Data Mining application to decision-making processes in University Management

JUAN CARLOS GARCÍA[1]
ALICIA INÉS ZANFRILLO[2]

UNMdP - Universidad Nacional de Mar del Plata
FI - Facultad de Ingeniería
Juan B. Justo 4302 – Mar del Plata (7600) – Argentina

FCEyS - Facultad de Ciencias Económicas y Sociales
Funes 3250 – Mar del Plata (7600) - Argentina
[1] jgarcia@fi.mdp.edu.ar
[2] alicia@mdp.edu.ar

Abstract
This article aims at identifying and characterizing the different configurations of the relevant factors related to administrative procedures, in order to learn about their behavior, and so about decision-making processes. Delays in administrative circuit concretion, users dissatisfaction and lack of autonomy are among the problematics leading to the diagnosis of the situation, so as to detect the elements playing key parts in said scenario.
Data analyzed was collected from files and notes submitted by the middle management, i.e., pedagogical areas and departments at a Higher Education institution in Argentina, the School of Economic and Social Sciences, FCEyS, of the Universidad Nacional de Mar del Plata, UNMdP, throughout 2002 - 2005.
Knowledge acquired through data mining techniques in their character of strategic knowledge, turns out to be potentially useful in the diagnosis of administrative circuit problematics in university environments. Besides, it is particularly beneficial in institutional decision-making processes, promoting plan of action deigns for an efficacious administrative management.

Keywords: Decision-Making Process - Data Mining – Knowledge Discovery- Administrative Circuits - Higher Education Institutions.

1. Introduction

Entering of documents such as files and notes and their corresponding transfer to the offices making up the administrative circuit demands the attention of groups of people working in different organization sectors, through an application process organized according to administrative process techniques [9].

In Academic Units, decisions are taken based on the proceedings appearing on files or notes, in which each area making an opinion does so based on a different rationality. The Teaching Department, for instance, does so on the basis of the staff: Administrative rationality; the Academic Department safeguards the quality of the teaching-learning process: Academic–Technical rationality; and finally the Coordination Department based on budget appropriation: Economic rationality [6]. The Academic Council, the highest decision-making body takes the final decision on the basis of all these recommendations as well as on its political rationality, always in view of the fulfillment of the Academic Unit final objective.

The Coordination Department[1] is in charge of entering in the information system the procedures initiated and decisions taken on a centralized basis, i.e., notes, files, passes and resolutions entered by the sectors involved in the administrative circuit. In this way, information is readily available for the whole academic unit through client/server applications. Paper documents follow their due course through all the offices so requesting them.

The need to enquire about the causes resulting in administrative procedures slow down as well as about procedures complexity arises from several queries at the Coordination Department and from users reiterative complaints for circuit delays and lack of autonomy in notes and files location and follow-up. Data mining is applied to that end [4].

A solution proposal based on the knowledge stemming

1.The Coordination Department is in charge of receiving and forwarding circuit notes and files to the other sectored in the Academic Unit in accordance with the nature of the document. It also handles the administrative proceedings resulting from written opinions by the Academic Council, the highest decision-taking body in the School.

from databases will allow not only to take into consideration the administrative procedures own characteristics, but to use different approaching strategies in view of the wide array of chances available, rather that a set of standard actions [5].

2. Knowledge Extraction from Databases

Technological advances, low costs in data storage and basic processes informatization have enabled great volumes of data availability in organizations.

This data constitutes the basis for the acquisition of useful information in decision-making processes. These data transformation processes into information are usually within the scope of statistics. Currently, new techniques such as data mining are being implemented, which allow to reveal behavioral patterns or hidden and implicit data relationships.

Data mining (DM) proceeds in the inverse order if compared to scientific method. To begin with, DM formulates hypotheses and then designs the experiment to collect the data either confirming or refuting the hypothesis constructed. In DM, data is first compiled and then possible valuable associations or links are awaited in order to make decisions.

Data mining, through predictive and descriptive relationships finding and quantification from raw data, allows to transform data into knowledge of use in decision making. A classic definition of data mining is as follows: "... a stage within knowledge discovery process. It refers to algorithms application to obtain data patterns" [2].

Ever since 1989, the term knowledge discovery starts to be applied to databases, making reference to the knowledge search process in data bases, especially with regard to optimum methods of data mining [2].

Discovery is a type of unsupervised knowledge induction, which, on the one hand, implies the search for interesting regularities among input data; and on the other, the formulation of rules describing them [7].

A classic taxonomy divides data mining techniques into two types: supervised or predictive and unsupervised or descriptive. Supervised techniques predict the value of an attribute or of a data set from data of known attributes. A relationship between the attribute and the already existing attributes is induced. Unsupervised techniques discover data behavioral patterns making no use of historic data [2].

Among the predictive algorithms are those which can be classified, such as decision and regression trees. Among unsupervised algorithms, on the other hand, are association rules and clustering mechanisms.

3. Description of the Employed Methods

3. 1 Data Preparation and Variables Definition

The data used in this article was collected from a file and note follow-up system of the FCEyS, UNMdP, corresponding to the 2002 – 2005 time period. The research was carried out on a total of 601 cases.

The procedures dealing with unprogrammed decisions were excluded, such as new degree programs, for being, above all, complex circuits. Analysis units with wrong or incomplete data were also excluded from this study.

Along these lines, work started with the notes submitted by the middle management to the Academic Unit: knowledge areas and departments, selecting those documents treated by the Academic Council. Middle management, so described, makes up the categories of the variable Sector, coded from 1 to 12 on account of information confidentiality reasons.

Variables selection was supported by the following: the *sector* in which the administrative procedure arises, the *character* of the note submitted, the *total delay,* meaning the time required to fulfill a given procedure in the Academic Unit, from the moment it is started to its resolution by the Academic Council, the *internal delay,* meaning the time elapsed between procedure initiation in the Academic Unit and its referral to the Academic Council; and the *order* or *complexity of the internal circuit,* meaning the number of passes the document undergoes through the different sectors until it is finally resolved by the Academic Council.

The taxonomy herein proposed of the documents entered to be treated by the Academic Council is as follows: notification, designation and resolution depending on the required decision-making level. The notification level encompasses notes reporting and informing the Academic Council on different issues. The designation level concerns with staff designations; and the resolution level with master agreements, extension activities and postgraduate studies.

As far as application domains are concerned, both *total* and *internal delay* constitute time periods whose values range from 0 to 265, and from 0 to 158 days, respectively; whereas *order* values range from 0 to 11.

3.2 Cluster Analysis

Once the relevant variables and categories, either latent or manifested, have been defined for the analysis, administrative procedures start being classified, grouping them in clusters through cluster analysis, based upon the scores of the variables employed [3].

This multivariate technique enables to classify objects or individuals when neither the groups to be formed nor the number is known a priori. Group formation is carried out in such a way that, with regards to the distribution of the variables values under study, each cluster is as much homogeneous as possible; still being different among

themselves. Objects, in this particular case, are the administrative procedures started in the Academic Unit.

Given the quantitative and qualitative character of the variables under study, the two-step analysis was employed available in the SPSS Statistics Software. The two-step analysis employs log-likelihood as distance measure, which assumes that variables in the model cluster are independent; categorical variables have a multinomial distribution; and continuous variables a normal distribution. Said variables are standardized in the analysis [10].

To determine the best number of clusters, each clustering solution is compared by means of Schwarz's Bayesian criterion, BIC [1] [11]. Said criterion is calculated for each potential number of clusters, the smallest BIC values representing the best models.

To validate the solution obtained, a discriminating analysis was employed, in which the dependent variable is a new variable generated from the two-step analysis, which determines each case belonging to the cluster created; and the independent variables those used in clustering definition [8].

A series of variables and categories is herein proposed, some of them latent and other manifested, which intend to represent the most relevant shades; and of which there is availability of administrative procedure information.

4. Exploratory Analysis

4.1 Cluster Analysis

Regarding document selection, the ones chosen were those corresponding to middle management, as the sector where administrative procedures get started, with conclusion date; i.e., the analysis is carried out on the population.

A weak to moderate correlation was found between the quantitative variables employed in the analysis. Given the fact that the analogy identified in the variables is not lineal, making up a points cloud instead; and given the analysis robustness, it is possible to consider this association between the variables under study.

As it could be a priori assumed that internal delay would be contained in total delay, a strong correlation could be established between them. In fact, such relationship is explained by the variable contemplating the delay between the procedure reaching the Academic Council and its actual resolution, a variable not considered in cluster analysis, with a positive correlation of 0.816.

The two-step analysis reveals six clusters depicted in the sectors diagram in Table 1. They are identified with:
Cluster 1: 5.2%

Cluster 2: 24.0%
Cluster 3: 19.1%
Cluster 4: 20.1%
Cluster 5: 12.1% and
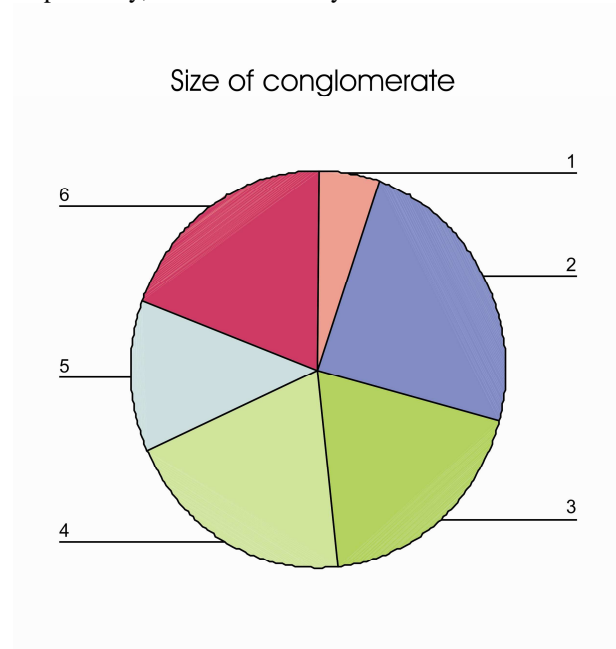Cluster 6: 19.5%,
respectively, of the cases analyzed.

Table 1. Two-step Analysis Cluster Configuration.
Cluster size

With a view to determining the qualitative variables categories distinguishable in the analysis, Tables 2 and 3 illustrate the diverse modalities of the categorical variables employed as well as their contribution to each cluster formation.

In the relevance analysis of each variable in cluster determination, qualitative are distinctly distinguished from quantitative variables and each of them is presented in order of decreasing importance. Once the variables exceed the critical line in a positive way, i.e., values above cluster mean; or negative way, i.e., values below cluster mean; they become meaningful in this analysis.

Qualitative and quantitative variables contribution in each conglomerate size can be appreciated in Tables 3, 4, 5, 6, 7 and 8. In accordance with the significance of each variable in the cluster formation, a name was given to each group on the basis of the main revealed profile.

		Conglomerate						
		1	2	3	4	5	6	Combined
1	Frequency	0	0	5	16	1	0	22
	Percentage	,0%	,0%	22,7%	72,7%	4,5%	,0%	100,0%
2	Frequency	2	12	10	0	0	0	24
	Percentage	8,3%	50,0%	41,7%	,0%	,0%	,0%	100,0%
3	Frequency	0	18	3	0	0	0	21
	Percentage	,0%	85,7%	14,3%	,0%	,0%	,0%	100,0%
4	Frequency	2	30	30	0	2	0	64
	Percentage	3,1%	46,9%	46,9%	,0%	3,1%	,0%	100,0%
5	Frequency	2	33	5	0	1	0	41
	Percentage	4,9%	80,5%	12,5%	,0%	2,4%	,0%	100,0%
6	Frequency	5	18	4	0	7	0	34
	Percentage	14,7%	52,9%	11,8%	,0%	20,6%	,0%	100,0%
7	Frequency	4	0	8	13	1	0	26
	Percentage	15,4%	,0%	30,8%	50,0%	3,8%	,0%	100,0%
8	Frequency	9	0	30	33	11	0	83
	Percentage	10,8%	,0%	36,1%	39,8%	13,3%	,0%	100,0%
9	Frequency	4	0	8	59	38	0	109
	Percentage	3,7%	,0%	7,3%	54,1%	34,9%	,0%	100,0%
10	Frequency	2	0	10	0	0	117	129
	Percentage	1,6%	,0%	7,8%	,0%	,0%	90,7%	100,0%
11	Frequency	1	19	2	0	11	0	33
	Percentage	3,0%	57,6%	6,1%	,0%	33,3%	,0%	100,0%
12	Frequency	0	14	0	0	1	0	15
	Percentage	,0%	93,3%	,0%	,0%	6,7%	,0%	100,0%

Table 2. Modalities adopted by the categorical variable Sector and its contribution to each cluster formation.

		Conglomerate						
		1	2	3	4	5	6	Combined
Notification	Frequency	8	0	115	0	0	0	123
	Percentage	6,5%	,0%	93,5%	,0%	,0%	,0%	100,0%
Designation	Frequency	13	144	0	121	0	66	344
	Percentage	3,8%	41,9%	,0%	35,2%	,0%	19,2%	100,0%
Resolution	Frequency	10	0	0	0	73	51	134
	Percentage	7,5%	,0%	,0%	,0%	54,5%	38,1%	100,0%

Table 3. Modalities adopted by the categorical variable Character and its contribution to each cluster formation.
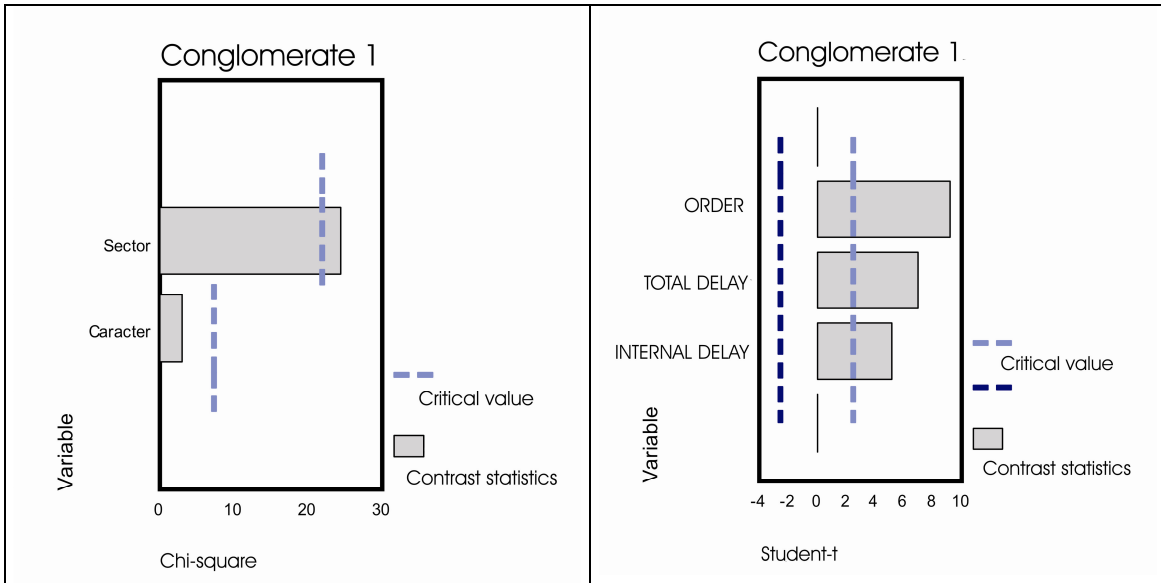
Table 4. Cluster 1: Qualitative and quantitative variables in cluster formation.

Complex and long-term sectorized procedures. They feature the highest total delay, internal delay and complexity values, with a minimum composition of the three types of notes, and, in a lesser extent, of most sectors. The most meaningful qualitative variable is the sector.
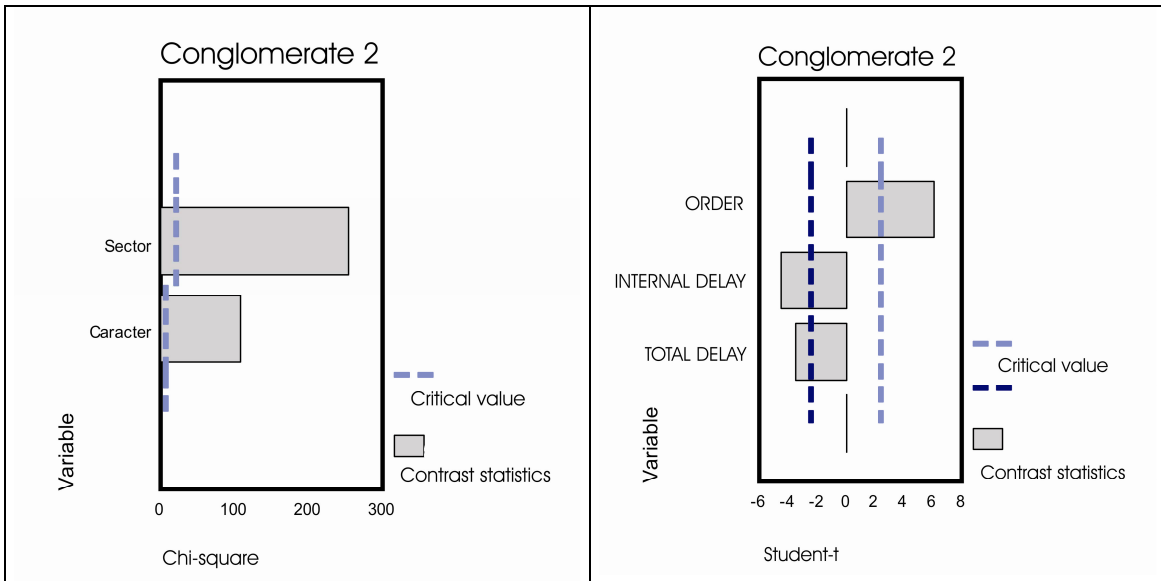

Table 4. Cluster 2: Qualitative and quantitative variables in cluster formation.

Sectorized complex designation procedures. They feature high complexity, low internal delay and low total delay. They include nearly half of the designation notes and sectors 2, 3, 4, 5, 6, 11 and 12, specially 3, 5 and 12. Both qualitative variables are meaningful, specially the sector.
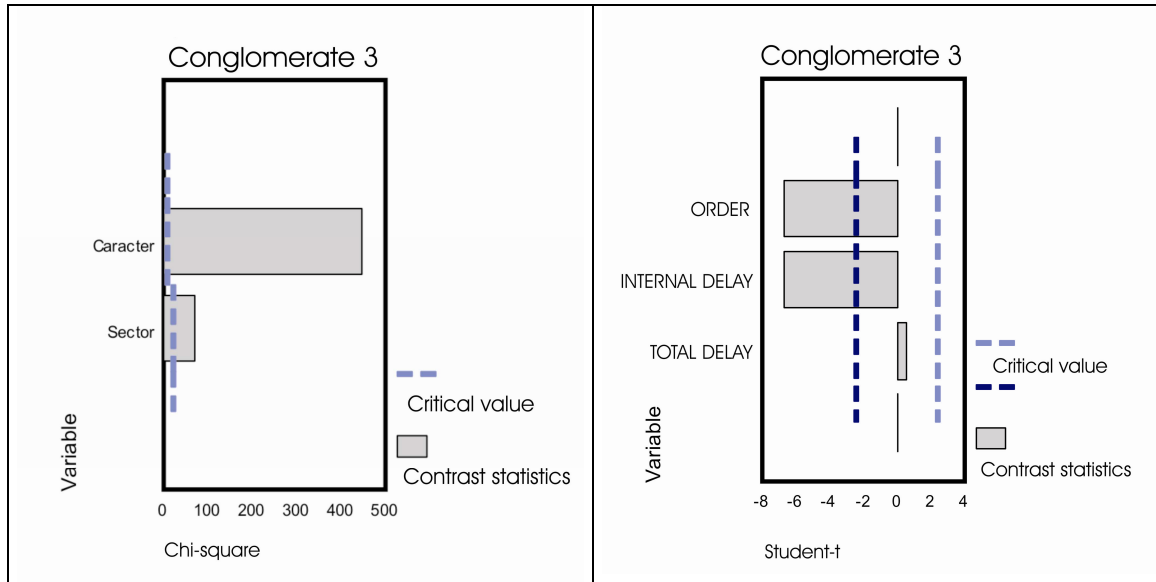
Table 5. Cluster 3: Qualitative and quantitative variables in cluster formation.

Quick and simple notification procedures in the internal circuit. They feature very low complexity and internal delay. They encompass most notification notes as well as most sectors, except for sector 12, especially 2, 4, 7 and 8. Both qualitative variables are meaningful, specially character.
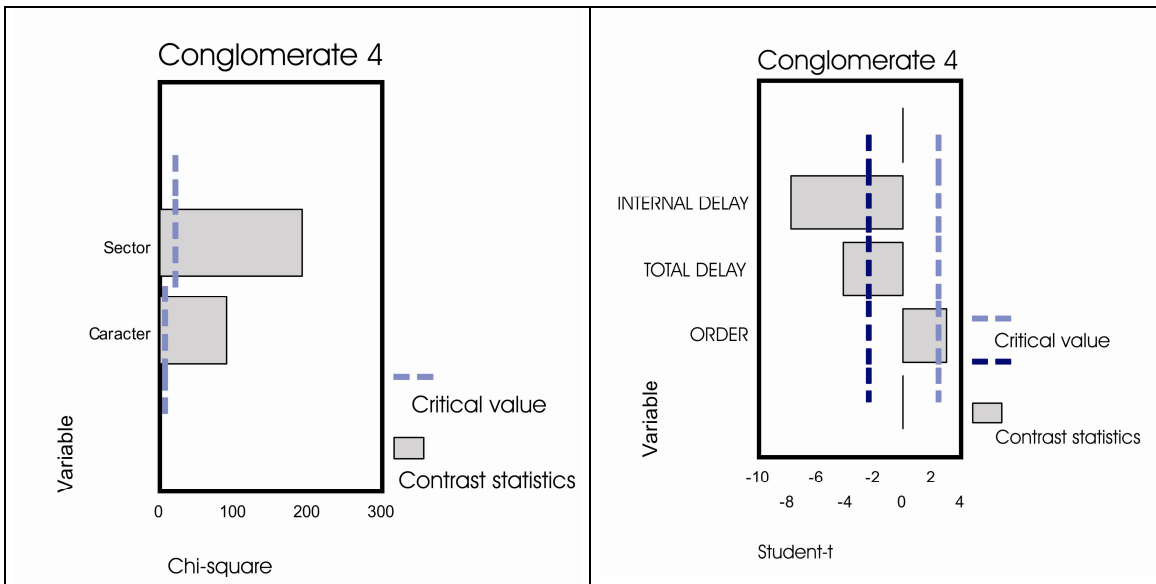

Table 6. Cluster 4: Qualitative and quantitative variables in cluster formation.

Sectorized quick designation procedures in the internal circuit. They feature very low internal delay, low total delay and moderate complexity. They comprise more than one third of designation notes and sectors 1, 7, 8 and 9, especially sector 1. Both qualitative variables are meaningful, specially sector.
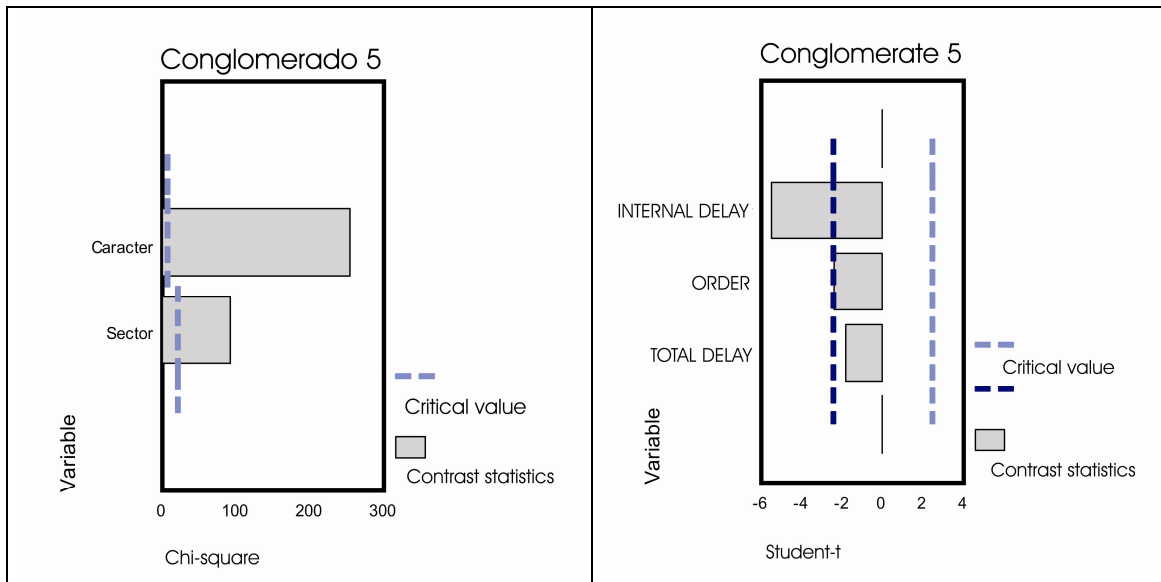
Table 7. Cluster 5: Qualitative and quantitative variables in cluster formation.

Sectorized quick resolution procedures in the internal circuit. They feature very low internal delay and moderate complexity. They comprise more than half of the resolution notes and are localized in sectors 1, 4, 5, 6, 7, 8, 9, 11 and 12, mainly in sector 9 and 11. Both qualitative variables are meaningful, specially character.
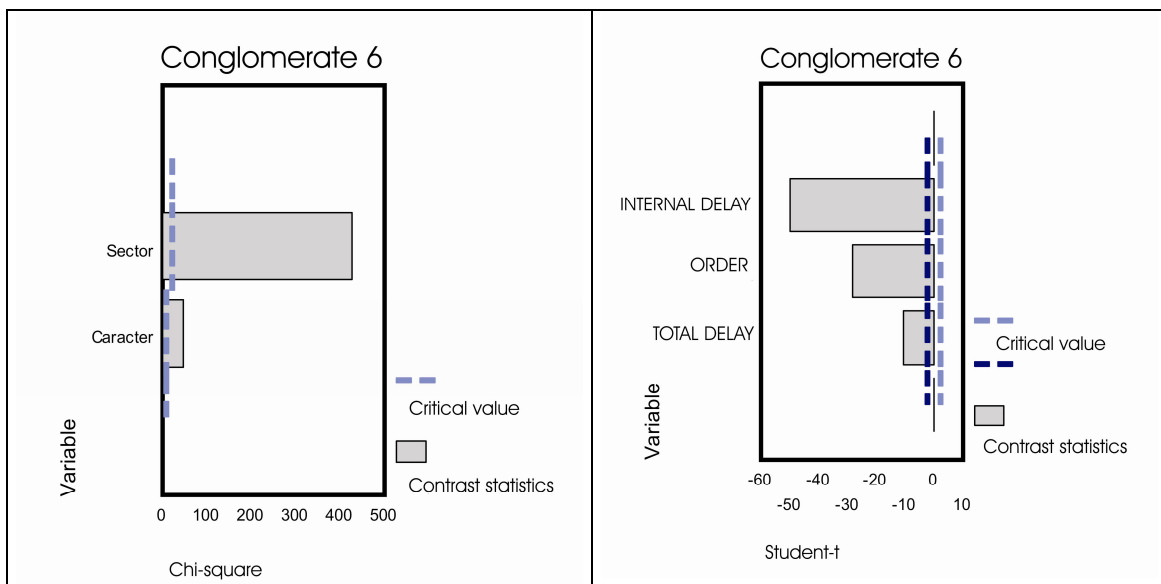

Table 8. Cluster 6: Qualitative and quantitative variables in cluster formation.

Sectorized rapid and simple procedures. They feature very low internal delay, low complexity and low total delay. They comprise approximately one fifth of the designation notes, and more than one third of the resolution notes. They almost exclusively include the totality of sector 10 notes. The most meaningful qualitative variable is the sector.

A total of 83% of the cases originally grouped were correctly classified. In order to validate the model obtained, a confusion matrix of the discriminating analysis (Table 9) was generated between the forecast clusters and the new discriminating analysis variable. This comparison yields an 83% of correct assignations, rendering an acceptable analysis goodness.

Clusters	Forecast Group						Total
Original	1	2	3	4	5	6	
1	24	2	2	0	3	0	31
2	1	104	0	31	0	8	144
3	0	0	115	0	0	0	115
4	0	34	0	84	0	3	121
5	0	0	0	0	73	0	73
6	0	0	0	7	11	99	117

Table 9. Confusion Matrix of the discriminating analysis.

Conclusions

This research carried out based on data mining for knowledge discovery has allowed the construction of an administrative procedure typology, taking into consideration its composition through relevant variables and categories. The relationship resulting from the data, despite yielding a diversity of associations, allows to adopt measures in agreement with the detected problematic and situations.

Below are the main conclusions drawn:

- Greater delay and complexity procedures arise especially from specific sectors.
- Designation notes, despite their great complexity, are promptly treated internally as well as in the Academic Council. Given the nature of the note, most of these procedures are delivered by hand. The complexity of these types of notes lies in the administrative procedure sector of origin.
- Notification notes are mainly characterized by the simplicity of the procedures required as well as by the rapidity in going along the circuit.
- Designation notes are promptly solved in some sectors; whereas in others, delays as well as circuit complexity are greater, proportional to the type of designation involved.
- Most resolution notes are quickly handled in the internal circuit; and most contributive sectors are those vested with greater decision-taking power in the organization.
- Sector 10, features very low values for most designation and resolution procedures. Being two clearly distinctive procedures, they are standardized, so explaining efficiency.

Possible solutions encompass specific training in those sectors in which longer and more complex procedures are completed, in formalizing knowledge about the sectors carrying out an efficacious management of the procedures and in the definition of a set of standards according to the character of the note to be submitted.

With the experience collected thanks to knowledge extraction, data mining techniques can be considered highly useful for the activities developed in Higher Education institutions, given the likelihood of applying said technologies. Knowledge resulting from administrative circuit characteristics will enable to count on a valuable tool in the decision-making process regarding area management.

To sum up, the significance of identifying lines of behavior differentiating clusters is noteworthy, so indicating the diversity of administrative procedures characteristics. Said diversity should be taken into consideration in the development and application of the organization policies and strategies, in order to attenuate present differences and improve the development of weaker sectors.

This paper intends to provide an answer to the queries arising from the characteristics and behavior of administrative procedures. A future line of study could concentrate on predictive nature questions made regarding already started procedures.

Bibliography

[1] CHIU, Tom; FANG, DongPing; CHEN, John; WANG, Yao; JERIS, C. A. (2001). *A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment*. Proceedings of the seventh ACM IGKDD international conference on knowledge discovery and data mining.

[2] FAYYAD, Usama, (edit.); PIATETSKY-SHAPHIRO, Gregory, (edit.); SMYTH, Padhraic, (edit.); UTHURUSAMY, Ramasamy, (edit.) (1996). *Advances in Knowledge Discovery and Data Mining*. Cambridge, MA. MIT Press.

[3] KAUFMAN Leonard; ROUSSEEUW, Peter J.; (1990). *Finding Groups in Data: an Introduction to Cluster Analysis.* John Wiley & Sons.

[4] LUAN, Jing. (2001). *Data Mining as Driven by Knowledge Management in Higher Education. Persistence Clustering And Prediction.* Keynote for SPSS Public Conference, UCSF.http://www.cabrillo.edu/services/pro/oir_reports/UCSFpaper.pdf> [Consultation: 2 May. 2005]

[5] MALHOTRA, Yogesh (1998). Knowledge Management for the Ner World of Business. "Deciphering the Knowledge Mangement Hype". The Journal for Quality & Participation, July/august 1998. Association for Quality & Participation.

[6] MINTZBERG, Henry (1984). *La estructuración de las organizaciones*. Barcelona. Ariel.

[7] ORALLO, José Hernández; RAMÍREZ QUINTANA, María José; FERRI RAMÍREZ, César (2004). *Introducción a la Minería de Datos*. Madrid. Pearson Educación.

[8] SANTESMASES, Mestre Miguel. (2001). *DYANE 2 Diseño y análisis de encuestas en investigación social y de mercados.* Dyane. 2 Version. Pirámide.

[9] SIMON, Herbert A. (1980). *El comportamiento administrativo. Estudio de los procesos decisorios en la organización administrativa.* Buenos Aires. 3rd Edition, Aguilar.

[10] SPSS (2001). *The SPSS TwoStep Cluster Component: A scalable component enabling more efficient customer segmentation* [on line]. In: White paper: technical report. <http://www.spss.com/downloads/Papers.cfm?List=all&Name=all>. [Consultation: 1 jun. 2005], available under registry.

[11] ZHANG, Tian; RAMAKRISHNAN, Raghu; LIVNY, Miron (1996). BIRCH: *An Efficient Data Clustering Method for Very Large Databases.* Proceedings of the ACM SIGMOD Conference on Management of Data. Montreal, Canadá.