectation, variance, confidence interval (see for example Efron and Tibshirani [1986]), adapting the approach to the regression framework (see for example Freedman [1981,1984]).

Zapranis and Refenes [1999] present an interesting analysis of the application of the bootstrap in the neural network fields. They apply it to the problem of model adequacy for multi-layer perceptrons (MLP). As they mention it, the main problem of applying the bootstrap approach to MLP is due to the fact that the minimized loss function is not quadratic. It can therefore exist numerous local minima in which the optimization algorithm (or so-called training algorithm) may remain blocked. To solve this problem, the authors introduce the concept of *local bootstrap* and *perturbed local bootstrap*. In *local bootstrap*, the MLP weight vectors obtained on the observed sample are used as an initialization for the learning on each bootstrapped data set. In *perturbed local bootstrap*, the same approach is used, but the MLP weight vectors are locally perturbed in order to avoid that the optimization algorithm remains blocked on this same solution. The authors show that these approaches allow avoiding the part of the variability due to convergence problem in the estimation of sampling distribution of the estimated parameters for the MLP.

3. A Bootstrap Scheme adapted to the SOM Algorithm

In real applications, the SOM algorithm is used on a finite data set, that can be seen as a sample from some unknown distribution. One of the important questions that raises about the resulting map is "Is it reliable?". We propose to use the bootstrap approach to evaluate the reliability of the map on both the point of view of *quantification* (evaluated by the sum of squares intra-classes, cf. eq. 1) and the *neighborhood significance* (evaluated by the stability of the observed proximity on the map).

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, the representative vector of the class they belong to). This sum is called *distortion* in the quantification theory, and *sum of squares intra-classes* by the statisticians. It can be expressed by :

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

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.

Let us recall 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*. But actually, in the last part of the iterations no neighbor is considered, and at the end, the SOM algorithm is equivalent to Simple Competitive Learning and minimizes exactly the *SSIntra* value.

The bootstrapped samples will help us to study the stability of the distortion by estimating it and its standard deviation, whatever the learning (which depends on the initialization, order of data presentation, decrease of the neighborhood size, of the adaptation parameter, etc.)

As to the stability of the neighborhood relation, it is simply evaluated by the number of cases where, during the bootstrap process, two observations are neighbor or not neighbor. The stability of neighborhood has therefore to be evaluated for couple of observations and, as classically, we have to define the radius of neighborhood at which the proximity is taken into account (see equation 2). For any par 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 be always 0 (never neighbor) or 1 (always neighbor).

The application of the bootstrap procedure to the SOM algorithm raises two specific problems :

- as for MLP, the minimized function has a lot 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] (cf. supra), we will therefore analyze the impact of the "convergence difficulty" on the stability of the estimations (see section 4 of the paper).
- to evaluate $NEIGH^{b}_{i,j}(r)$, it is needful to say that x_i and x_j must be part of the bootstrap sample b, which is by 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 part of the bootstrap samples that contains the observations x_i and x_j.

The proposed bootstrap procedure is resumed at figure 1. The terminology that 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 only due 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 of 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 if we add a small random perturbation to the weight vectors, we will talk of Local Perturbed Monte Carlo (**LPMC**) or Local Perturbed Bootstrap (**LPB**).

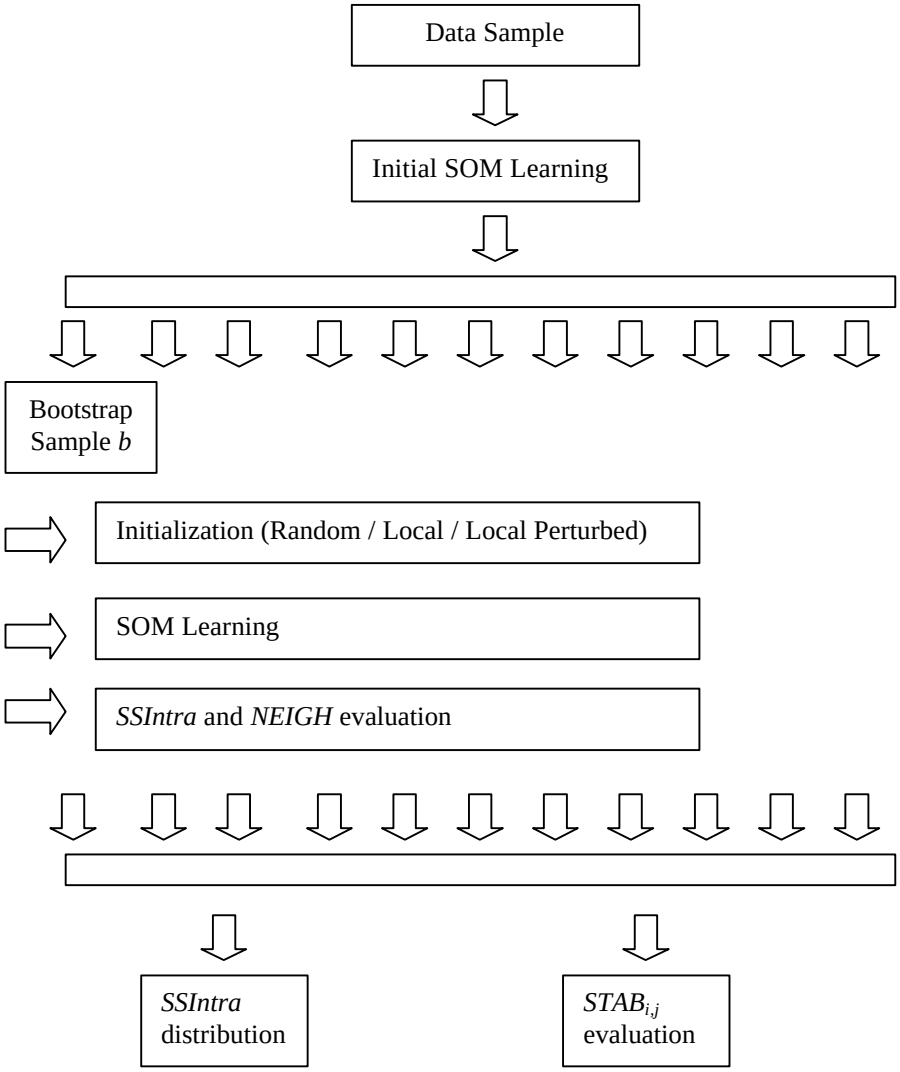

Figure 1 : Bootstrap procedure of the SOM algorithm

We can study the significance of the statistics $STAB_{ij}(r)$, by comparing it to the value it would have if the observations would fall in the same class (or in two classes distant of less than r) in an completely random way.

Let be U 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 par of observations x_i and x_j, if the repartition would be at random, the probability they are neighbor 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 x_i and x_j are neighbor or are 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, for 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 $B\frac{v}{U} + 1.96\sqrt{B\frac{v}{U}\left(1-\frac{v}{U}\right)}$.
This allows to give a level of significance to the presence/absence of the neighborhood relations.

4. Applications

4.1. Data set and SOM Map

The results that we present and analyze here have been obtained on three simulated data set[3], each of one representing a specific situation. We will call them Gaussian_1, Gaussian_2 and Gaussian_3. In each case, they are two-dimensional data sets, obtained by random drawing in uncorrelated Gaussian distribution. They are represented respectively at fig. 2, fig. 3 et fig.4. The first data set shows a situation where there is only one cluster of observations. The second contains three clusters with equal variance and some overlap. The third one is also composed of three clusters but with different variances and no overlap. Each data set is composed of 500 individuals and, for data sets Gaussian_2 and Gaussian 3, observations 1 to 166, 167 to 333 and 334 to 500 are in the same cluster.

[3] Complementary results have been obtained on several real data set but the simulated ones allow us to put into light more clearly the interesting results.

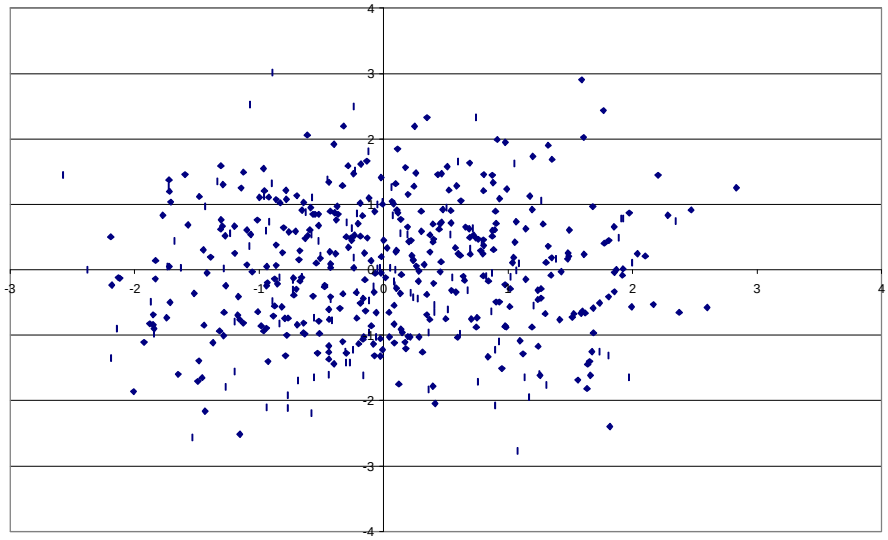

Figure 2 : Gaussian_1 data set

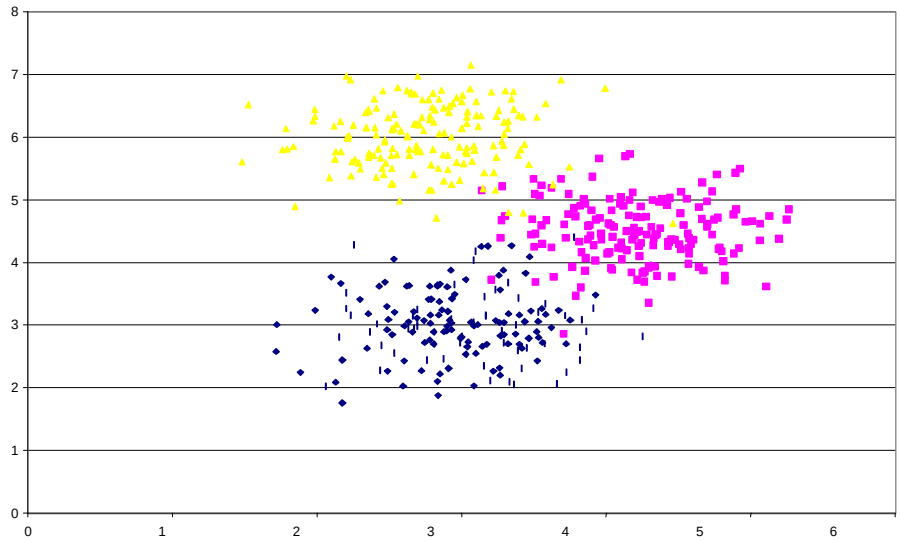

Figure 3 : Gaussian_2 data set

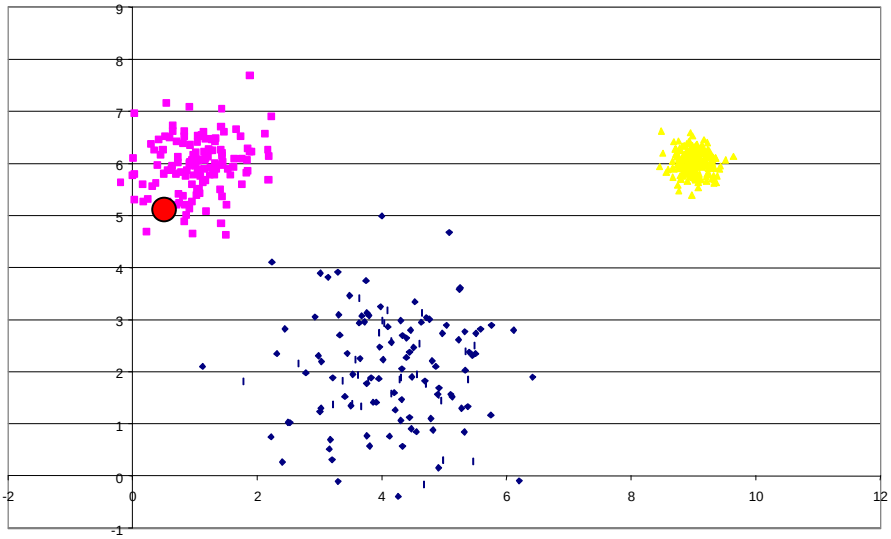

Figure 4 : Gaussian_3 data set

Always 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. As classically, the neighborhood and the learning rate are decreasing during the learning.

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

The first point about which we present some results is the variability of *SSIntra* due to convergence of the SOM algorithm. The point is here to see if the existence of local minima can introduce variability in the estimation of *SSIntra*. Table 1 summarizes the coefficients of variation (CV[4]) of the distribution of *SSIntra* obtained by CMC (no re-sampling and random initialization at each iteration), Table 2, those obtained by LMC (no re-sampling, fixed initialization at each iteration) and Table 3, those obtained by LPMC (no re-sampling, small random perturbation of the fixed initialization). Each result presented here has been established with 5000 independent samples[5]. The results for $STAB_{i,j}$ are not presented here for the sake of conciseness. The comparison shows quite clearly that the stability of the *SSIntra* estimation does not rely on the mode of initialization of the bootstrap procedure. By switching from CMC to LMC or PLMC, that is to say, by fixing the initialization of

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

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

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 to 3 is the important impact of the number of units on the CV in Gaussian_3 cases. As it can be seen in figures 2 and 3, Gaussian_3 is the only case with well separated asymmetric clusters. It is clear that the "natural" number of units should be 3 and that, in some sense, a SOM Map with 6 units is over parameterized. The stability of *SSIntra* seems at first sight to be an indicator of this wrong choice of number of units. This is the point that we will explore in the next section of the paper.

	3 units	6 units
Gaussian_1	0.052	0.045
Gaussian_2	0.051	0.046
Gaussian_3	0.076	0.101

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

	3 units	6 units
Gaussian_1	0.053	0.044
Gaussian_2	0.049	0.045
Gaussian_3	0.064	0.103

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

	3 units	6 units
Gaussian_1	0.052	0.045
Gaussian_2	0.051	0.046
Gaussian_3	0.067	0.101

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

4.3 Assessing the right number of units in a SOM Map

Tables 4 shows the CVs of *SSIntra* obtained on the three simulated data set presented at section 4.1 as well as on a real data set called POP, presented at the annex 1 of the paper[6]. The results have been obtained using 100 bootstrap samples. They confirm those highlighted in the previous section. For Gaussian_1, where there is only one

[6] These real data (extracted from official public statistics for 1984) were 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.

natural cluster, the CV of *SSIntra* exhibits oscillations around 0.45. For Gaussian_3, as expected, the addition of a fourth unit generates a high increase of the CV. As shown in table 4 and 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. The result can seem to be surprising for the Gaussian_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).

Number of units	Gaussian_1	Gaussian_2	Gaussian_3	POP
1	0.052	0.043	0.055	0.046
2	0.045	0.060	0.089	0,079
3	0.059	0.054	0.065	0.073
4	0.055	0.049	0.144	0.068
5	0.044	0.066	0.152	0.085
6	0.051	0.047	0.120	0.088
9	0.054	0.047	0.109	0.147
12	0.037	0.049	0.092	0.180
15	0.040	0.040	0.080	0.187

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

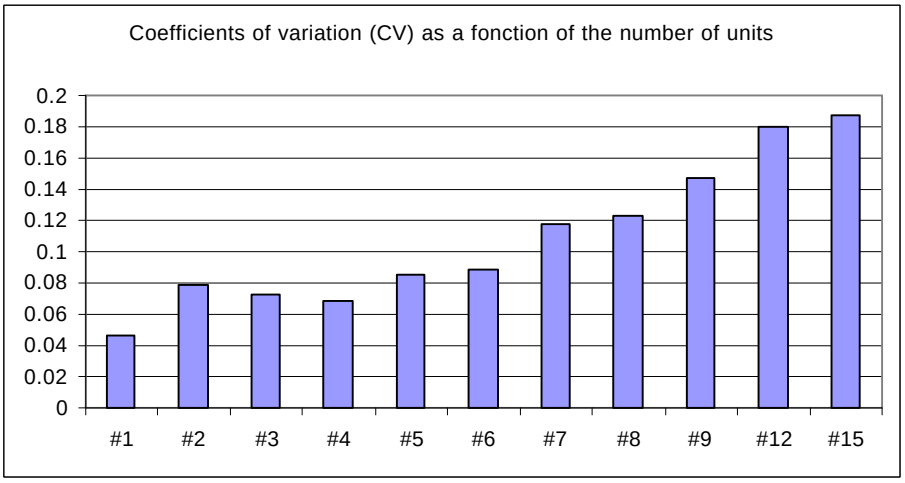

Figure 5 : Evolution of the CVs of *SSIntra* for the POP data set when increasing the number of units of the one dimensional SOM Map

4.4 A statistical test of the proximity relations among observations in the SOM Map

We present in this section some results about the stability of the neighborhood relations that appears in the SOM maps. The three first par of columns concern the neighborhood with radius r=0, (i.e. the observations are considered as neighbor only if they belong to the same class). The two last ones show the results for the POP data set, and a radius neighborhood of 1 (i.e. the observations are neighbor if they belong to the same class or to two adjacent classes).

Par of obs.	Gauss_2 3 units	Par of obs	Gauss_3 3 units	Par of countries	POP (r=0) 6 units	POP (r=1) 6 units
137/43 Cl1/Cl1	1	137/43 Cl1/Cl1	0	49/21 Turquie/Haute Volta	0.04**	0.65**
137/255 Cl1/Cl2	0	137/255 Cl1/Cl2	1	49/13 Turquie/Cuba	0***	0.22***
137/437 Cl1/Cl3	0	137/437 Cl1/Cl3	0	49/47 Turquie/Sweden	0***	0.05***
137/70 Cl1/Cl1	1	137/70 Cl1/Cl1	0	49/19 Turquie/France	0***	0***
137/278 Cl1/Cl2	0	137/278 Cl1/Cl2	0	49/20 Turquie/Greece	0***	0.25***
43/255 Cl1/Cl2	0	43/255 Cl1/Cl2	0	21/13 Haute Volta/Cuba	0***	0***
43/437 Cl1/Cl3	0	43/437 Cl1/Cl3	0	21/47 Haute Volta / Sweden	0***	0***
43/70 Cl1/Cl1	1	43/70 Cl1/Cl1	1	21/19 Haute Volta / France	0***	0***
43/378 Cl1/Cl1	0	43/378 Cl1/Cl1	0	47/19 Sweden/France	1***	1***
255/437 Cl2/Cl3	0	255/437 Cl2/Cl3	0	13/47 Cuba / Sweden	0.02**	0.81***
255/70 Cl2/Cl1	0	255/70 Cl2/Cl1	0	13/19 Cuba / France	0.02**	0.78***
255/378 Cl2/Cl3	0	255/378 Cl2/Cl3	0	13/20 Cuba / Greece	0.69***	0.97***

Table 5 : Frequencies of neighborhood obtained by Local Bootstrap

** significant at 5% - significant at 1%

Table 5 shows the results concerning $STAB_{i,j}$. In column "Par of obs", the number of two observations and, for the data sets Gaus_2 and Gaus_3, the cluster ownership are mentioned (for example, the first par of observations of Gaus_2 data set is 137/43; Cl1/Cl1 means that the observation 137 is member of cluster 1 and observation 43 is member of cluster 1). For the POP data set, we mention the country names. The number of units is mentioned in the title of the columns. All the estimations have been computed with 100 bootstrap samples. The levels of signification have been calculated from a Binomial distribution with p=1/6 (cf section 3). The main results are the following:

- for the Gaus_2 data set, we strictly obtain what was expected: if two observations are in the same cluster, the probability to belong to the same unit is 1 (and vice-versa). We have to remind here that the SOM algorithm is a stochastic one…
- for the Gaus_3 data set, the conclusions are the same as those obtained for the Gaus_2 data set, except for observation 137, which is wrongly associated with observations of the second cluster. On figure 4, we mark this observation with a red point. As we can see, it is located in the second cluster (while issued from the first one). It corresponds to an error of classification, due to its location and the results obtained by bootstrap are fully coherent.
- for the POP data set, the observed similarities between the countries agree with the economic situation in the year 1984, as long as we know. It would be necessary to study the map in a more detailed way to fully interpret the results, but it is out of scope of the paper. However, it is evident that France is completely different from Haute-Volta (nowadays Burkina-Faso), and that France and Sweden are very similar, with respect to the considered variables (see in the appendix).

5. Conclusion

These results are preliminary, but are very promising. We intend now to pursue these tracks

- by studying in a systematic way 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,
- by analyzing more deeply the stability of the neighborhoods according to the number of units, as we saw that the stability disappear when the number of units is over-dimensioned,
- by applying these methods to real numerous data and applying, in this context, well-know numerical optimization to the Monte-Carlo procedure.

We think that this kind of work can supply to the innumerable users of the SOM maps a new tool, which can make them more and more confident in the power and the 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] Zapranis A. & Refenes A.P., Principles of Neural Model Identification, Selection and Adequacy, Springer, 1999.

Annex 1

The POP Data Set (1984)

Country	ANCRX	TXMORT	TXANAL	SCOL2	PIBH	CRXPIB	ID
Afrique du sud	2,9	89	50	19	2680	-2,9	1
Algerie	2,9	114	58,5	47,9	2266	0,1	2
Arabie Saoudite	4,2	111	75,4	39,7	10827	-10,8	3
Argentine	1,2	44	5,3	69,5	2264	2	4
Australie	1,3	10,4	0	86	9938	-1,2	5
Bahrein	3,8	57	20,9	76,3	8960	-10,1	6
Bresil	2,2	75	23,9	62,3	1853	-3,9	7
Cameroun	2,4	106	55,1	44,5	939	6,5	8
Canada	1	10	0,9	93	9857	3	9
Chili	1,7	42	7,7	85,2	1853	-0,5	10
Chine	1,4	71	31	44	231	10	11
Coree du Sud	1,6	33	8,3	82,1	1716	9,3	12
Cuba	0,7	16,8	8,9	78,7	2046	5,2	13
Egypte	2,7	74	58,1	45,8	626	6	14
Espagne	0,9	9,6	6,8	88	5316	2,3	15
Etats Unis	1	11,2	0,8	91	11732	3,3	16
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
RDA	-0,2	11,4	0,5	89	5103	4,2	42
RFA	-0,1	12	0,7	87	12176	1	43
Royaume Uni	-0,1	10,1	0,8	83	8655	3,5	44
Senegal	2,6	152	77,5	19,2	430	2,3	45
Suede	0,1	7	0,6	85	13920	1,8	46
Suisse	0,6	8	0,9	88	15522	-0,1	47
Syrie	3,8	60	46,3	50,7	1717	5,8	48
Turquie	2,1	119	31,2	42	1491	3	49
URSS	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
Yougoslavie	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.