ANNALS of Faculty Engineering Hunedoara
International Journal of Engineering
ISSN: 1584 - 2665

COMPUTATIONAL TAXONOMY FOR IDENTIFICATION OF "MOVING GROUPS". APPLIED TO OPEN CLUSTERS

Rosa Beatriz ORELLANA[1], Gregorio PERICHINSKY[2] , Luís Ángel PLASTINO[3]

[1.] Faculty of Astronomical and Geophysical Sciences, Institute of Astrophysics of the National University of La Plata (CCT La Plata - CONICET)
[2.] Faculty of Computer Sciences, Institute of Physics of the National University of La Plata
[3.] Faculty of Sciences, Institute of Physics of the National University of La Plata
La Plata, Buenos Aires, ARGENTINA

Abstract
A new method is presented here to identify the members of an open cluster, based on computational taxonomy: “the spectral method”. The characters used for the analysis are position and proper motion of all the stars belonging to the cluster’s region. Taxonomy allows to groups the stars together (OTU’s) in clusters, based on its genotypic characteristic, which shows the similarity, between two or more Stars (OTU’s). The method has been applied to open cluster NGC2516. The outcome list of members agrees very well with the one obtained by applying other methods.

1. INTRODUCTION

The first statistical studies of stellar proper motion to determine their ownership to a cluster were suggested in 1958 [10]. The model consists on the sum of two Gaussian distributions corresponding one to the stars of the cluster and the other one to the stars of the field. Later on in 1971 [9], the method is standardized applying the principle of maximum similarity constituting the "parametric method". Other authors have carried out improvements to this method with the purpose of diminishing the error in the identification of the members.

In 1990 [1] and in 1998 [3] a denominated method "not-parametric method" is applied, it consists in to determine empirically the function of distribution of the proper motion.

In this work that we present it proposes a new method to identify the members of an open cluster applying the Computational Taxonomy to the positions and motions characteristic of the stars.

In this work we intend a new method to identify the members of an open cluster applying Computational Taxonomy to the positions and characteristic motions of the stars. To this method we have denominated it "spectral method" because as a result a spectrum is obtained that will be common to all the members of the cluster [4].

2. SPECTRAL METHOD

The open cluster are stellar concentrations where their members have similar characteristics and the identification of the same ones is necessary for to approach, for example, studies on the dynamics of the galaxy.

The spectral method that was developed in extensive [5] [6], it uses taxonomic procedures to identify the members of a cluster by means the positions and proper motions of the stars of the region. This task is fulfilled gathering the stars according to the degree of similarity and resemblance in function of the values of its characters (positions (α y δ) and their corresponding proper motions (μ^{α} y μ^{δ})).

It is assigned each star a number (i) and each one of their characters will be identified with the star's number and another number (j) that will vary from 1 to 4, according to the character.

It is defined the degree of similarity among the stars by means of a Euclidean distance that one obtains starting by means the values of the normalized characters. The normalization of the values of

each character is obtained starting through its mean value and of its variance. The Euclidean distance normalized (d_n) it allows to visualize the degree of a star's similarity regarding the other ones in its "characteristic spectrum" (Figures 1 and 2), where the ordinate represents the distance dn. The members of the cluster are obtained applying the theorem of Tchebycheff and the inequality from Bienaymé-Tchebycheff to the Euclidean distances normalized, $d_n = \sqrt{k}.\ \sigma_d$, where σ_d is the variance of the distance d_n and k the constant of Tchebycheff that applying the Principle of Entropy Maximum is calculated [7] [8]. All the stars members of the cluster will be those for which ones $d_n\ \sqrt{2}.\ \sigma_d$, (k = 2). All the doubtful stars will correspond to the uncertainty region for those which ones

$$\sqrt{2}.\ \sigma_d < d_n \leq 2.\ \sigma_d,\ (2 < k \leq 4).$$

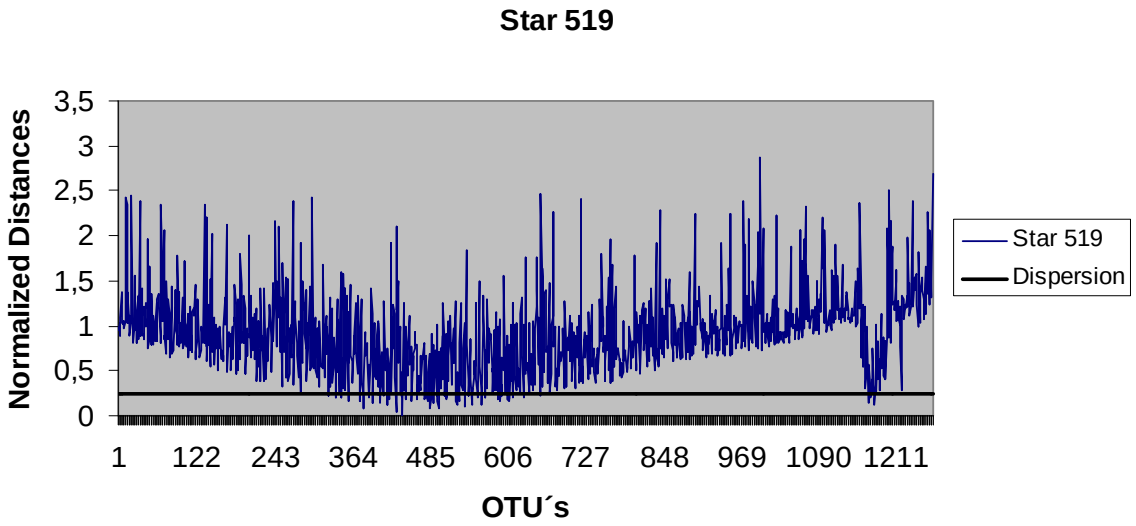


Figure 1. Spectrum of a star member of the cluster NGC2516. The line indicate the value of d_n for k = 2, boundary

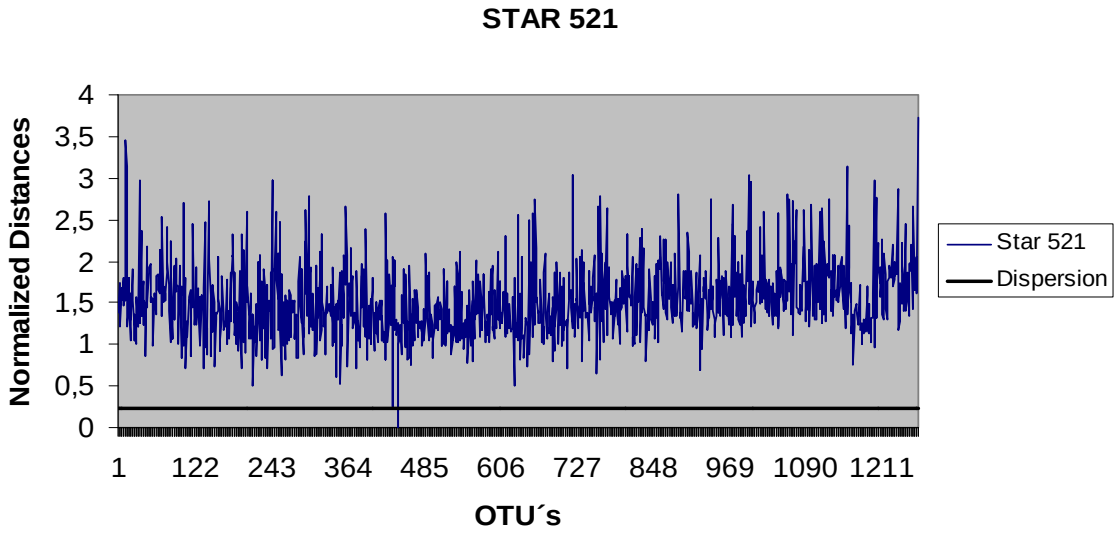


Figure 2. Spectrum of a star that belong to the field of NGC2516, not a member of the cluster. The line indicate the value of d_n for k = 2, boundary

In the Figures 1 and 2 the line indicates the value of d_n corresponding to k = 2, it is denominated boundary. The characteristic spectrum of a star non member (Figure 2) it doesn't show values of d_n under of the boundary line, what indicates that this star is not linked with some other star of the region. The characteristic spectrum of a star member (Figure 1) it shows values of d_n below the boundary line, what indicates that this star is linked with those stars of the region, having all them characters with value similar, constituting the members of the cluster. Therefore, the characteristic spectrum of a star member of the cluster shows all its members clearly [4].

2.1. Spectrum of taxonomic evidence

The classification is a technique of abstraction used for grouping objects with common properties. This allows defining the domain of objects, under the hypothesis that each object belongs to a class (and alone one class) and that, for each class, there is an object that belongs to her at least.

The search of classificatory concepts that it allows a classification structure that it is not modified with the incorporation of new information (stability of the classification), neither is it altered for the incorporation of new entities, which is a motivation for us to search new analytic tools. The association of concepts in systematic way, to classify, using numeric variables, they are technical mathematical which ones belong to the Computational Taxonomy, resemblance and similarity between taxonomical units and the cluster of those units in taxons (taxa as taxon plural, cluster, family), based on the state of its characters.

The objects, to those that are denominated operational taxonomical units (OTU), they are classified according to a process based on the properties of the same ones. The difference among them is the source of the taxonomic evidence.

A character could be defined as any property that characterizes the OTU in study. The states are the set of possible values of their characters. To estimate the taxonomic similarity we use the coefficient of similarity to quantify this resemblance, which is to say to obtain the resemblance for each pair of OTU's of the basic matrix of data.

In order to work in a novel way with this matrix we will introduce here Technical of Theory of the Information.

The Computational Taxonomy allows to group, through the numerical method called structural analysis of clusters (clustering), taxonomic operational units (OTU's) in taxa or groups of OTU's in function of its characteristic states: value, domain, attribute. The clusters constitutes families whose first structural analysis is based on its phenotypical characteristic, taxonomical structures that show the relationship between two OTU's or groups of OTU's in degree of similarity.

In this way, the method allows to define the domain of the objects under the hypothesis that each object belongs to a class, and in each class there is an object that belongs to her at least. The classificatory concept gives a structure that has stability, and the incorporation of new information doesn't alter the same one [5].

The treble constituted by the value, the domain and the attribute in taxonomic methods allows to group the OTU's, in families or clusters whose structural analysis has as base the phenotypic characteristic. We infer the analogy between the taxonomic representation, entity-relation model and databases (in dynamic relational databases) and develop an algorithm to obtain clusters or families. Moreover, we introduce new and original concepts about characteristic spectra of OTUs and families [6]. (See The Annexed of Analogies).

The clusters associated for their degree of similarity, based on the Euclidean distance among OTU's, the technique of "nearest neighbors" among OTU's and pairs of OTU's. In the taxonomical space this clustering method defines taxonomic groups that can be visualized as "characteristic spectra of an OTU" and "characteristic spectra of the families" and it allows the obtaining of Invariants (centroid, variance and radio). The groups are defined through the theorem of Tchebycheff and of the maximum of the inequality of Bienaymé-Tchebycheff [7].

The OTU's take values of the dynamic domains of attributes that form entities that go changing according to the taxonomical necessities: to classify to form families or clusters.

The conformation of groups (clusters) based on the matrix of similarity, Euclidean distances among OTU's, metrics of Minkowski or Manhattan and the application of the technique of "nearest neighbors" it is carried out by means of OTU's that are associated for its degree of similarity. The objective is not to only show relationships of similarity among pairs (pair-group method) of OTU's, but among all the OTU's (construction of the matrix of similarity).

The source of the taxonomic evidence is achieved beginning with the quantification of the coefficients of similarity, the resemblance for each pair of OTU's of the basic matrix of data.

The concept of spectrum of the states of the characters of the pairs of OTU's emerge regarding the total and the formation of the spectrum of families for the overlapping principle when processing the spectra to the right and to the left of the pairs of OUT´s and the obtaining of Invariants (centroid, variance and radio).

2.1.1. Normalization

It is useful to consider the effects taken place by the operations of changes of scale for operators of distances and angles regarding coordinated of reference and their correlation regarding the change of the origin.

In the normalization of characters the average and the standard deviation of each row are computed (the states of each character) and we express each state like a deviation of the average in

units of standard deviation. The normalization of the states of the character does that the average of all character is of value zero and variance of unitary value.

$$\overline{Xj} = (\sum_{i}^{n} \text{Xij}) / n$$

$$\sigma j = ((\sum_{i}^{n} (\text{Xij} - \overline{Xj})^2) / (n - 1))^{1/2}$$

$$\overline{X' ij} = (\text{Xij} - \overline{Xj}) / \sigma j$$

If we want to add a new OTU we can calculate the values standardized from the previous average and the respective standard deviations, although the resulting value won't really be corrected. When ones few OTU's are added they don't constitute a serious problem, because the average and the variance would not see each considerably altered.

In a more general sense we can argue that the variation contributes to most of the information, and that the gross size of the character and the variation range should contribute little to the phenotypical resemblance, from the point of view of the relative information to the taxonomy (the equiprobability produces the maximum entropy) [7].

2.1.2. Matrix of Similarity

During a procedure of accumulative sequential clustering the arbitrary values decrease in a predetermined way; and we extend this method to define a generalized function of distance

$\mathbf{D_{j-k} = [(\overline{X}_k - \overline{X}_j)' S_J^{-1} (\overline{X}_k - \overline{X}_j) |S_J|]^{(1/2)}}$ where $\overline{X}_j$, and $\overline{X}_k$ they are the vectors column that represent the average for the clusters J and K, respectively, for n variables, S_j it is the matrix of variance of these variables for the cluster J, and $|S_J|$ it is their determinant, the generalized variance, for several OTU's.

Using Euclidean distances (or with metric of Manhattan) we can compute the matrix of Taxonomic Distance or of Similarity or of Resemblance or Matrix of Coefficients of Similarity, Matrix by means of which we want to find the taxonomic structure {Sij} of dimensions (t x t) where t is the number of OTU's.

The clusters are the sets of OTU's in the hyperspace, for patterns phenotypical.

The center of the cluster or centroid represents an average object that is simply a mathematical construction that allows the characterization of the Density and the Variance, and the radius and range of the taxon.

The positions of the OTU's can be represented in a system of coordinated, if this positions they are near, the distance decrease until being made zero if they coincide, the distance can be seen this way as the complement of the similarity, being able to prove algebraically that the theorems of the geometry are completed in a Euclidean hyperspace of n dimensions.

Starting from the normalized domains the taxonomic distance is calculated where they can be considered the metric of Minkowski and of Manhattan.

2.1.3. Thus it arrives to the Matrix of Similarity

Each row j of the matrix of similarity contains the distances among the one OTU_j and all the t-1 OTU's remaining.

These similarities depend on the values or states of the characters for the contribution that they do at the distance among those OTU's.

The distribution of the OTU's in the taxonomic hyperspace allows us to visualize the accumulation of the same ones, for vicinity that is to say, next or near neighbors (nearest neighbor) for the method of relationships of similarity among pairs (pair-group method) of OTU's.

In the structuring analysis, clustering, the families for **agglomerate** or assembled of the OTU's, they go taking place by means of the aggregation, until covering them integrally, of a quantity of smaller subsets that t. associate Partitions that are **overlapped** not in all the cases, that is to say that if an OTU belongs to a partition disjointed or associated.

2.1.4. Characterization of the Spectra of similarity

Considering a characteristic spectral to the states of the characters or attributes of the OTU's, under conditions defined by the **Superposition Principles and Interference**, the new concepts of **Spectra of objects** and **Spectra of families are introduced**.

In the taxonomic space this clustering method defines taxonomic groups that can be visualized as characteristic spectra of an OTU and characteristic spectra of the families.

We define as **spectrum individual taxonomic** to the group of distances of an OTU regarding the others OTU's of the group, where each one contributes with the states of the characters and

therefore it is constant for each OTU, under the same taxonomic conditions (in analogy with the phasors).

We define as **spectrum of taxonomic similarity** to the group of distances of the OTU's regarding the others OTU's of the group that determine the constant characteristics of a cluster or family, under the same taxonomic conditions.

The first importance of having a spectrum individual taxonomic, emerge of the fact of being able to study the properties of the individuals through the contribution of the states of its characters and in second instance, the study of the taxonomic structure when conforming clusters or families.

It is solved in this way the taxonomic evidence and they can be invariants that characterize to each cluster, such as: variance, radio, density and centroid.

These invariants are associated to the spectra of taxonomic similarity that identify each family.

2.1.5. Dispersion

It is necessary to have parameters that of an idea of how scattered, or concentrated, are their values regarding the average in variance units (σj).

The variance is a moment of second order and represents to the moment of inertia of the distribution of objects (mass) with respect to their center of gravity (the so-called centroid).

$$\overline{X'ij} = (Xij - \overline{Xj}) / \sigma j$$

is a normalized variable that represents the deviation of the Xij with respect to their average (in units of σj).

As usual, is takes the dispersion to be given by the variance σ^2_d , the mean-squares method is applied.

Let g (Xij) be a not negative function of the variable Xij, for all k > 0 will have the probability function:

$$P [g (Xij) \geq K] \leq (E (g (Xij)) / K.$$

Let S be the set of all the Xij that satisfy the inequality g (Xij) ≥ K, that emerge truly of the theorem Tchevicheff stems from the relationship (valid in any number of dimensions):

$$Eg(Xij) = \int_{-\infty}^{\infty} g(Xij)dF \geq K\int_{S} dF = KP(S)$$

If g (Xij) = (Xij - $\overline{Xj}$)² , K = k² σj², which leads, for all k > 0, to the inequality of Bienaymé-Tchevicheff.

In particular, for a distribution of an average value $\overline{Xj}$ and deviation σj that has one mass 1 / 2.k² located at each the points Xij = $\overline{Xj}$ ± k. σj and [1 - (1 / k²)] and in a point of mass OTU Xij = $\overline{Xj}$ one has: P (| Xij - $\overline{Xj}$ | ≥ k. σj) = 1 / k² a maximal limit (upper) value of the probability that can not be improved (from the Tchebycheff theorem and the inequality of Bienaymé-Tchebycheff).

This inequality shows that the quantity of mass of the distribution is to be found in the interval.

The inequality allows fixing the levels limits of the distribution and it allows fixing the radius of a cluster and the value of k allows fixing iterations in the algorithm [8].

2.1.6. Variation Range Normalization

There exist sound reasons, that considering that the weight of a character should be inversely proportional to its variability. For normally distributed quantitative characters their information content (in the information theory sense) is proportional to the variance. If the variances are equals, then each character contributes an equal informational amount. An uniform probability (**equiprobability**) produces, of course, is possible a maximum **entropy**.

In a more general sense we may argue that the variation contributes most of the information, and that the gross character size and range of variation should contribute little toward phenotypical resemblance, in terms of that information relevant for taxonomic purposes.

3. RESULTS AND DISCUSSIONS

The Open Cluster NGC2516 ($\alpha_{2000,0}$ = 7h 58m, $\delta_{2000,0}$ = - 60° 45') that it is find to some 400 pc of the sun it has a diameter of 35', and it was reexamined applying the spectral method, their parameters were determined and their members were identified..

The position data and proper motions of the 1554 stars of the region were obtained of the catalog UCAC2 [11]. Once applied the spectral method, 135 stars were found in the region defined by the boundary where inside it they are members of the cluster, 13 uncertain stars as part of the uncertainty

of the region and the following parameters for the coordinates and proper motions of the cluster

$$\alpha = 7h\ 58m,\ 0{,}3164^s \pm 0{,}132^s \qquad \delta = -60^o\ 45'\ 39{,}69'' \pm 2{,}19^s$$

$$\mu_\alpha = -2.48 \pm 0{,}27 \text{ mas/year}, \qquad \mu_\delta = 9.74 \pm 0{,}22 \text{ mas/year}$$

The method also allows obtaining the characteristic spectrum corresponding to each star of the region. The Figures 1 and 2 show the spectra corresponding to a star that is a member of the cluster and to a star that is not a member of the cluster. It is notorious that when a star is member her spectrum crosses the line corresponding to k = 2 and when a star is not a member her spectrum does not reach the line. The analysis of all the spectra concludes that a unique spectrum that characterizes the members of the cluster exists for all the stars of the field, and that the same one is similar to the Figure 1. It is notorious that when a star is member her spectrum it crosses the line corresponding to the dispersion and when it is not it doesn't reach to the same one.

3.1. Error subject

From the point of view of the Tchebycheff theorem and their inequality with Bienaymé the variance quantify the dispersion medium from the values of the variable regarding their central value (average, centroid, $\overline{Xj}$), it is the existence of a relationship among the probability of being "far" of the average and the variance or the deviation or standard deviation, in equiprobability for application of the Principle of Maximum Entropy, reason why the "k" of Tchebycheff it takes the value 2, ($\sqrt{2}$), although their maximum pass per 3, ($\sqrt{3}$) and finally 4, ($\sqrt{4} = 2$) for boundary problems, because to maintain constant the density in front of the contraction or dilatation of the space when objects have been added or they have gotten lost (stars= objects).

In a space with metric, or measurable, with an interval ($\overline{Xj}$ $\pm \sqrt{k} \cdot \sigma$), for the range formed by the product of the constant of Tchebycheff (k), for $\sqrt{k}$ for the radius of the cluster (σ) plus the average of the distance $\overline{Xj}$, that it produces an error E if we subtract him the space occupied by the cluster, given by the sum of the radio (σ) plus the average distance $\overline{Xj}$:

$$E = [\ (\sqrt{k} \cdot \sigma + \overline{Xj}\ ) - (\sigma + \overline{Xj}\ )\ ]$$

this for k = 2 and the cluster treaties with 135 stars, in our case, it gives 0.035 (3.50%) that is a quantity of 3 object-stars in the computation (138), they were carried out several tests for errors of 1.0%, 1.5% and 1.75% do not produce changes, newly with a cardinal of 135 objects with a 2.0% it produces one object-star of uncertainty and like was said with a 3.50% it gives an uncertainty of 3 object-stars.

It is necessary to make notice that 1.75% corresponds a k = 3 of Tchebycheff that it includes the circular crown for overlapping the uncertainty, of a cluster in the boundary and vicinity of another region is, to say in the computed range gives 148 object-stars, with 13 object-stars in the border vicinity more far from centroids, for the 135 objects. This value is if we take 3.5% of 148 it gives 5 objects and not 13 objects, that is to say that the method produces less uncertainty than the expected Error.

It is necessary to consign that the systematic error one didn't keep it in mind because it gives less than 10^{-4}.

Remember that the method applied to other cases and in particular to asteroids and using Data Mining to verify the Robustness of the method gives approximately 3% too.

Finally: Error = Range of the Region - Space of the Cluster.

When is compared the proper motions obtained for the cluster applying the spectral method, with those obtained by other authors, it is observed that our results are coherent with those found ones in the literature (see Table 1).

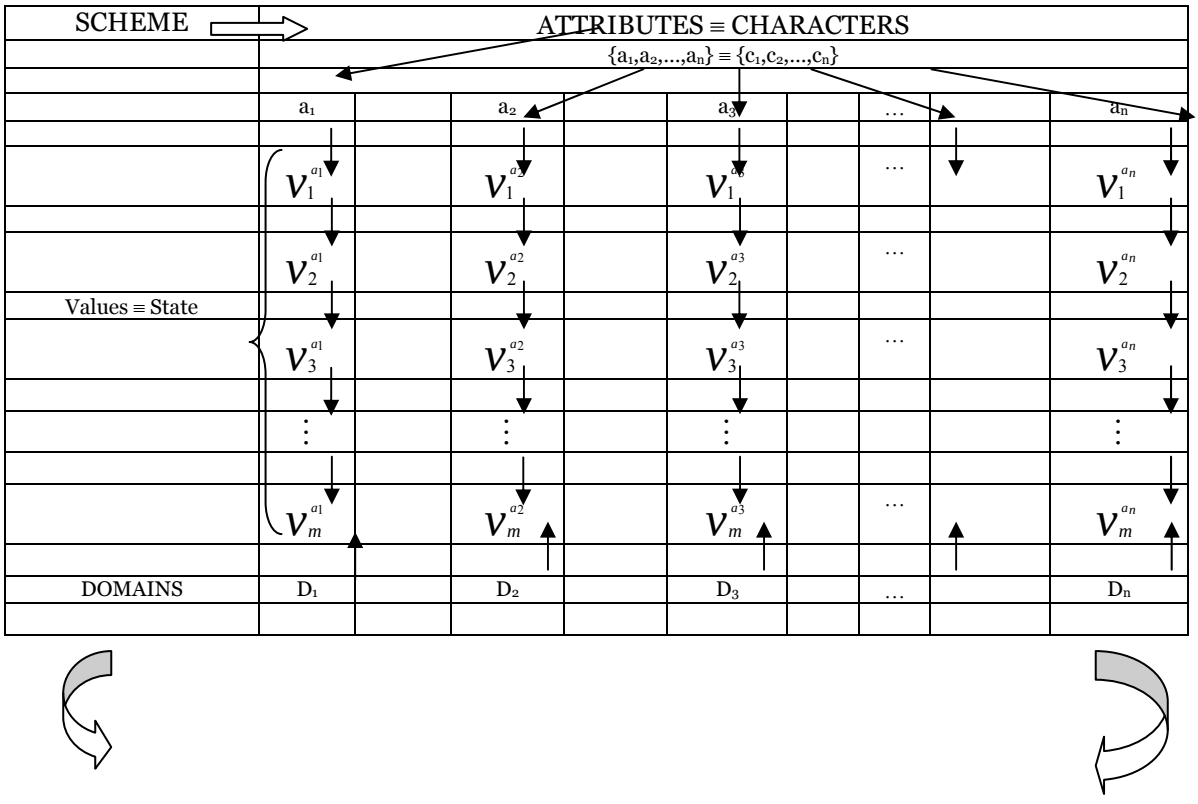
Table 1. Proper Motions of the cluster NGC2516

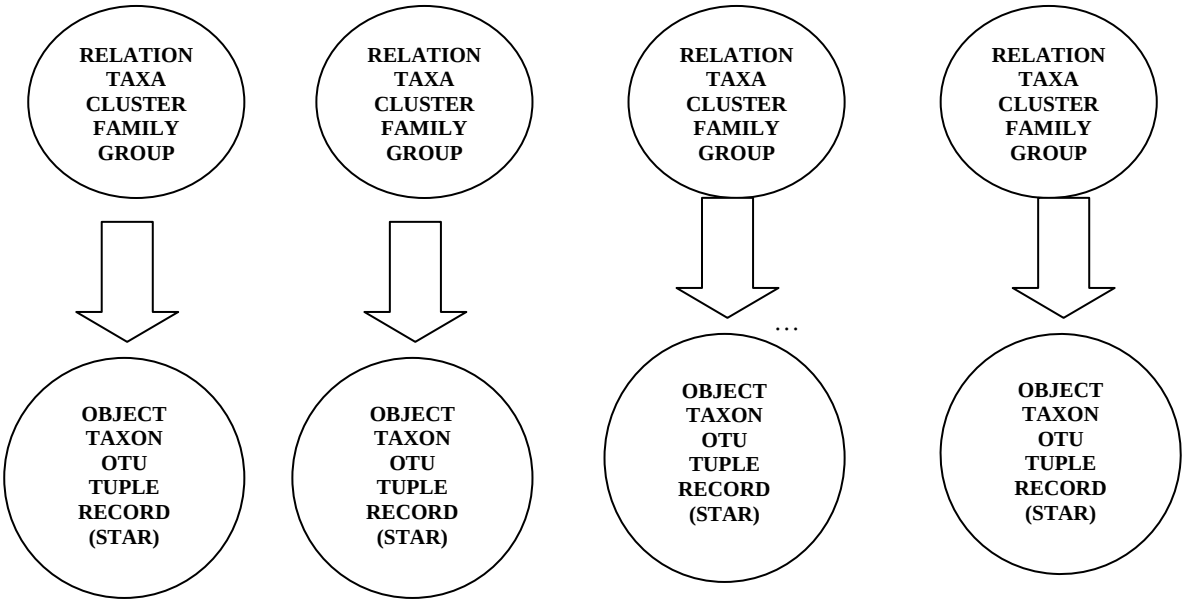
Authors	$\mu_\alpha \cos\delta$	μ_δ
	[mas/year]	[mas/year]
Robinson et al.	-4.04± 0,22	10.95 ± 0,20
Baumgardt et al	- 4.08 ± 0,27	10.98 ± 0,24
Dias et al	- 3.20 ± 0,25	10.10 ± 0,25
Orellana & Perichinsky	- 2.49 ± 0,27	9.74 ± 0,22

ANNEXED ANALOGIES

Analogy among Model of Entity Relationship (E-R), Databases and Taxonomy.
"More character-attributes are possessed best identified and qualified will be the objects"

Model of Entity Relationship	≡	Databases	≡	Taxonomy
↓	→	↓	→	↓
Entity Representation of Objects Through attributes	≡	Entity Representation of Objects Through attributes (Virtual)	≡	Representation of Objects Operational Taxonomic Units (OTU) Through characters
↓		↓		↓
Set of attributes $E=E\{a_1,a_2,...,a_n\}$	≡	Set of attributes $E=E\{a_1,a_2,...,a_n\}$	≡	Set of character s $O=O\{c_1,c_2,...,c_n\}$
↓		↓		↓
Domain Group of Values of an attribute $Da_i=Da_i\{v_1,v_2,...,v_n\}$	≡	Domain Group of Values of an attribute $Da_i=Da_i\{v_1,v_2,...,v_n\}$	≡	Domain Group of Values of a character $Dc_i=Dc_i\{v_1,v_2,...,v_n\}$ Value ≡ State of a character
↓		↓		↓
Scheme Set of attributes of an entity that identify a set of objects of the total of the field	≡	Scheme Set of attributes of an entity that identify a set of objects of the total of the system	≡	Cluster – Taxa – Family. Set of characters-attributes of Objects, OTU's, that they qualify them and they identify in the taxonomic space.
↓		↓		↓
Record Set of values of a proper domain of attributes that characterize to one object of a set in particular of a file.	≡	Tuple Set of values of a proper domain of attributes of an entity that characterize to one virtual object of a set in particular of a relation.	≡	States Set of values of a proper domain of an OTU, Object (e.g. star) that characterize it in a set, cluster or group in particular, according to the state-value of the character.
↓		↓		↓
File (Transparent) Set of records whose cardinal is equal to the total of objects qualified by the scheme.	≡	Relation (Virtual – Transparent) Set of tuples whose cardinal it is equal to the total of objects qualified by the scheme.	≡	Clustering Method of structural analysis that allows with the state-values to group Objects (e.g. Stars), in clusters that according to the cardinal of the set, the space contracts or it dilates to maintain a constant density of objects.
	→		→	





4. CONCLUSIONS

A new method has been presented, "the Spectral Method", for identify members of a cluster using the positions and the proper motions of the stars of the region, based on the Computational Taxonomy as evolution of the Numerical Taxonomy.

The stars characteristics spectra are obtained and are clearly shown the members stars of the cluster. It has been applied to the cluster NGC2516 satisfactorily.

Field stars will be processed in the next works, with the supposition that not alone those stars belong to an open cluster, but can have more than a group in movement, experimentally are observed two groups, while the spectral method shows to have a separator power that the other methods don't have, being been able to determine three groups in movement, in that field.

REFERENCES

[1] *Cabrera-Caño y Alfaro,1990,AJ,63, 387*
[2] *Dias et al. 2006, A&A, 446, 949*
[3] *Galadí-Henríquez 1998, AJ, 63,387*
[4] *Orellana R. B. y Perichinsky G. 2008. The Numerical Taxonomy, a tool to identify "Moving Groups". Application to Open Clusters. 51 Annual meeting of the Association Argentina of Astronomy. Astronomical Observatory R. Félix Aguilar. 22-27.*
[5] *Perichinsky, G. et Al. 2000. "Spectra of Taxonomic Evidence in Databases". Proceedings of the XVIII International Conference on Applied Informatics. Innsbruck. Austria.*
[6] *Perichinsky, G. et Al. 2002. "Spectra of Taxonomic Evidence in Databases". Proceedings of International Conference on Computer Science, Software Engineering, Information Technology, e-Business, and Applications (CESITeA´02). Foz do Iguazú. Brazil.*
[7] *Perichinsky, G. et Al. 2003. "Taxonomic Evidence Applying Algorithms of Intelligent Data Mining.". Proceedings of International Conference on Computer Science, Software Engineering, Information Technology, e-Business, and Applications (CESITeA´03). Rio de Janeiro. Brazil.*
[8] *Perichinsky, G. et Al. 2005. "Taxonomic Evidence Applying Intelligent Information Algorithm and the Principle of Maximum Entropy". Electronic magazine of Systems of Information (RESI). Edition 6 - Year IV - Volume IV - Number 2. Department of Computer Science and Statistic. Federal University of Santa Catarina. Brazil.*
[9] *Sanders,WL, 1971, A&A, 14,226*
[10] *Vasilevskis, S et al. 1958, AJ, 63, 387*
[11] *Zacharias et al. 2004, AJ. 12 7, 3043*