THE SELF-ORGANIZING MAP OF NEIGHBOUR STARS AND ITS KINEMATICAL INTERPRETATION

M. Hernandez-Pajares, R. Cubarsi, [1] *E. Monte* [2]

Abstract: The Self-Organizing Map (SOM) is a neural network algorithm that has the special property of creating spatially organized *representatives* of various features of input signals. The resulting maps resemble real neural structures found in the cortices of developed animal brains. Also, the SOM has been successful in various pattern recognition tasks involving noisy signals, as for instance, speech recognition and for this reason we are studying its application to some astronomical problems. In this paper we present the 2-D mapping and subsequent study of one local sample of 12000 stars using SOM. The available attributes are 14: 3-D position and velocities, photometric indexes, spectral type and luminosity class. The possible location of halo, thick disk and thin disk stars is discussed. Their kinematical properties are also compared using the velocity distribution moments up to order four.

Key words: *Neural network, self-organizing map, pattern recognition*

Received: October 20, 1992
Revised and accepted: February 23, 1993

1. Introduction

As [14] indicated, most of the methods currently used in observational Astronomy are rather old, and need an urgent updating before they can be used with confidence for the treatment of high quality material provided by orbital observatories. So it is necessary to look carefully at the new trends and tools in the field of Statistics and Information Theory.

One relevant aspect of the studies that are carried out from stellar catalogues, is the segregation of stars in populations in terms of spectral, photometric or kinematic criteria. For instance [17], [22], [24], [5], [6] have also worked on this subject

[1] M. Hernandez-Pajares, R. Cubarsi
Departament de Matemàtica Aplicada i Telemàtica,
ETSETB, Universitat Politècnica de Catalunya, Apartat 30002 - Barcelona, Spain
[2] E. Monte
Departament de Teoría del Senyal i Comunicacions,
ETSETB, Universitat Politècnica de Catalunya, Apartat 30002 - Barcelona, Spain

recently. The viewpoints adopted by most of these have been statistical, numerical or dynamical approaches.

We present in this paper the application of one recent and powerful statistical tool to the problem of classifying one real stellar sample between several astronomical populations: *thin disk, thick disk* and *halo* (see [10]). This tool is the SOM: a classification scheme with an unsupervised competitive learning algorithm proposed by T. Kohonen in the 80's within the artificial neural network field (for instance [15], [16]). The main advantage of the algorithm is that it arranges the resulting groups in an associated bidimensional map, where proximity means similarity between the *global* group properties.

The final suggested groups are studied and tested from a kinematic point of view. Indeed, assuming the superposition of Gaussian distributions for the residual velocity it is possible to estimate properties of the mixed populations at the moments of the separate distributions [6].

2. The Self-Organizing Map

2.1 Fundamentals

SOM is an unsupervised neural classifier that has been applied to astronomical data in [12], [13]. The basic aim of this classifier is finding a smaller set $C = \{c_1, \ldots, c_p\}$ of p centroids that provides a good approximation of the original set S of n stars with m attributes, encoded as "vectors" $x \in S$. Intuitively, this should means that for each $x \in S$ the distance $\| x - c_{f(x)} \|$ between x and the closest centroid $c_{f(x)}$ shall be small. However, the main advantage of the algorithm is that it also arranges the centroids so that the associated mapping $f(.)$ from S to A maps the topology of the set S in a least distorting way. Usually A is a bidimensional set of indexes named *Kohonen map* where proximity between them means similarity between the global properties of the associated groups of stars.

From a detailed point of view, the neural network is composed of a set of p nodes or neurons. Every neuron will represent after training, a group of stars with similar features and the weight vector will be approximately the centroid of these associated stars. Following the concise description of [1], the training process consists of presenting sequentially all the training data in parallel to all nodes. For each training vector, each node computes the euclidean distance between its weight and that vector, and only the node whose weight is closest to the vector, and its neighbours will update their weights by approaching them to the presented datum. So the nodes compete approaching as many as possible the training vectors. Also updating the neighbours' weight instead of just that of the winning mode, assures the ordering of the net [15]. Finally we will have p good representatives of the input space after training with the associated p groups of input data. In addition, weights of nodes which are close within the grid will also be close within the input space.

The detailed algorithm scheme is:

1. We initialize the weights of the p nodes of the grid with small values: $C = \{c_1, \ldots, c_p\}$

2. For each of the n training vector of the overall database, x_i:

   (a) We find the node k whose weight c_k best approach x_i : $d(c_k, x_i) \leq d(c_l, x_i)$, $\forall l \in \{1, \ldots, p\}$.

   (b) We update the weight of the winner node k and its neighbours, $N_k(i)$:

$$c_l(i) = \begin{cases} c_l(i-1) + \alpha(i)(x_i - c_l(i-1)) & l \in N_k(i) \\ c_l(i-1) & l \notin N_k(i) \end{cases} \qquad l = 1 \ldots p$$

   being:

   - $\alpha(i)$ a suitable, monotonically decreasing sequence of scalar-valued gain coefficients, $0 < \alpha(i) < 1$. A good choice is a rapidly decreasing function during, let's say, the first 1000 iterations between 0.9 and 0.1 (ordering period): this function can be lineal. After the initial phase, $\alpha(i)$ should attain small values (≤ 0.01) over a long period. A valid dependence is $\alpha(i) \propto 1/i$.
   - The radius of the activated neighbourhood $N_k(i)$, a monotonically decreasing function of the iteration i. It can begins with an initial fairly wide value, for $N_k(0)$ (e.g. more than half the diameter of the network), and letting it shrink with time during the ordering phase to, say, one unit; during the fine adjustment phase the radius can be zero (only the winner neuron is activated).

3. The process 2 is repeated for the overall database until a good final training is obtained. A *rule of thumb* is that for good statistical accuracy, the number of steps must be at least 500 times the number of nodes.

2.2. Calculations

The observational data considered is the [7] stellar catalogue (see also [8]). It was made from the S.A.O. catalogue that contains all the kinematic and astrophysical information available about more than 250000 stars [19], [20]. The final catalogue contains 12824 stars with enough information to estimate the spatial velocity.The most relevant data for our purposes are the galactic longitude, latitude and the heliocentric distance; the spectral type and luminosity class; the Johnson photometric magnitude and indexes m_v, $B-V$, $U-B$; the spatial residual velocities taking out the simple rotation model in a galactic heliocentric reference frame; and finally, the velocity components in the same reference system as the residual velocity components.

SOM has been applied working in a 14 dimensional characteristic space, i.e., the space formed by the 14 properties described above. We assume the symmetry referred to the galactic plane for the galactic latitude and for the perpendicular to galactic plane residual velocity component, that means considering its absolute values $|b|$ and $|W_1|$ respectively. In the calculations we have taken $8 \times 8 = 64$ centroids to be determined after $4 \cdot 10^6$ training iterations of the neural network

($\approx$ 330 presentations of the entire database). So, the resultant Kohonen map consists of a two-dimensional grid of $8 \times 8 = 64$ neurons, with 14 dimensional centroid vector and an associated group of stars for every node. To evaluate the results, it is interesting to keep in mind that if the j-th characteristic is significant in the segregation problem, then a systematic trend for that characteristic appears in the Kohonen map. The centroids obtained for the stellar catalogue present as the main significant characteristics, the distance and the absolute value of the residual velocity component perpendicular to the galactic plane $|W_1|$. The distance is directly correlated with other significant characteristics such as the spectral type and the luminosity class.

Using the Kohonen map we can segregate the catalogue from an astronomical point of view. Indeed, $|W_1|$ gives us the maximum perpendicular distance to which the star can climb away from the plane. This parameter is related directly with its metallicity and with the population to which the star can belong: disk and halo populations with low and high values, and the recent proposal of a third population, the *thick disk*, with intermediate values of $|W_1|$ [9].

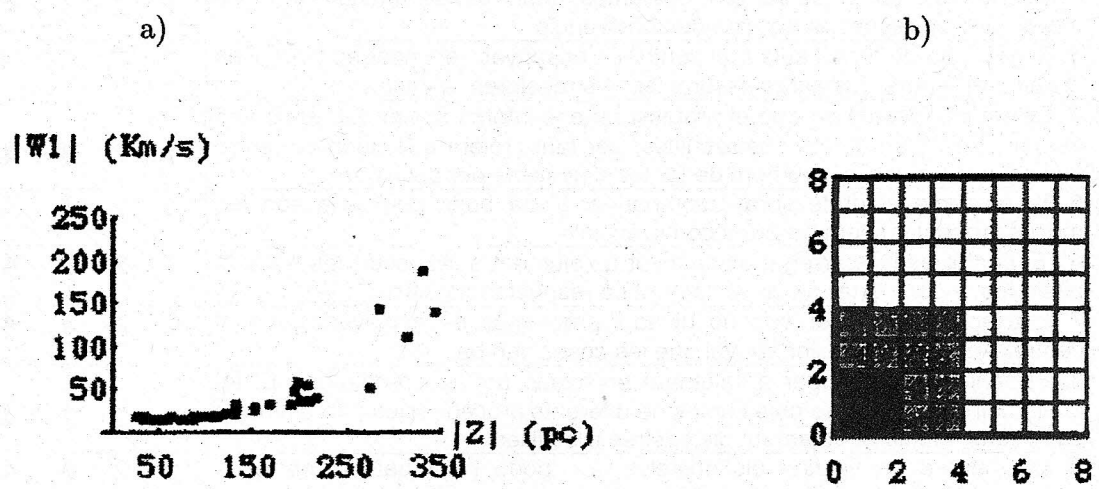

Fig. 1. *In a) $|W_1|$ is plotted again the height above the galactic plane, $|Z|$, for the 64 centroids calculated. In b) the distribution of the three families of centroids in the Kohonen map, regarding the $|W_1|$ characteristic is represented; (A) with $|W_1| \leq 24$ Km/s (white squares), (B) with $24 < |W_1| \leq 60$ Km/s (gray squares) and (C) with $|W_1| > 60$ Km/s (black squares).*

In Fig. 1.a) the $|W_1|$ in function of the distance perpendicular to the galatic plane $|Z|$ appears, calculated as $r \sin |b|$, for the 64 centroids obtained. We can distinguish between three groups of neighbouring centroids in the Kohonen map: (A) with $|W_1| \leq 24$ Km/s (distances basically lower than 550 pc), (B) with $24 < |W_1| \leq 60$ Km/s (distances between $\approx$ 380 and 1400 pc) and (C) with $|W_1| > 60$ Km/s and distances in general greater than 1400 pc (Fig. 1.b). These intervals agree with the kinematic bins considered by [3]; and related to the metallicity and to the Galaxy populations: region (A) with a predominant thin disk component, (B) with the thick disk and (C) with the halo population. On the other hand, the residual velocity moments for the groups (A), (B) and (C) are listed in Tab. 1. The

2nd. order moments are in principle compatible with the accepted values for the thin disk, thick disk and halo (for instance [6]). A detailed study of each group from these moments is done in the next section.

	Group A (a)		Group B (b)		Group C (c)	
	Moment	Error	Moment	Error	Moment	Error
U_0	10.87	0.34	4.5	1.3	27.1	4.7
V_0	18.09	0.26	13.0	1.1	3.6	3.7
W_0	7.65	0.20	5.0	1.2	21.6	5.7
μ_{11}	1257	31	2820	150	8000	700
μ_{22}	735	24	2110	130	5040	470
μ_{33}	435	16	2440	130	11900	890
μ_{12}	115	18	-	-	1200	400
μ_{112}	-237	23	-	-	-	-
μ_{222}	-363	39	-	-	-	-
μ_{233}	-87	14	-	-	-	-
μ_{1111}	1187	95	4820	580	24600	4300
μ_{1122}	328	30	1570	220	6180	880
μ_{2222}	674	82	3310	580	10600	2000
μ_{1133}	181	18	1310	190	8300	1100
μ_{2233}	158	18	1060	120	5230	640
μ_{3333}	305	60	3460	420	43400	5700

Tab. 1. *The mean residual velocities, U_0, V_0, W_0 and the non-vanishing central moments (3-sigma level) of order two, three and four $\mu_{ij}, \mu_{ijk}, \mu_{ijkl}$ respectively, with the associated errors are listed for the groups A, B and C of stars (those with residual velocity greater than 300 Km/s have not been taken into account). The units for the 2nd, 3rd and 4th order moments are* $(\mathrm{Km/s})^2$, $10^2\ (\mathrm{Km/s})^3$ *and* $10^4\ (\mathrm{Km/s})^4$ *respectively.*

3. Kinematical Interpretation of the Stellar Groups

From the total stellar catalogue, three stellar groups have been obtained with some common features. The interpretation of the resulting subsamples has been done according to the average properties of the stars belonging to each group. In particular, the spatial velocity perpendicular to the galactic plane and the heliocentric distance. Nevertheless, since the central velocity moments up to fourth-order of each stellar group have been computed, it is also possible to analyze their kinematical features by means of other statistical criterions and, thus, to compare and interpret the resulting SOM segregation.

The method of analysis used in this section is based on the fact that, according to gas dynamics, the velocity distribution of a galactic component can be locally described with a multivariate Gaussian function of the peculiar velocities. It is easy to identify this type of distribution, since they satisfy the following two main properties [21]: (1) The odd-order central moments are zero; in particular, this is

significant for the third moments. (2) The excess of the distribution is null in the three velocity components.

However, according to the features of the foregoing stellar groups, we shall focus in the kinematical interpretation of the groups A and B, which can be clearly considered as local samples. In fact, the characteristic scale heights over the galactic plane for the galactic components of the disk, thin disk and thick disk populations, are [10] around 0.3 kpc and 1.3 kpc respectively. Moreover, the central moments of these stellar groups (Tab. 1.) can not be explained from a Gaussian velocity distribution alone, since the third moments are not null, and the second and fourth moments lead to a positive excess of the distribution in all the velocity components. Thus we may suspect that the A group corresponds to a typical local stellar sample containing a main population from the thin disk, being contaminated by a significant number of stars belonging to the thick disk population [4], [25]. Similarly for the B group central velocity moments, we may think in a main thick disk population with some contamination of halo population.

The central velocity moments resulting from a superposition of two Gaussian distributions are not arbitrary. The main properties of such distributions [6] can be summarized as follows: (1) Concerning the odd-order moments (in this case they are not null), there exist four vanishing linear combinations of the third moments. (2) The second, third and fourth moments are constrained by a set of fourteen relationships, providing the fundamental superposition parameters, which are associated with the skewness and the peakedness of the distribution. These properties lead us to compute the partial second moments of the stellar components, the mean velocity difference between them, and the percentage of mixture.

If we apply this method by assuming that the A group is a superposition of two Gaussian components, a mixture with a percentage of 90 % ($\pm$ 6%) for the first stellar component is obtained. This takes into account around 9535 stars for the first component and 1060 stars for the second one. The non-vanishing partial second central moments and the non-vanishing vector components of the mean velocity difference are listed in Tab. 2a. Let us note that the non-diagonal partial moment corresponding to the indices 12 is nearly zero, and the other nondiagonal partial moments are null. This fact must be interpreted as an almost insignificant vertex deviation of the velocity ellipsoids associated with the stellar components. Moreover, it must be pointed out that, besides the expected differential movement in the rotational direction between the stellar components, a significant differential radial movement has been obtained. The velocity dispersions (square root of the diagonal moments) of the first component (27:15:11) are in accordance with thin disk population [23], and those corresponding to the second component (72:55:56) are quite in agreement with the thick disk population. However, this second component is biased towards old disk stars, or is likely contaminated by some halo stars.

Similarly, if we assume that the B group is a superposition of two Gaussian distributions, a mixture with a percentage of 95% ($\pm$ 4%) for the first component is obtained. The non-vanishing partial second central moments and the non-vanishing vector components of the mean velocity difference are listed in Tab. 2.b. In this case also the non-diagonal partial moments are nearly null. The velocity dispersions of the first component (46:40:42) are in total accordance with the typical thick

disk population [10]. For the second component, the dispersions (130:110:110) correspond to the values of halo population [2], [18]. The result implies that the stellar group B consists in a nearly pure thick disk population with 1590 stars, contaminated with less than 85 halo stars. Moreover, let us note that the z-component dispersion is nearly the same than in the stellar group C, a typical halo population value.

At this point it seems interesting to know what would be the percentage of mixture in the second component of the A group in terms of both components obtained in the B group. By applying the superposition criterion we find that this A group second component is composed of a 82% (± 5%) of pure thick disk stars, with dispersions 46:40:42, and the remaining halo stars, with dispersions 130:110:110. Therefore, the stellar group A should be composed of three stellar components with 9535 thin disk stars, 870 thick disk stars, and 190 halo stars.

	Group A (a)		Group B (b)	
	t	$T+h$	T	h
% of population	90	10	95	5
μ_{11}	710	5200	2100	17000
μ_{22}	220	3000	1600	12000
μ_{33}	130	3100	1800	12000
μ_{12}	45	130	-	-
μ_{13}	-	-	-	-
μ_{23}	-	-	-	-
w_1	12		-	
w_2	50		55	
w_3	-		-	

Tab. 2. *The non-vanishing partial second central moments and the vector components of the mean velocity difference are listed for the stellar groups A and B. The symbols for the subsystems are t =thin disk, T =thick disk and h =halo stars.*

4. Conclusions

In this paper we have applied the Self-Organizing Map method to the study of a stellar catalogue that contains 3D positions, jointly with spectral, photometric and kinematic data for a total of more than 12000 stars in the solar neighbourhood. We have found the existence of three regions of neighbouring centroids in the resulting Kohonen map from the $|W_1|$ attribute. Another important feature in the classification has been the distance. Hence the resulting groups present properties related with the *locality*. These three regions seem to correspond basically to the thin disk (A), thick disk (B) and halo populations (C) also taking into account the respective residual velocity moments. This result has been contrasted for A and B assuming that every sample is a superposition of two Gaussian components for the velocity distribution. Under these assumptions, the analysis gives a purity of 90 and 95 percent for the samples A and B. Also is possible to characterize the

efficiency of SOM in this kind of astronomical problems applying that algorithm to synthetic samples (see [11]).

Acknowledgments

This work has been supported by the D.G.C.I.C.I.T. of Spain under Grant No. PB90-0478.

References

[1] Cabrera A., Cid J., Hernández A.: Artificial neural networks. Lecture Notes in Computer Science, **540**, Springer-Verlag, Berlin, 1991, 401.

[2] Carney B.W., Latham D.W.: Astron. J., **92**, 1986, 60

[3] Carney B.W., Latham D.W., Laird J.B.: Astron. J., **97**, 1989, 423.

[4] Chiu L.T.G.: Astron. J., **85**, 1980, 812.

[5] Cubarsí R.: Astron. J., **99**, 1990, 1558.

[6] Cubarsí R.: Astron. J., **103**, 1992, 1608.

[7] Figueras F.: Ph.D. thesis. University of Barcelona, Barcelona, 1986.

[8] Figueras F., Núnez J.: Astrophys. and Space Sci., **177**, 1991, 483.

[9] Gilmore G., Reid N.: Monthly notices roy. Astronom. Soc. **202**, 1983, 1025.

[10] Gilmore G., Wyse R.F.G.: The galaxy. D.Reidel Publishing Company, Dordrecht, 1987, 247.

[11] Hernández-Pajares M., Comellas F., Monte E., Floris J.: In: Heck and Murtagh, Eds., Astronomy from Large Databases II, Eur. Southern Obs., 1992 (forthcoming - proc. of conf., Haguenau, France, 14-16 Sept. 1992).

[12] Hernández-Pajares M., Monte E.: Artificial neural networks. Lecture Notes in Computer Science, **540**, Springer-Verlag, Berlin, 1991, 422.

[13] Hernández-Pajares M., Monte E.: The Stellar Populations of Galaxies (B.Barbuy and A.Renzini, eds.), IAU Symposium 149, Kluwer Academic Press, Dordrecht, 1992, 430.

[14] Jaschek C.: HIPPARCOS: Scientific Aspects of the Input Catalogue Preparation II (J.Torra and C.Turon, eds.), 1988, 97.

[15] Kohonen T.: Self organizing and associative memory. Springer Series in Information Sciences, Springer-Verlag, Berlin-Heidelberg, 1989.

[16] Kohonen T.: Proceedings of the IEEE, **78**, 9, 1990, 1464.

[17] Kondratev B.P., Ozernoy L.M.: Astrophys. and Space Sci., **84**, 1982, 431.

[18] Norris J.E.: The galaxy, 297, D.Reidel Publishing Company, Dordrecht, 1987, 297.

[19] Ochsenbein F.: CDS Inf. Bull., **19**, 1980, 74.

[20] Ochsenbein F., Bischoff M., Egret D.: Astronom. Astrophys. Suppl. Ser., **43**, 1981, 259.

[21] Orús J.J.: Apuntes de dinámica galáctica. University of Barcelona, Barcelona, 1977.

[22] Ros R.M.: Rev. Mexicana Astronom. Astrofís., **11**, 1985, 23.

[23] Strömgren B.: The Galaxy, 229, D.Reidel Publishing Company, Dordrecht, 1987, 229.

[24] Wyse R.F.G., Gilmore G.: Astronom. and Astrophys. **60** (1986), 263.

[25] Wyse R.F.G., Gilmore G.: Astron. J., **91**, 1986, 855.