A NEW CRITERION TO EVALUATE THE STABILITY OF SOM

Jérôme Blancher[1,2] and Guy Lamarque[2]
[1]*Navtel Systems*
2 bis rue Muette. 28700 Houville-La-Branche, France.
Jerome.Blancher@univ-orleans.fr
[2]*Laboratoire d'Electronique, Signaux, Images (LESI)*
Université d'Orléans, 12 rue de Blois, BP 6744. 45067 Orléans Cedex 2, France.
Guy.Lamarque@univ-orleans.fr

Abstract – *A new stability criterion (γ_{50}) is proposed to evaluate the reliability of SOM. This criterion estimates the confidence that we can have on the topological neighborhood relationships shown on a map. This indicator is non graphical and gives more accurate information on the instability of SOM than the existing stabilities histograms methods. We show that SOM present particular statistical properties that are used with the γ_{50} criterion to determine the instability values of unorganized maps. This criterion is tested on two databases and provides more accurate results on the network's geometry than the mean quantization and the mean Kohonen errors.*

Key words – stability analysis, topology preservation, geometry of a map.

1 Introduction

The main objective of SOM [1] is to map high-dimensional data onto a two-dimensional (2-D) lattice of neural units by preserving the local neighborhood relationships of the dataset. Many criteria have been proposed in order to choose, between several different maps, the one that best preserves the topology of the input data, a list of such criteria can be found in [2] and [3]. Another efficient measure is the Kohonen cost function, as stated in [4]. However, none of these criteria can measure the reliability of the results after several learning procedures of the same dataset.

Averaging topographic errors obtained over all the maps generated can help to have a better measure of the topology preservation for each network configuration. Yet, this measure only indicates how local topology is preserved globally even if from one learning phase to another the effective local topology that is preserved does not correspond to the same part of the dataset. This problem was already raised in [5] where a stability measure was proposed. Histograms are drawn to compare the stability of the pairs of data for a given neighborhood distance with the stability of the pairs mapped in a random way. However this method is only graphical and several stability histograms are needed (one for each neighborhood distance).

In this paper we show through a statistical study of the stability of unorganized maps that they present particular properties that can be used to define a new stability criterion. This criterion can be applied to choose, from several network configurations, the one that gives the most reliable information on the dataset. Section 2 presents the basic concepts of the stability measure. Section 3 describes the statistical properties of unorganized maps. Section 4 introduces the new stability criterion. Section 5 gives experimental results on the stability of two datasets. Section 6 concludes the paper.

2 Stability measures

The aim of a stability criterion is to study the reliability of the neighborhood relationships shown on a map. Indeed, given two maps generated from the same dataset, with the same network parameters and with the same topology preservation, it may be still not possible to know which one best reflects the neighborhood relationship of two given data if the distance of these data on each map is different. It may also be possible that neither of these two maps correctly reflects the neighborhood relationship of these two data. Topography criteria only indicate the global topology that is preserved on each map. Hence, it is important to know the confidence that we can have in the neighborhood distances shown on a map.

A Stability criterion has been proposed in [5] so as to study the reliability of the topology preservation of data in SOM over several bootstrapped learning procedures. The bootstrap technique is used to generate a new sample of the dataset for each learning procedure in order to take the variability of the dataset into account. The stability criterion is based on a neighborhood distance function $\Gamma^{b}_{k_1,k_2}(\alpha)$, defined as follows:

$$\Gamma^{b}_{(k_1,k_2)}(\alpha)=\begin{cases}0 & \text{if} \quad \delta\left(\Psi\left(Z_{k_1}\right),\Psi\left(Z_{k_2}\right)\right)>\alpha \\ 1 & \text{if} \quad \delta\left(\Psi\left(Z_{k_1}\right),\Psi\left(Z_{k_2}\right)\right)\le\alpha\end{cases}, \tag{1}$$

where b refers to a bootstrapped sample, Ψ is the affectation function which finds the neuron that best matches the input vector Z and $(k_1,k_2)\in N^2$ are the data numbers of the data for which the neighborhood distance δ is calculated. This function evaluates for each pair of data if its neighboring distance on the map is smaller or equal to α. A stability function $\Xi_{(k_1,k_2)}(\alpha)$ then gives the probability for the data Z_{k_1} and Z_{k_2} to be neighbors within radius α over B bootstrapped samples. This function is defined by:

$$\Xi_{(k_1,k_2)}(\alpha)=\frac{1}{B}\times\sum_{b=1}^{B}\Gamma^{b}_{(k_1,k_2)}(\alpha). \tag{2}$$

Histograms of the stabilities $\Xi_{(k_1,k_2)}(\alpha)$ over all pairs of data are then drawn for each neighborhood distance α. A map is perfectly stable with radius α if the histogram only shows two peaks; one at 0 and one at 1. A map is unstable with radius α if the histogram is close to the one that would be obtained in the random case (unorganized maps). The probability $p(\alpha)$ for two data to be neighbors by chance with a neighborhood distance smaller or equal to α, on a map with sizes (ξ_1,ξ_2), can be approximated by:

$$p(\alpha)=\frac{(2\alpha+1)^2}{\xi_1\xi_2}, \tag{3}$$

if the edge effects of the map are not taken into account. Hence, the stability histogram of an unorganized map over B bootstrapped samples for each neighborhood distance α is assumed to follow a binomial distribution with parameters B and $p(\alpha)$. Thus, the stability histogram of a given

map configuration can be compared graphically with the unorganized map stability histogram for each neighborhood distance. This criterion is very useful since it allows the reliability of a map configuration to be evaluated. However, the proposed method is graphical, several stability histograms are needed (one for each neighborhood distance) and the function $p(\alpha)$ is only valid for small values of α.

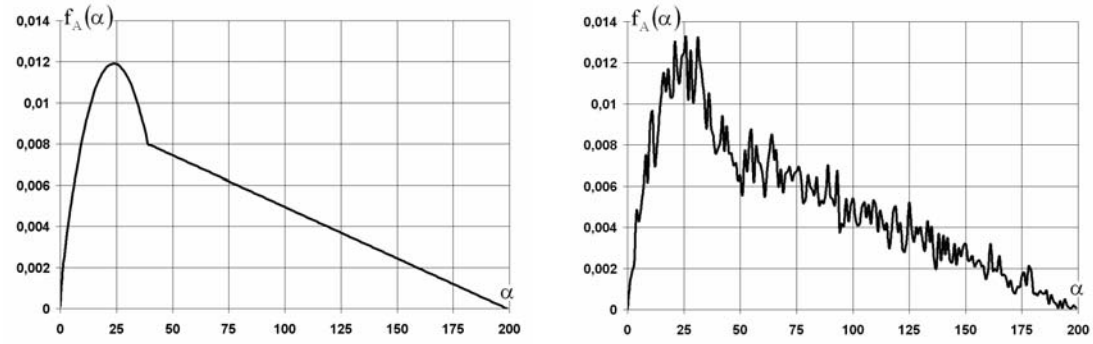
In the following lines, we present the γ_{50} criterion that gives a more accurate measure of the stability of SOM by taking the edge effects of the maps into account. To introduce this criterion, we need to study the statistical properties of SOM first.

3 Statistical properties of SOM

After a long and complex demonstration, we are able to show that the probability function $f_A(\alpha)$ of two data to be neighbors by chance on a 2-D map with a neighborhood distance α is defined, for $\alpha \in N$ and $0 \le \alpha < \xi 2$, by:

$$f_A(\alpha) = \begin{cases} \dfrac{1}{\xi_1 \xi_2} & \text{if} \quad \alpha = 0 \\ \dfrac{2}{{\xi_1}^2 {\xi_2}^2}\left(4\xi_1\xi_2\alpha + 2\alpha^3 - 3\alpha^2(\xi_1 + \xi_2)\right) & \text{if} \quad 1 \le \alpha \le \xi_1 - 1 \\ \dfrac{2(\xi_2 - \alpha)}{{\xi_2}^2} & \text{if} \quad \xi_1 \le \alpha \le \xi_2 - 1 \end{cases}, \tag{4}$$

This probability function has been drawn on Fig. 1 for a map with sizes $(\xi_1 = 40, \xi_2 = 200)$.



(a) (b)

Fig. 1. Probability function of two data to be neighbors by chance on a 2-D map. (a) Theoretical distribution. (b) Experimental results after 10 000 random projections of two data.

We are also able to establish that the mean value of $f_A(\alpha)$ is given by:

$$\mu = \frac{1}{30{\xi_1}{\xi_2}^2}\left(5{\xi_1}^3\xi_2 - {\xi_1}^4 + 5{\xi_1}^2 - 15\xi_1\xi_2 + 10\xi_1{\xi_2}^3 - 4\right), \tag{5}$$

and that the variation value of $f_A(\alpha)$ can be calculated by:

$$\sigma^2 = \frac{1}{900\xi_1{}^3\xi_2{}^4}\Big[-55\xi_1{}^7\xi_2{}^2 + 10\xi_1{}^8\xi_2 - 80\xi_1{}^6\xi_2 + 300\xi_1{}^5\xi_2{}^2 - 100\xi_1{}^5\xi_2{}^4 + 190\xi_1{}^4\xi_2 - \xi_1{}^9 + 10\xi_1{}^7 + 140\xi_1{}^6\xi_2{}^3 - 33\xi_1{}^5 - 400\xi_1{}^4\xi_2{}^3 + 40\xi_1{}^3 - 345\xi_1{}^3\xi_2{}^2 + 50\xi_1{}^3\xi_2{}^6 + 150\xi_1{}^3\xi_2{}^4 - 120\xi_1{}^2\xi_2 + 260\xi_1{}^2\xi_2{}^3 - 16\xi_1\Big]. \quad (6)$$

It has to be noted that if $\xi_1 = \xi_2 = \xi$, (5) and (6) can be rewritten as:

$$\mu = \frac{1}{15\xi^3}\left(7\xi^4 - 5\xi^2 - 2\right), \quad (7)$$

and

$$\sigma^2 = \frac{1}{900\xi^7}\left[44\xi^9 - 20\xi^7 + 72\xi^5 - 80\xi^3 - 16\xi\right]. \quad (8)$$

Moreover, if $\xi \geq 4$, we can write:

$$\mu \approx \frac{7}{15}\xi, \quad (9)$$

and

$$\sigma \approx \sqrt{\frac{11}{225}} \cdot \xi. \quad (10)$$

Finally, we can also show that the distribution function $F_A(\alpha)$ is given by:

$$F_A(\alpha) = \begin{cases} \dfrac{1}{\xi_1{}^2\xi_2{}^2}\left[\xi_1\xi_2(2\alpha+1)^2 - (\xi_1+\xi_2)\left(\alpha+3\alpha^2+2\alpha^3\right) + \alpha^2 + 2\alpha^3 + \alpha^4\right] & \text{if} \quad 0 \leq \alpha \leq \xi_1 - 1 \\ \dfrac{1}{\xi_2{}^2}\left[\xi_2 + \alpha(2\xi_2 - \alpha - 1)\right] & \text{if} \quad \xi_1 \leq \alpha \leq \xi_2 - 1 \end{cases} \quad (11)$$

This distribution function has been drawn on Fig. 2 for maps with sizes $(\xi_1 = 40, \xi_2 = 200)$ and is compared to the distribution function $p(\alpha)$ defined in (3). On these graphics we can see that $p(\alpha)$ is only valid for values of α smaller than 10. Hence, for larger values of α, it is preferable to use the $F_A(\alpha)$ function with the $\Xi_{(k_1,k_2)}(\alpha)$ criterion. However, this criterion is graphical and several stability histograms are needed to study a map (one for each neighborhood distance α).

4 The γ_{50} criterion to evaluate the stability of SOM

As stated in paragraph 2, the aim of a stability criterion is to evaluate the reliability of the neighborhood relationships of data on a map over several learning phases. Indeed, each pair of data has a particular neighborhood distance on a map which may be different from one learning to another.

Hence it is possible to draw for each pair of data, after several learning procedures, a histogram $H_A(\alpha)$ that shows the number of times the data were neighbors with the neighborhood distance α. A pair of data is perfectly stable if its corresponding neighborhood distance on the map remains unchanged over all the learning procedures, that is to say if $\sigma_{H_A(\alpha)} = 0$. The value of $\sigma_{H_A(\alpha)}$ that corresponds to an unstable pair of data, on a map with sizes (ξ_1, ξ_2), is given by (6).

Therefore, an interesting criterion to evaluate the instability of a map is to draw a histogram $H_S(\sigma_{H_A(\alpha)})$ that shows the number of pairs that are unstable with the value $\sigma_{H_A(\alpha)}$ over all the learning phases. However, the use of the criterion $\sigma_{H_A(\alpha)}$ as an indicator of the instability of a pair of data is too restrictive.

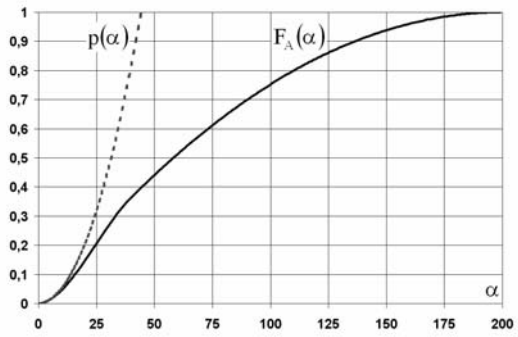
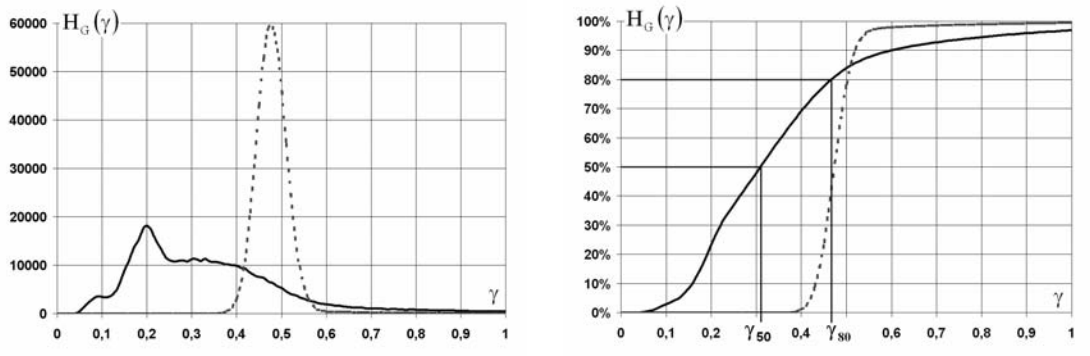


Fig. 2. Distribution functions $F_A(\alpha)$ and $p(\alpha)$ of two data to be neighbors by chance on a 2-D map with a neighborhood distance smaller or equal to α. These functions have been drawn for maps with sizes $(\xi_1 = 40, \xi_2 = 200)$.

In fact, the Kohonen algorithm only tries to preserve the local neighborhood relationships of data. In terms of stability this means that the smaller $\mu_{H_A(\alpha)}$ the smaller $\sigma_{H_A(\alpha)}$ should be. Therefore, a better indicator of the stability of a map is to draw the histogram $H_G(\gamma)$ where $\gamma = \frac{\sigma_{H_A(\alpha)}}{\mu_{H_A(\alpha)}}$.

Instability histograms have been drawn on Fig.3.a for an artificial dataset of 1 000 data, that is to say 499 500 pairs of data. A map is stable if $H_G(\gamma)$ is concentrated close to a peak in $\gamma = 0$. A map is unstable if $H_G(\gamma)$ is close to (or located after) the peak obtained in the random case (unorganized maps). The corresponding theoretical γ value can be calculated for a map of sizes (ξ_1, ξ_2) by using (5) and (6). If $\xi_1 = \xi_2 = \xi$ and $\xi \geq 4$, $\gamma \approx 0.47$.

To obtain non graphical indicators of the stability of a map we use the γ_{50} criterion which represents the maximum instability value of 50% of the most stable pairs of data. It is also possible to use other criteria such as the γ_{70}, γ_{80} and γ_{90} criteria that represent respectively the maximum instability values of 70%, 80% and 90% of the most stable pairs of data. The γ_{50} and γ_{80} criteria have been plotted on Fig. 3.b.



(a) (b)

Fig. 3. Experimental instability histograms $H_G(\gamma)$ that show the number of pairs of data that were instable for each γ value. An artificial dataset of 1 000 data was used. (a) The plain line corresponds to the instability histogram computed for organized maps with sizes $(\xi_1 = 50, \xi_2 = 50)$, the dashed line refers to unorganized maps with same sizes. (b) Cumulated representation of the instability histograms. The γ_{50} and γ_{80} criteria have been plotted for organized maps.

5 Experimental measures of the stability

The γ_{50} criterion defined above has been applied to study the influence of map sizes on the stability of SOM. Indeed, the choice of the right values for the parameters ξ_1 and ξ_2 is often difficult and can modify the topology preservation of the data considerably.

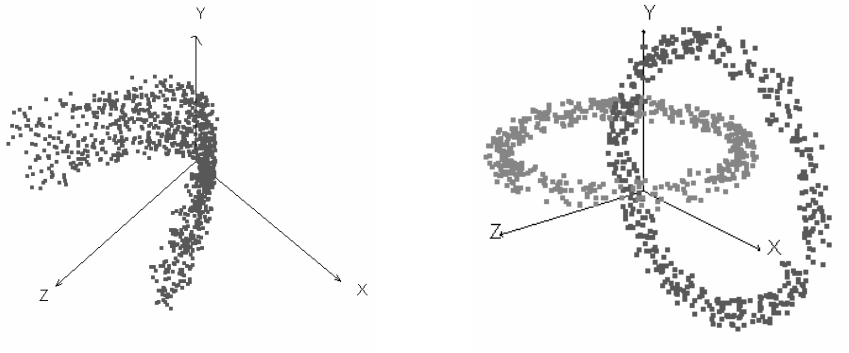


Fig. 4. D1 and D2 datasets

Two artificial datasets have been considered. The first dataset (D1) represents a horseshoe distribution and the second dataset (D2) is composed of two rings (Fig.4). Both of the datasets were generated with 1 000 data. In order to study the influence of the map's sizes, the parameters ξ_1 and ξ_2 were varied with the values {10, 20, 30, 40, 50, 60, 70, 80, 90 and 100} (with $\xi_1 \leq \xi_2$). Hence, up to 55 map configurations have been tested with both datasets. For each configuration, the γ_{50} criterion has been computed over 100 learning procedures and compared to the μ_{E_K} criterion that is an average of the Kohonen cost function defined as:

$$E_K = \sum_i \sum_j \Lambda\left(\delta\left(j, j^*\right)\right) \times \left\| Z_i - W_j \right\|^2, \tag{12}$$

where W_j refers to the synaptic weight vector of the neuron j and j^* corresponds to the neuron that best matches the input vector Z_i.

with $$\Lambda(\delta) = e^{\frac{-|\delta|}{T}}, \tag{13}$$

where the T parameter was fixed to $1/3$ of each network's diagonal. The results obtained for the two datasets have been plotted on Fig. 5. The first row refers to the values of μ_{E_K} and γ_{50} computed with D1 and the second row corresponds to the results obtained with the same criteria and D2.

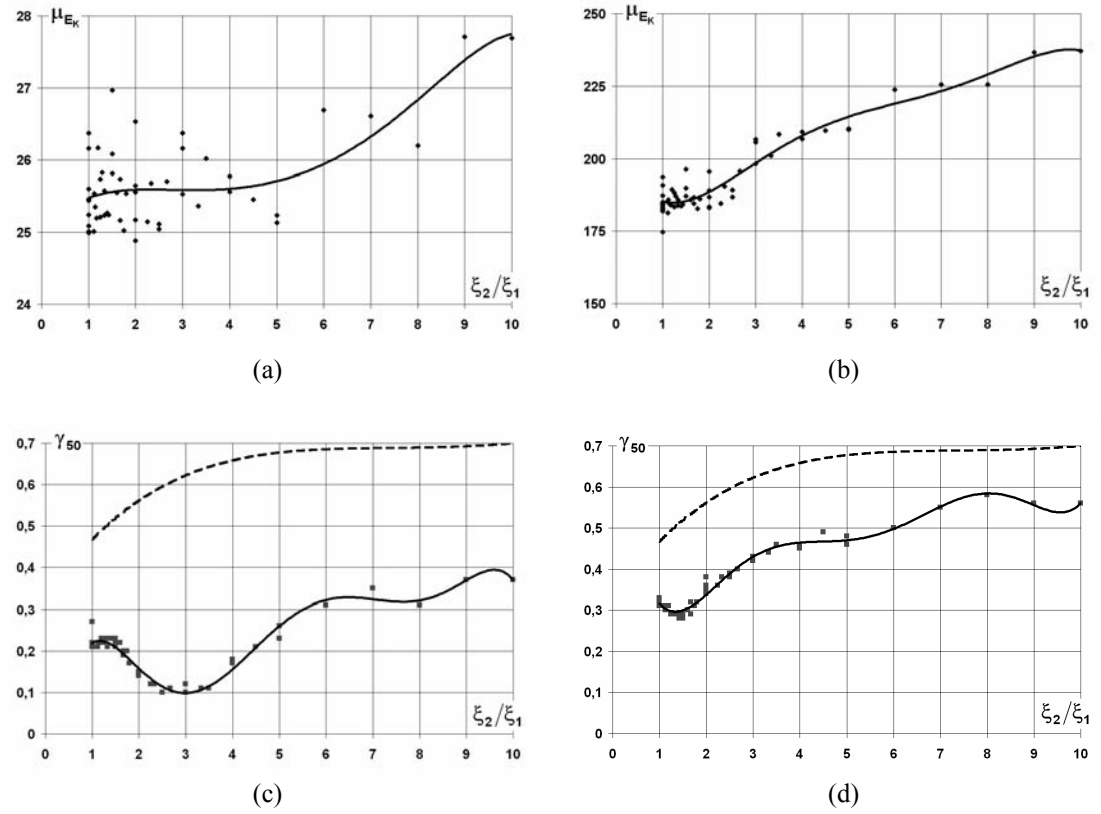


Fig. 5. Influence of SOM sizes (ξ_1, ξ_2) on the Kohonen cost function (μ_{E_K}) and the instability criterion (γ_{50}). The dashed line on graphs (c) and (d) shows the theoretical instability value of unorganized maps.

The figures Fig. 5.a and 5.b indicate the mean topological conservation of the data for a given geometry of the map (ξ_2/ξ_1). Looking at these figures, it is not possible to choose the right geometry for the map. On both graphics a tendency curve has been added but is not significant since the points are very dispersed, especially on Fig. 5.a. Anyway, we can see that the geometry that best preserves on average the topology of the data lies in an interval from 1 to 5 for the D1

dataset and seems to be contained between 1 and 2.5 for the D2 dataset. The use of the γ_{50} criterion gives more accurate information as shown on the last graphics (Fig. 5.c and 5.d) that present the evolution of the instability with the network's geometry. Indeed, we can see that the best stability measures were obtained for size ratios ξ_2/ξ_1 close to 3 for the D1 dataset and for ratios close to 1.5 for the D2 dataset. The lowest instability values that were obtained for each database are quite low: 0.1 for D1 and 0.3 for D2. Hence, the γ_{50} criterion allows the best stable geometry for the network to be chosen. It also gives important information on the instability of the map configurations. Indeed, the theoretical instability values of unorganized maps have been represented in dashed line on Fig. 5.c and 5.d. Thus, it can be seen on Fig. 5.d that organized maps with size ratios higher than 5 seem to be more stable than unorganized maps but the information provided by these maps is globally less reliable than the information given by a network with a size ratio of 1 that has its units positioned randomly. More generally, all the map configurations that give an instability value higher than 0,47 have to be considered as unstable.

6 Conclusion

A new stability criterion γ_{50} was presented to study the reliability of the neighborhood distances shown on a map over several learning procedures of the same dataset. The mean and the standard deviation of each pair of data is computed and used to build a histogram of the γ instabilities. The γ_{50} criterion is then determined and used to evaluate the global instability of the data for a given maps configuration. This indicator is non graphical and gives more accurate information on the instability of SOM than the existing $\Xi_{(k_1,k_2)}(\alpha)$ criterion. Moreover, it allows choosing the geometry of the network that gives the most stable maps. We have shown that SOM present particular statistical properties that are used with the γ_{50} criterion to determine the instability values of unorganized maps. Besides, we have also established that the instability value of an organized map has to be smaller than 0,47.

References

[1] T. Kohonen (1995), *Self-Organizing Maps*, Berlin, Springer.

[2] H.-U. Bauer, M. Herrmann and T. Villmann (1999), Neural maps and topographic vector quantization, *Neural Networks*, **vol. 12**, p. 659-676.

[3] G. J. Goodhill and T. J. Sejnowski (1997), A unifying objective function for topographic mappings, *Neural Computation*, **vol. 9**, p. 1291-1304.

[4] S. Kaski (1997), *Data exploration using self-organizing maps*, Ph.D., dissertation, Acta Polytechnica Scandinavica, Mathematics, Computing and Management in Engineering Series No. 82, Espoo.

[5] E. de Bodt, M. Cottrell and M. Verleysen (2002), Statistical tools to assess the reliability of self-organizing maps, *Neural Networks*, **vol. 15**, p. 967-978.