The use of Weighted Metric SOM Algorithm as a Visualization Tool for Demographic Studies

Elio Villaseñor, Humberto Carrillo, Nieves Martínez de la Escalera, and Valeria Millán

UNAM, Mexico City, Mexico

Abstract. Unsupervised neural networks provide a useful resource for exploratory data analysis. Here, we present an application of the SOM with weighted metric as an automatic tool to explore a large base of digital data of information about the student population of the National Autonomous University of Mexico,in order to look for gender differences impress in the academic performance. Our study proves the usefulness of this technique to visually analyze the performance and academic paths of several courts of students.

1 Introduction

Finding interesting structures and novel relations hidden in vast multidimensional data sets, be they textual documents, experimental data, or statistic information, is difficult and time-consuming. For these data mining tasks, information visualization techniques are valuable to display and analyze the discovered knowledge and therefore has become a topic of significant development and research.

The use of neural networks have proved to be useful in the data mining and knowledge discovery processes [1]. They are particularly valuable for the automatic generation of knowledge maps, that compactly convey information and are easy to analyze. By means of a "self-organizing" procedure, unsupervised neural networks based on mathematical algorithms, are capable of automatically explore large data bases. These neural networks are effective, not only to discover the structure of the data set, but also to reveal these patterns in a well organized knowledge map. This process involves two fundamental tasks: (i) the classification and cluster analysis; (ii) the geometric projection from the multidimensional space of data towards a two dimensional map.

The Self-Organizing Maps (SOM) algorithms [2] constitute a family of neural networks models that are widely used for exploratory data analysis and classification and clustering tasks. These algorithms due their popularity to the following facts: (i) by the use of reference vectors they allow an easy coding of multidimensional data that can be projected into a plane map by means of

© G. Sidorov (Ed.)
Advances in Artificial Intelligence: Algorithms and Applications
Research in Computing Science 40, 2008, pp. 13-25

a topology preserving transformation of the multidimensional space; (ii) supervised training is not required and (iii) the high performance of the algorithms allow the treatment of large databases.

During the unsupervised training of the SOM algorithm the neural network recognizes the structure of the data set. The following face of the process involves a visualization technique to generate a knowledge map over the two dimensional array of neurons.

The use of a weighted metric play an important role to obtain a visually adequate representation of the information and knowledge contained in the map. In this paper we discuss and illustrate, in the context of demographic study, the usefulness of a *variable scaling technique.* In this application the use of a weighted metric introduces a competition criteria that incorporates a hierarchical order of the variables that constitute the multidimensional data space.

In this work we present an investigation using a SOM algorithm with the purpose of finding gender differences in the academic performance of students in the Universidad Nacional Autonoma de México (UNAM).

2 A SOM Based Visualization Technique

A Basic SOM is a training neural network model that consider a two dimensional regular processing grid of neurons $\mathcal{N}$, with a set of reference vectors (synaptic weights) $W = \{w_\eta\}_{\eta \in \mathcal{N}}$ such that $W \subseteq \mathbf{X}$. Where $(\mathbf{X}, d)$ is a n-dimensional metric space. For each $\eta \in \mathcal{N}$, its reference vector has the form $\omega_\eta = (\zeta_1^\eta, \zeta_2^\eta, ..., \zeta_n^\eta,)$. If we consider a data set $X \subseteq \mathbf{X}$, at the end of the SOM training process, it is defined a projection mapping of the form:

$$\varphi : X \rightarrow \mathcal{N}, \tag{1}$$

given by the condition

$$d(x, \omega_{\varphi(x)}) = \min_{\eta \in \mathcal{N}} \{d(x, \omega_\eta)\}.$$

The main property of 1 is named "topology preserving", i.e. if $x, y \in X$ are close in $\mathbf{X}$ then $\varphi(x)$ and $\varphi(y)$ are close in $\mathcal{N}$. *Topology preserving maps, were originally created as a visualization tool; enabling the representation of high-dimensional data sets onto two-dimensional maps and facilitating the human expert the interpretation of data* [3].

A SOM-based data visualization method consist in a coloring of the cells in the two dimensional SOM's grid $\mathcal{N}$. One example is the U-matrix. For each $\eta \in \mathcal{N}$ consider U_η a neighborhood of η and the average distance u_η between the reference vector ω_η and the reference vectors of the neurons in U_η:

$$u_\eta = \frac{1}{\#U_\eta} \sum_{\nu \in U_\eta} |\omega_\eta - \omega_\nu| . \tag{2}$$

Where $\#U_\eta$ is the cardinality of the neighborhood. In the U-matrix method, u_η is used as a measure of accumulation of similar data. Assuming the topology preserving property, the information given by the projections of the $\{u_\eta\}_{\eta \in \mathcal{N}}$ over the grid $\mathcal{N}$.,. provides insight about similarity relations present in data.. For the visualization of this projection is used a bijection ϑ among $[\min\{u_\eta\}_{n\in\mathcal{N}}, \max\{u_\eta\}_{n\in\mathcal{N}}]$ and a monochromatic (e.g. gray-scale) color bar. For each $\eta \in \mathcal{N}$, its color is given by $\vartheta(u_\eta)$.

Although, the information that gives the U-matrix concerns only to the clusters structure present in the data but does not give information about the correlations between data variables. By studying this correlations it is possible to find out causality relations among data components. One technique that is useful to find correlations between variables is called Component Planes.

For each component, consider its values $\{\zeta_k^\eta\}_{\eta\in\mathcal{N}}$ on the the grid, the $\zeta_k -$ *Plane* is a coloring given for a bijection ϑ_k between the interval $[\min_{\eta\in\mathcal{N}}\{\zeta_k^\eta\}, \max_{\eta\in\mathcal{N}}\{\zeta_k^\eta\}]$ and a chromatic color bar. Therefore each variable ζ_k of the data set is projected using a coloring $\zeta_k -$ *Plane*. The comparison of two distinct component planes is useful for finding correlations of corresponding variables. *Correlations between component pairs are reveled as similar patterns in identical positions of the component planes. Pattern matching is something that the human eye is very good at* ...[4]. In the section of computational results these kind of analysis are presented.

3 Weighted Metric in the Multidimensional Space

The quantitative and qualitative information contained in the input data set has to be mathematically modelled in order to automatically explore the database. For this, each individual data is represented by a vector in an abstract multidimensional space. To analyze the structure of the *row data set* the use of a metric is in order. Typically the standard Euclidean (homogeneous) metric is used, however in certain applications it is convenient to consider an alternative metric in the row data set. The use of weighted metrics is a convenient resource to incorporate an element of competition during the training process, among the variables that represent the data in the space $\mathbb{R}^n$. Accordingly, the associated weights, pondering the relative relevance of the individual variables, establish a hierarchy in the multidimensional space.

A weighted metric is a function $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$. For each $x, y \in \mathbb{R}^n$ with components $x = (x_1, ..., x_n)$ and $y = (y_1, ..., y_n)$ the weighted distance between x and y is given by

$$d(x,y) = \sqrt[2]{\sum_{k=1}^{n}(w_i(x_i - y_i))^2},$$

where $w = (w_1, ..., w_n)$ is the vector of weights associated to each dimension.

The use of a weighted metric affects the induced projection, φ, from the row data set, $X \subset \mathbb{R}^n$, to the two dimensional neural grid, $\mathcal{N}$. The way in which the global ordering of the points in X projects over the $\mathcal{N}$ grid is determined principally by those components with heavier weights meanwhile the more local relations of similarity of the data are ordered in $\mathcal{N}$ by those components with lighter weights. Thus the training might differentiate first regions that correspond to the variables with the heavier weights and a further differentiation of these regions will subdivide them in subregions determined during the training by the variables with lighter weights.

In the following section, we illustrate the application of the SOM algorithm with a weighted metric in the frame of a demographic study of five student cohorts (39,893 students) of the UNAM. In this study each student is represented by a 28-dimensional vector, according to the order of magnitude of the associated weights to each of the vector components, are divided in three classes. One of the vector components is the variable sex to which, due to the purposes of our investigation, we have assigned the largest weight (2). The effect of this weighted metric in the projection map is the identification of two clearly separated gender regions (figure 2(a)). We point out how, inside these two regions, the neural net finds difference patterns, for each gender, associated to the 20 variables that measure the career-progress percentage of the students; a weight of 0.1 was applied to these variables. Using these twenty components we calculate a new discrete variable of the data vector that constitutes an indicator of each student's final status on finishing her or his studies: *Normative Graduation, Deadline Graduation, Terminal Graduation and Dropout*; a weight of 0 was applied to this variable. The next component correspond to the area of knowledge in which the career selected by the students is classified: *Physics, Mathematics and Engineering, Biological and Health Sciences, Social Sciences and Arts and Humanities*; this variable was weighted by 0.01. Finally the five following components contain the information on the quantified categorical variables, which, in our investigation, constitute the presumed achievement factors: marital status, children, job, the mother's academic background, whether the student's family owns a car or not, etc. We give this variables a weight of 0.01.

4 Practical Application: A gender study of student population of UNAM

As shown in figure 1, during the1980-2005 period the UNAM student population became stable with approximately 140,000 students. However, during the last 25 years there has been an important change in the sex ratio. In 1980, the male population was twice as large as the female population. From then on, the female population has increased notoriously, while male population has decreased at the same rate. Because of this sustained trend, in 1994 the two populations became equal with values around 70,000 until 1999. During this time a strike that lasted almost a year paralyzed the Institution and made enrollment of a whole generation impossible, which resulted in a momentary reduction of the population. After this numerical fall, both populations have recovered, reaching figures similar to those before the strike, where an interesting phenomenon can be observed: the recovering rate of female population notably overcomes that of the male. From the year 2000, this sex ratio has kept up the disparity, and

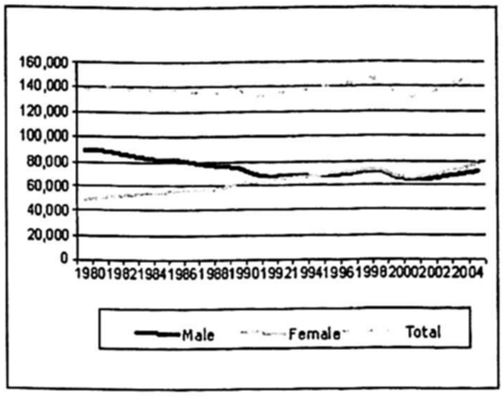

Figure 1. Dinamics of students population at UNAM.

in 2005, male population reached around 73,000, which was overcome by the female population of 80,000. If we calculate the change of these two population during the 1980-2005 period, there is a decrease in male population of over 25% and a growth of female population in more than 40%.

This behavior, [5] show: "Different gender social values and cultural norms based on masculine and feminine identity and specificity, as well as inequity conditions between the sexes are matters built and socially reproduced in different ways and in a dynamic manner". This phenomenon of substituting the male space by the feminine one should be studied because it is not the result of any special institutional effort to support this sector.

In a recent comparative study [6], about college education and gender in Latin America and the Caribbean, 'feminine' and 'masculine' careers are identified and typically associated to social roles and gender. These roles affect individual decisions and may create cultural barriers preventing an equal enrollment in higher levels as well as in the job market. As stated in a [7] study on Education in the XXI Century: "teaching and education are a source of social segregation according to gender, to the extent that the selection of professional and career paths are usually made before entering the job market".

In some articles, the UNAM case has been given particular importance. In the study [8] coordinated by the Programa Universitario de Estudios de Género (PUEG) entitled "The Presence of Men and Women in the UNAM: an X ray" gives a panoramic view of differences between women and men in three sectors of the UNAM: students, academicians, and administrative staff. Among the first results for the student population, the existence of careers that may be classified as typically masculine or feminine was validated. The evolution of gender predominance and detected variations tending to feminization is also analyzed. Besides, a more favorable positioning of women regarding the speed of curricular progress, higher averages and higher subject approval is noted.

In another gender study, social background and performance in the UNAM, [9] analyzes an enrollment cohort in 1997 and shows a consistent correlation between school performance and social factors. It can be concluded from the study that those differences noted regarding performance in individuals cannot be understood only as a product of innate skills but also as a product of other advantages and disadvantages which have an accumulative effect, among which the gender factor plays a complex role. In this study the importance of 'double shift' is also analyzed, the combination of study and work (generally males with salaries and women doing unpaid housework), and even in this scenario a better female academic performance and profit has being shown.

The database compiled for the analysis was obtained from two complementary sources: the academic record and questionnaires answered by candidates to the institution (Encuesta de Aspirantes de ingreso a licenciatura por concurso de selecciốn y pase reglamentado). We discarded all those students who did not answer the questionnaire or did not answer one of the questions regarding our study. The consolidated sample has a total of 39,893 students which represents 59.2% of the population from the five studied cohorts. We analyzed the progress of these five cohorts for 20 semesters (most of the careers are planned for 10 semesters). After 20 semesters very few students finish their careers. In this experiment we get the same conclusions of the mentioned works by visually analyze the component maps of a SOM training with this database.

4.1 Computational Results

The following maps are conformed by a flat square hexagonal grid of 3,969 neurons, following the recommendation of Kohonen for the number of neurons to use one order of magnitude less than the cardinality of X [2]. We used a SOM algorithm implemented in ViBlioSOM system that is the Batch-Map variation with a Gaussian neighborhood function with a lineal radius decreasing. This is a software system that is in development by the "Laboratorio de Dinámica No-Lineal" of the Sciences School at UNAM. This system implements a SOM algorithm for the visualization of bibliometric information. In order to execute exploratory data analysis, taxonomy studies or Clustering, as well as generate cartographies through projecting data from the multidimensional space in a plane. This system is useful for informetric analysis who produce automatic knowledge maps representing the structure and information included in the data basis. The system also produces quantitative results and offers a variety of graphic scenarios for their representation. In this work, we only use the SOM-Based visualization capabilities of the system. In fact. the results of this applications are already presented in [10].

For example, Sex-Plane (figure 2(a)) we established a chromatic spectrum that goes from blue to red (as in rainbows). In this case, the red color represents women, and blue represents men. Thus, in this map we can identify two distinct areas (clusters of cells) in which men and women have been distributed. Both, the red and blue regions of this map have been clearly differentiated by a diagonal division. We call these regions the feminine zone and the masculine zone. They both span areas that are almost equal in size within the neuronal grid which has been created by the map, although the feminine zone is slightly larger than the masculine zone. This suggests that the feminine population is slightly larger than the masculine one, and this reflects a numeric reality: there are 22,127 women, whereas there are only 17,765 men.

In the Performance-plane (figure 2(b)), this process has created four different zones. These four classes are determined according to the time in which the students covered an equal or superior percentage of 90% of the semesters which are required in order to finish their Bachelor's Degree studies. The term Normative Graduation corresponds to both male and female students who accredited at least 90% of their studies during the first 10 semesters. The term Deadline Graduation corresponds to the period of time in which this 90% percentage of credits has been reached between the 10th and 15th semesters. The term Terminal Graduation corresponds to students who have finished their studies between the 15th and 20th semesters. And the term Dropout corresponds to students who do not manage to finish their studies, and thus do not graduate, even after being registered at UNAM for 20 semesters. We give a weight of 0 to this component because it has been considered in the previous twenty components. The Normative Graduation zone (which is red) corresponds to both male and female students who have "finished or completed" their career studies (having covered 90% of their credits) during the allotted time. In similar fashion, the Deadline Graduation zone has been highlighted in green, the Terminal Graduation zone has been highlighted in yellow, and the Dropout zone has been highlighted in blue. Any one can clearly see that this map is quite symmetrical in relation to the diagonal line which descends from left to right. Nevertheless, the Normative Graduation zone (highlighted in red) covers a greater extension of space in the feminine region. Besides, one can observe that the Deadline Graduation zone (yellow), the Terminal Graduation zone (green), and the Dropout zone (blue) cover areas of similar size in the masculine area as well as in the feminine one.

In Job-plane (figure 3(a)), we can perceive an outstanding asymmetry: the yellow and red tonalities predominate (these correspond to students who were working at the time when they answered the questionnaire) in the masculine zone. The greater part of these cases appear distributed in the Dropout zone. The color red does not appear in the feminine zone, but we can observe green regions. As the color of the cells reflects the average value of the component which is being analyzed, one must interpret that in the green cells (which stand between the red and the blue zones) we have grouped both the women who hold jobs as well as those who do not. It must be noted that these tonalities predominate in the Dropout zone.

In the Children-plane (figure 3(b)), blue regions indicate the place within the map held by students who already had children when they begun their studies. The greenish-yellowish tonalities are distributed homogeneously across the map, but we can observe that these colors predominate in the feminine Dropout zone, with some exceptional cases in the feminine Normative Graduation zone. The Marriage Status-plane (figure 3(c)) shows a very similar distribution pattern that is quite similar to the previous component plane.

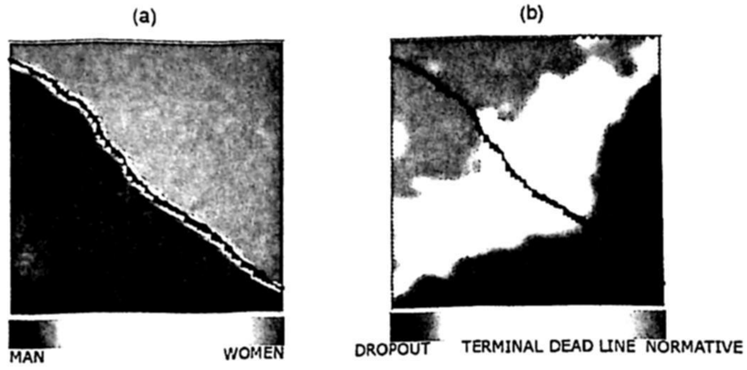

Figure 2. (a). Sex-plane (b). Performance-plane

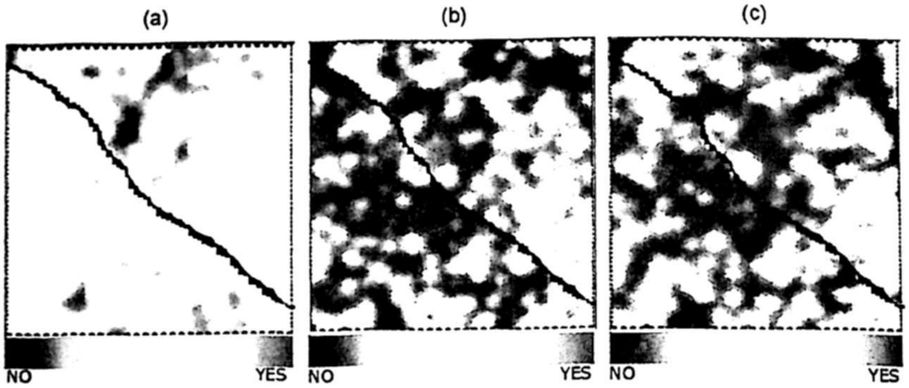

Figure 3. (a). Job-plane (b). Marriage Status-plane (c). Childen-plane

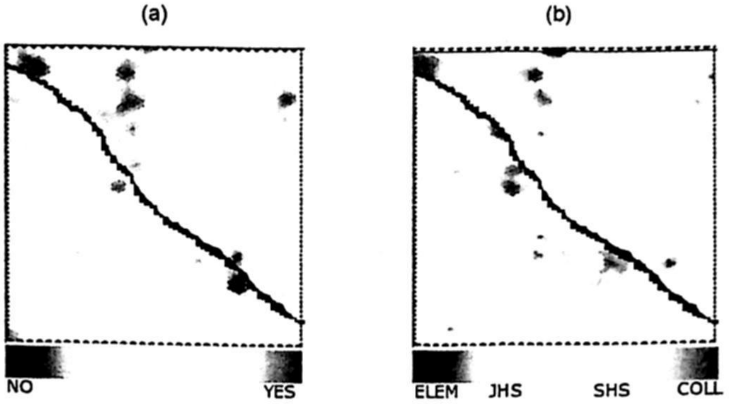

Figure 4. (a). Automovil-plane. (b). Mother. Instruction-plane

In the stratification of Mother Instruction-plane (figure 4(b)): The Mother's Academic Background, we can see a greater concentration of yellowish-reddish blots (students whose mothers have a higher academic level) can be seen in the feminine Normative Graduation zone. This same type of organization is manifested in Automovil-plane (figure 4(a)): Car, although they are more dispersed throughout the whole map, and there are some yellowish-reddish blots in the masculine Normative Graduation zone.

The distribution of colors that correspond to the Areas of Knowledge in which students chose their career studies (Area-plane figure 5) is less disperse if compared to previous maps. However for the correct interpretation for this component plane, it has to be considered that there are neurons that do not have the exact color that is assigned to an specific area. In this neurons, it is sure that there are students whose career do not belong to the same area of knowledge. Taken into account this considerations, it is noteworthy that the region that corresponds to Physical and Mathematical Sciences as well as Engineering (blue) is made up of two unrelated conglomerates. This highlights the existence of two different classes of typical trajectories within this subpopulation group, and which should be conceptualized. One of them inhabits the Normative Graduation zone with a notoriously dominant component in the masculine zone. The second blue conglomerate is slenderer than the previous one, and is enclosed, almost completely, in the masculine Dropout zone. Much in the same manner, in the feminine Normative Graduation zone we can see a red conglomerate which corresponds to female students belonging to Arts, Philosophy and Literature academic careers. In this same zone we can also observe the dominant presence of yellow, which corresponds to the Social Science area.

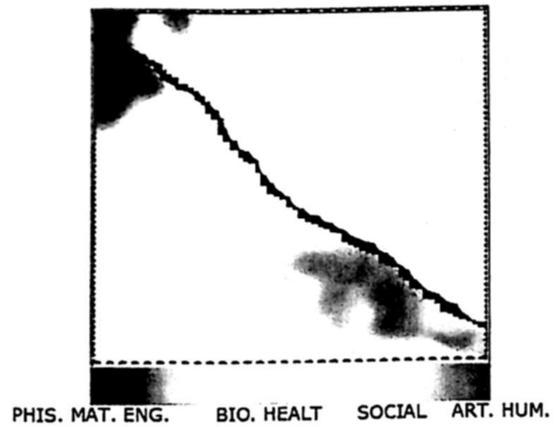

Figure 5. Area-plane

4.2 Discussion and Interpretation

The computational results we have obtained by means of the ViBlioSOM system are consistent with available information as well as with the results of previous investigations. This validates the procedure we have followed.

This experimental investigation has served us well in verifying the different work hypothesis, thus confirming the prevalence of noteworthy gender differences in academic achievement. These differences are clearly visible within the different time modes in which students manage to finish their studies, or not, as well as in the different career studies that students in the UNAM choose to study:

a. If the categorical variables we have proposed were not legitimate achievement factors, the regions of different colors which appear on the component maps would not fit regularly, as we have already seen, within the different Graduation and Dropout zones.

b. On the other hand, given the almost perfect symmetry that can be seen in the zones pertaining to gender classification, if any of the proposed achievement factors did not have a differentiated effect on gender issues or if students' preferences when choosing the area of knowledge in which they wish to study would not be marked by the difference in gender, one could expect that the coloring in the corresponding component maps would be distributed in a symmetrical manner in reference to the diagonal line that divides the feminine zone from the masculine one.

Gender differences were expressed in the creation of the maps in the following manner:

- The map of the Gender component clearly shows the feminization of the enrollment pattern within the UNAM. This could mean a progress towards gender equality and it also constitutes a potential factor that can contribute to the balance of opportunities for both female and male graduates when looking for a job and as an equal means to achieve social standing.

- From the analysis of Areas of Knowledge map, we can confirm the presence of academic spaces which are considered typically masculine, such as Physics, Mathematics, Engineering, as well as others which are considered typically feminine, such as Arts, Philosophy and Literature. One possible explanation for the dominance of women in the Normative Graduation zone is the "double shift" characteristic, which in the case of men implies holding a job besides studying. This conjecture is supported by the information obtained from the Job map: when students begin working at an early age, this can have a negative effect on the time they take to graduate from the university, or it even forces them to finally drop out. Thus we can conclude that for men holding a job at an early age, compelled to earn money in order to assume their traditional gender role, their role as breadwinners competes with their role as students, making them lag behind in their studies.

- Entering the university either alredy married or with children is not a common condition within the freshman population (freshman average age is 19.97 years, std. deviation 3:02; men 20:30 years, std. deviation 3.09; women 19.72 years, std. deviation 2.93). As is to be expected, both the Marital Status and the Children maps coincide quite a bit. These maps reveal a differentiated tendency by gender for those students whom, upon enrolling were already married or already had children: in the masculine zone, the scarce presence of students married or with children does not have a notorious influence on graduation timing, and it is distributed evenly in the different graduation zones. On the other hand, married women or women who have children are concentrated preferentially in the Dropout zones. This is reminiscent of the traditional roles played by both genders, according to which women must take care of their homes and their families, while all professional aspirations are deposited in men.

- An additional confirmation of the validation of the methodology we have used is offered by The Mother's Academic Background Map and the Car map. Both express the supposed relationship between the socioeconomic status of students and their academic achievements. It is commonly thought that a higher socioeconomic standing will necessarily mean a competitive advantage for academic performance. This is confirmed in maps which show a prevalence of yellow and red spots in the Normative Graduation zone.

- Furthermore, the exploratory analysis of data we have carried out by means of the ViBlioSOM system allows us to make the following predictions, which should be the object of future investigations:
  - Economic status carries less weight in graduating during the normative time than the mother's academic level.
  - The fact that a student's mother has a high academic level can certainly have an outstandingly positive effect on the academic trajectory of female students, but not so much in the case of male students.
  - The differentiated effect that a mother's academic achievements have on female students, together with the current process of substituting masculine spaces by female ones, clearly suggests that this feminization trend will be maintained in the long term. This conjecture must become the subject for future research.

5 Conclusions

In this work we present an experimental demographic application of the basic SOM model with weighted metric. The interpretation of the maps coincide with the results of other researches of gender differences at the UNAM and new hypothesis were formulated. With the puropose of getting meaningful maps it is useful to incorporate hierarchical orders in the variables via weighting of variables. Once the logic principle of map's interpretation is understood, the visual analysis is very easy to be done. So it is possible to get very valuable information by simply viewing the maps without considering statistics. We think that this kind of applications can be extended to other demographic studies and it can be a great support to many important decisions in social researches.

Acknowledgement *"Macroproyecto. Tecnologías para la Universidad de la Información y la Computación".*

References

1. Bigus J., "Data Mining with neural networks", McGraw Hill, USA, 1996. Buquet, Anaetal. Presencia de Mujeres y Hombres en la UNAM: una radiografía. Programa Universitario de Estudios de Género (PUEG)-UNAM, México (2006).
2. Coñeen T., "Self-Organizing Maps", 3ra Edición, Springer-Verlag, 2001.
3. Baruque B., Corchado E., Yin H., "ViSOM Ensambles for Visualization and Classification", F. Sadoval et al. (Eds.): IWANN 2007, LNCS 4507, Springer-Verlag 2007.
4. J. Vesanto, "SOM-based data visualization methods," Intelligent Data Analysis, vol. 3, April 1999.
5. Szasz, I. and Susana L. "Aportes teóricos y desafíos metodológicos de la perspectiva de género para el análisis de los fenómenos demográficos" in

Susana Lernery Alejandro I. Canales (eds.). Desafíos teórico-metodológicos en los estudios de población en el inicio del milenio. COLMEX, UDG, SOMEDE, México (2003), pp.177-209.

6. Papadópulos, Jorge and Radakovich, Rosario. "Estudio comparado de edu-cación superior y género en América Latina y el Caribe" Educación superior y género en América Latina y el Caribe, IESALC, Unión de Universidades de América Latina-UDUAL.(2003). Ch. 9, pp.118-128.
7. Internacional Labour Organization. La educación permanente en el siglo XXI: nuevas formas para el personal de educación, Geneve (1998).
8. Buquet, Anaetal. Presencia de Mujeres y Hombres en la UNAM: una ra-diografía. Programa Universitario de Estudios de Género (PUEG)-UNAM, México(2006).
9. Mingo, Araceli. ¿Quién mordió la manzana? Sexo, origen social y desem-peño en la universidad, Universidad Nacional Autónoma de México-Fondo de Cultura Económica (2006).
10. Millán V., Villaseñor E., Martínez de la Escalera N and Carrillo H, "Informetrical Visualization of Gender Differences in College Performance: an application of ViBlioSOM", sended to Resources for Feminist Research.