Perspectives of Statistical Analysis in the Study of Scientific and Professional Migrations

Alvaro Montenegro

Introduction

This paper presents the methodological study to carry out the survey “Colombia Network”, within the framework of the research project “Brain gain revisited through the Colombian case. Study of the Caldas Network” [1]. The objective of the project is to make a detailed analysis of the population of expatriate Colombian researchers, organized mainly around the Caldas Network [2]. The survey was sent to expatriate Colombian researchers that this group of researchers located in 24 countries. Different means were used for sending the survey, among them mainly the air mail and the electronic mail.

The tools

The statistical analysis of the information collected, requires at least the use of the following tools

Data base system

The costs for launching and gathering the information in an international survey is very high, and thus the information collected must be sufficiently wide, in such a way that the most advantage is taken of the effort made. This implies that the survey will have information of very varied types and for the analyses complex reports will be required. Thus, it is indispensable to build in a technical manner a standardized data base to store the information. The information must be entered to the data base already coded, so it will be necessary to use standard classification tables. In the case of the survey “Colombia Networks”, UNESCO discipline tables and the Pascal base classification system were used for coding the schooling, research fields and labor activities data.

Programming language

Experience has shown us that notwithstanding counting with a sophisticated data basis and having experience in its use, the preparation of the data for statistical analysis, after they are out of the data basis is in many cases very complex. According to the kind of analysis required, the data must be reorganized and coded again, before entering to the statistical tools.

Electronic sheet

The modern electronic sheet is of invaluable help for the manipulation of information. The best statistical graphs may be obtained from this tool. In addition, a good part of the new coding and data reorganization work may be done with this tool.

Statistical tools

The election of statistical tools depends on the type of analysis required. In social studies, the most usual recommendation is to use tools based on data analysis techniques. As discussed in section 2, the purpose of the French approach of data analysis is to discover and describe the characteristics present in a population, as opposed to the parametric methods, the purpose of which is to model and then test.

Techniques

In addition to the descriptive analysis techniques, the data analysis techniques are the most indicated techniques for social studies. For the analysis of the survey “Colombia Networks” the following data analysis techniques have been used :

Simple Correspondence Analysis (SCA)

By means of this technique it is possible to compare two categorical variables. In the SCA answer profiles are built from the cross matrix of the two variables, and their purpose is to find an space in which the total inertia of the data may be broken down only along the new axes. The initial data matrix is of a size *n*m*, where n and m are respectively the number of categories of each variable, and the element *ij* of the matrix is the number of individuals who simultaneously present the categories *i* of the first variable and *j* of the second variable. In the calculation the distance chi-square is used, which has the characteristic of giving less weight to the more frequent categories, and more weight to the less frequent ones, which just because of their low frequency are the ones that determine the analysis. The final result is the obtaining of the charts (factorial charts) in which you can observe the relationships between some categories and others, taking into account the following interpretation principles :

- The two variables can be represented simultaneously in the same factorial chart.
- The most frequent modalities are represented close to the new coordinated origin. These are the common characteristics of the population. The less frequent modalities appear far from the origin. These are the characteristics that differentiate the population.
- The modalities that appear relatively close to each other are characteristics of a same group of individuals and therefore characterize same.
- The position of any modality *i* (respectively *j*) in a factorial chart is the baricenter of all modalities (respectively *i*) of the other variable which were selected simultaneously with the modality *i* (respectively *j*) in the answers to the survey.
- Additional modalities (which will be explanatory) may be plotted in the factorial chart in order to complete the characterization of the groups present in the population

Multiple Correspondence Analysis (MCA)

The MCA is the natural extension of SCA, to analyze simultaneously multiple variables. The initial matrix for the analysis is now an n*p matrix, where n is the size of the sample and p the total number of categories present, including all the variables for which the analysis will be made. The objective of MCA is the same as in SCA, that is, it seeks to find a space in which the inertia of the cloud of dots will be totally broken down along the new coordinated axes. The chi-square distance is also used and has the same effects as before. The principles of interpretation are similar to the SCA, taking into account that in this case the origin of the factorial charts is the baricenter of all categories.

Analysis of textual data

It is the most recent data analysis technique. It derives from SCA, with the characteristic that it adapts to the management of very disperse huge matrixes. This technique was developed for the analysis of literary texts and particularly for the analysis of open questions of surveys. The method base is in the creation of the lexical variable where its categories are each one of the different words present in the text. The matrix for the analyses is generally an *n*p* matrix where *n* is the number of texts and *p* the number of categories of the lexical variable, and the analysis is a SCA. In the study of the survey "Colombia Networks" it has been found that the tool may be successfully used in the analysis of biographical information and in obtaining scientific charts from key words.

Analysis of clusters

The three above mentioned techniques are sufficiently descriptive so that with some experience the researcher may have an idea of what happens in the population, form the direct observation of different factorial charts. However, in order to have an exact description of the population studied, it is necessary to make cluster analyses on the date from their locations on the factorial charts. This process will permit at the end to describe the population such as it is without losing the multi-variant reality present in the data.

The associate words analysis

This technique is used in scientometrics to analyze the contents of a documental data corpus data that is built by key words used to index the original documents.

The base of the method is to calculate the association coefficient between two words. Let´s be *i, j* two words of the corpus. The association coefficient between *i* and *j* is defined by

$$E_{ij} = \left(\frac{c_{ij}}{c_i}\right)\left(\frac{c_{ij}}{c_j}\right)$$

where c_i and c_j are the frequences of *i* and *j* words respectively in all corpus, and c_{ij} is the frecuency of coocurrence of the words *i* and *j* in a same text.

The association coefficient is calcutated for all couples of words that have a frecuency greater than a threshold. From the association coefficients a cluster analysis is made, and clusters of word are obtained. Each cluster determines a theme present in the corpus.

Two index are calculate now for each cluster. The centrality index and the density index. The cen-

trality index of a cluster of words is the mean of the association index betwen the words of the cluster and the words of the others clusters. In other words this is a relational index inter clusters. A theme is more central if it´s more connected with the others. The density index is the mean of the the association coefficients inside the cluster. A theme is more dense if it has more development. The two inedex allow to built a two dimensional strategic diagram, by cross of the centrality index (first axis) and density index (second axis) of each cluster. The diagram´s origin is the median of centrality index and the density index respectively. The diagram has four quadrants that can be interpreted.

- First quadrant (greater centrality, smaller density) presents the development themes, the reference themes.

- Second quadrant (smaller centrality, greater density) presents las temáticas desarrolladas con poca influencia global.

- Third quadrant (smaller centrality, smaller density) presents the new themes, the nascent themes.

- Fourth quadrant (greater centrality, smaller density) presents the developing themes, the link themes, the promising themes.

Data types

The data types of the survey must be carefully defined before designing the survey and must correspond to the purposes of the research. In the case of the survey Colombia Networks, the object in general is to describe the situation of the identified group of expatriate scientific Colombians in terms of :

Socio-demographic data

- Age, sex, nationality (ies) of the surveyed person and his close relatives, schooling, current activity.

Biographic data

- Residential history
- Academic history
- Labor history

Research Field data

- Fields and key words that describe the contents of the research.

Factual Data

- Type of active relationships maintained with Colombia.
- Type of entity in which he works.
- Communication media used.
- Belonging to Colombian association networks

Opinion data

- Level of satisfaction and expectations with respect to current work.
- Benefits and contributions expected with respect to the Caldas Network.
- Favorableness and difficulty conditions to establish active relations with Colombia.
- Successes and problems of the Caldas Network, evolution of science and technology in Colombia (open questions)

Publication data

- References data.

Methodology

Figure 1. shows the methodological process that has been followed in the survey "Colombia networks". The methodology is applicable to any study of this type. Each one of the steps is described below.

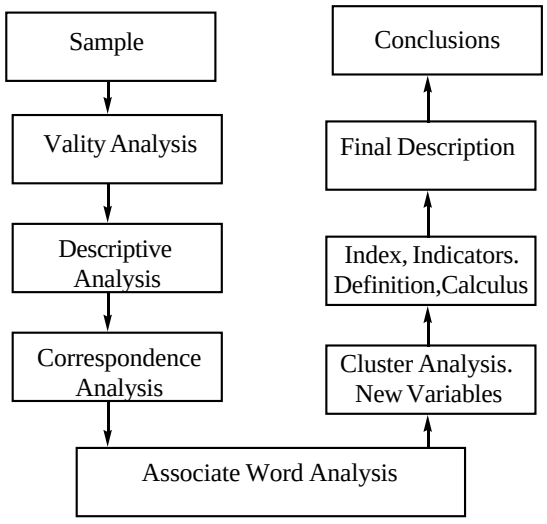


Figure 1. Scheme of methodology for the statistical analysis of scientific and proffesional migrations

The sample

In a type of survey such as the "Colombia Networks" survey, it is highly unlikely that a sampling can be made following some of the traditional sampling models because of the following reasons :

- There is not a complete sample framework.
- It is not possible to interview the surveyed people directly.
- There is no certainty of locating each individual, due to their mobility.
- The probability of an answer is not the same in each country.
- The object of the study is very dynamic.

The surveys for the study of scientific and proffesional migrations must be sent to different parts of the world. The primary source of information to locate the individuals to be studied is

found in the government files which generally are incomplete and are not always updated. Therefore new sources must be located, in such a way that it will be possible to find other individuals and thus increase the framework. For the "Colombia Networks" survey, the start was Colciencias data base, an entity that orients the policies of science and technology in Colombia, in which were recorded and totally identified 826 Colombian researchers and proffesionals abroad, in the year 1994. Our researchers were located thanks to the information received from some of the people surveyed, and others from lists of node coordinators of Network Caldas at several countries.

The methodology for obtaining the sample must take into account the difficulties mentioned above, in order to guarantee to the extent possible the reliability of the results obtained. The steps proposed are the following :

Decide the number of forms to be sent

With the information available on the locations of people, it is necessary to decide, in accordance with the available budget, what will be the number of forms sent, and which the method for sending them. All available information must be used, in order to obtain a sample that will be really significant of the population studied.

A prior estimate of some parameters results useful for decision making prior to sending and after receipt of the forms. The hypothesis that can be handled for example is that the government data base represents a good sample of the population studied. If this is assumed, then it is possible to make for the case of the scientific and proffesional migrations the following estimates :

- The population is considered geographically divided, and thus it is possible to assume a multinomial model, where each category is a receiving country. Then it is possible to estimate the proportion of individuals in each country. Note that it is not possible to establish confidence intervals, since the size of the population is unknown.
- If there is any additional information available, for example the work area, research, schooling, etc., it may be used additionally. For example, in the case of the study "Colombia Networks", the classification from the research programs of Colciencias was available. This additional information permits to estimate parameters of interest for the study. For the case of scientific migrations it is more important that the sample will be very representative of the scientific areas, even though it may get a little (although not too much) far from the geographic distribution.

Receive the surveys and store the data

All the information coming from the surveys must be stored in the data base built for that purpose. Figure 2 presents the main menu of the data base of the survey "Colombia Networks".

Analysis of the validity and representativeness of the survey

It is the first important step of statistical analysis and can not be omitted. From the result of this analysis will depend whether a subsample should be taken or give different weights to some individuals. In addition, it should be indicated whether the conclusions are really applicable to all the population under study. Within the study of the survey "Colombia Networks" the chi-square conformity tests described below were performed.

- A comparison was made of the distribution of sending by countries and receiving by countries. The distribution of receiving was perfectly in conformity. Figure 3 shows the comparison.

- A comparison was made of the distribution by countries of Colciencias data base and the subsample resulting from the surveys obtained from the Colciencias data base. The subsample was perfectly in conformity. Figure3 shows de comparision.

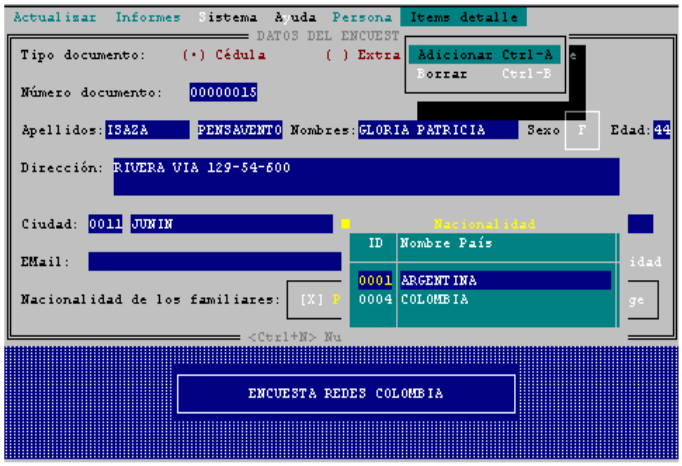


Figure 2. Main menu of “Colombia Networks” Survey

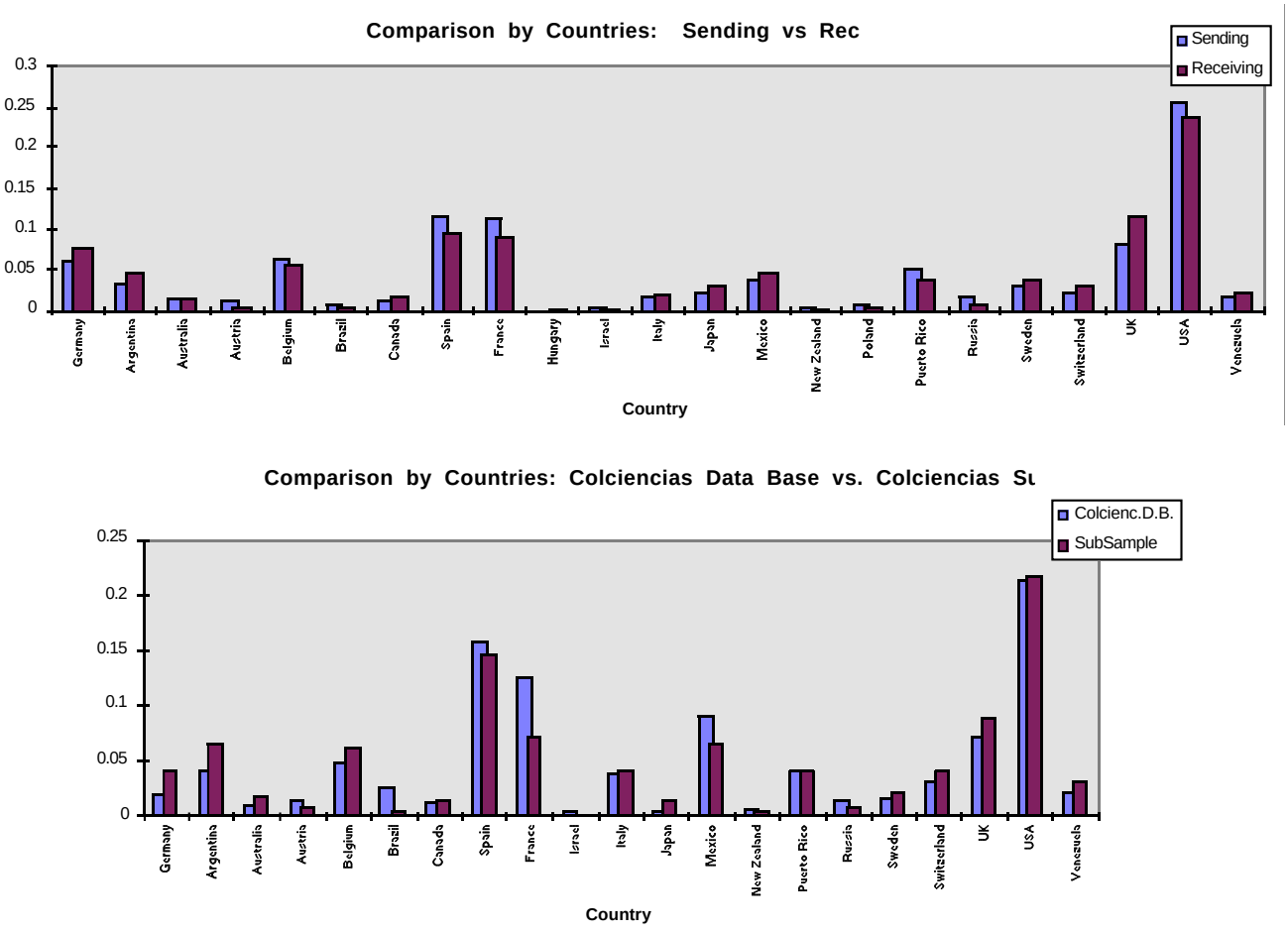


Figure 3.

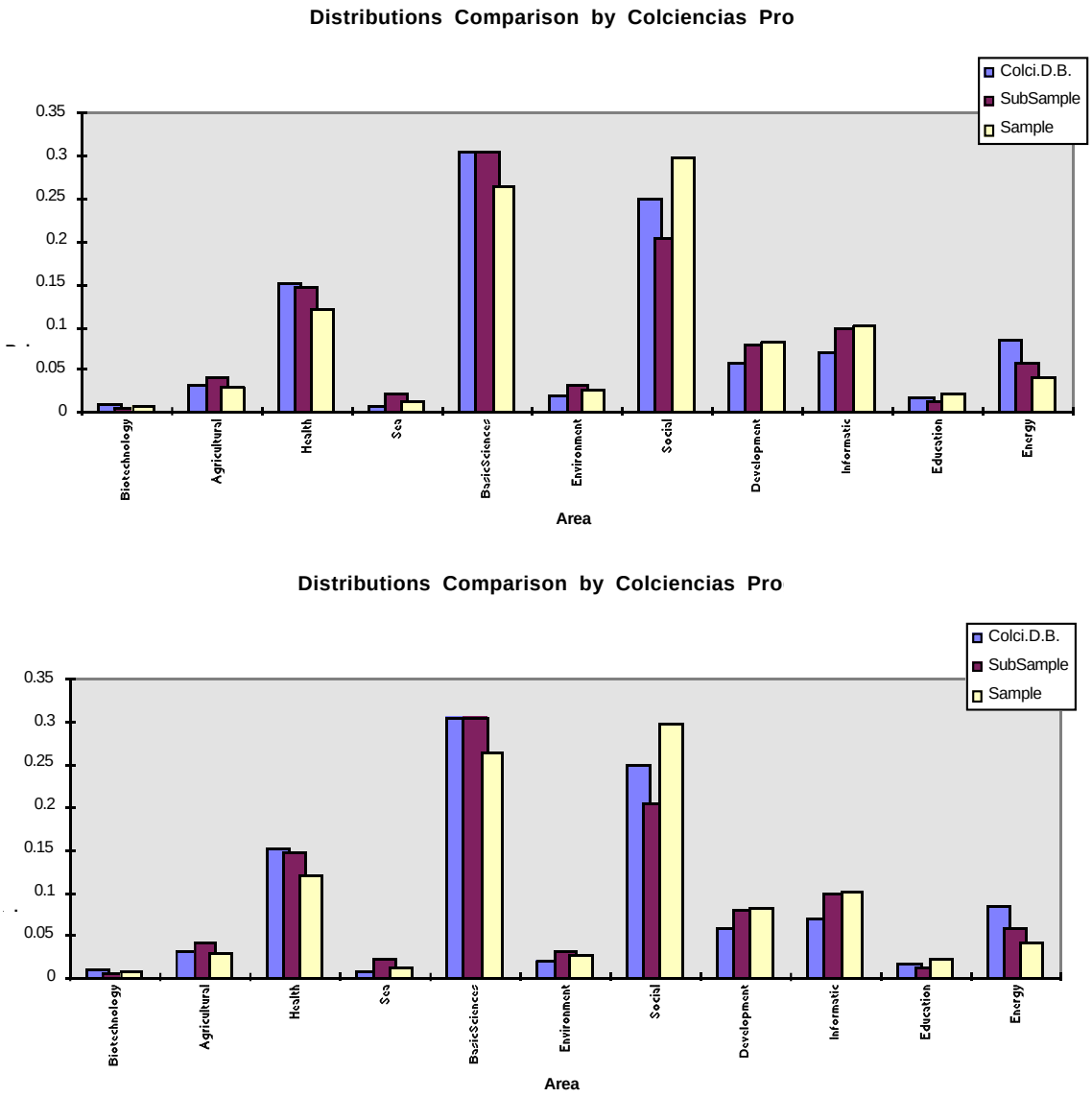


Figure 3. Sample Analysis of Survey “Colombia Networks”. Comparison of distributions

- A comparison was made of the distribution by research programs, of the subsample resulting from the surveys obtained from the Colciencias data base, with the parameters from Colciencias, with positive result.

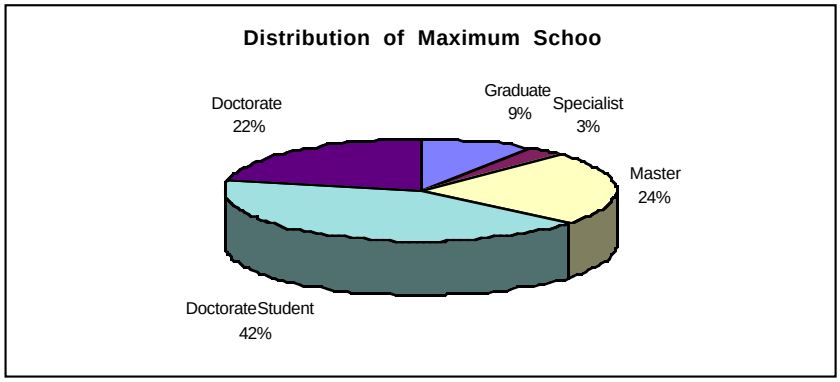
- The almost full independence of answer within each country was verified, comparing the distribution, by research programs for the sample and subsample respectively. In the case of the total sample only in two countries, the parameters did not conform well. He have in this cases more individuals located, coming from sources other than Colciencias.

The conclusion is that the sample is very representative of the population studied and therefore, how the study is principally descriptive (exploratory) it is possible to continue with the following step of the analysis without any modification. Figure 3 shows the comparison of the distributions by research programs of the total sample and of Colciencias data base of 1994, in the survey “Colombia Networks”, in general and the internal comparison within the United States.

Finally, it´s necessary to understand that the results obtained should be interpreted how trends that is presents in the populations, and they are not parameters.

First descriptive analysis

This first step is necessary in order to have a first impression of the global characteristics of the population even in an isolated manner. Some first conclusions may be obtained already. For example, it is possible to say that the population examined in the survey "Colombia Network" is made up by 69 % men, that most of the population distributes their time between work and research (35 %) and between study and research (25 %). That 87 % of the population surveyed has to do with research processes, either as students who pursue a Masters or Ph.D. degree or as professional researchers, that 83 % have a schooling level between a Masters and a Ph.D. degree, that the most frequent labor migrations occur between the countries of the United States, France, Spain, England, Mexico, Germany, Argentina, Brazil, Venezuela and obviously Colombia. That most people migrate for the first time for school reasons, etc. This information is in fact useful for some decision making. In effect, it may be strategic information for some entities. For us, it represents really the first contact with the data. Figure 4 shows two examples:



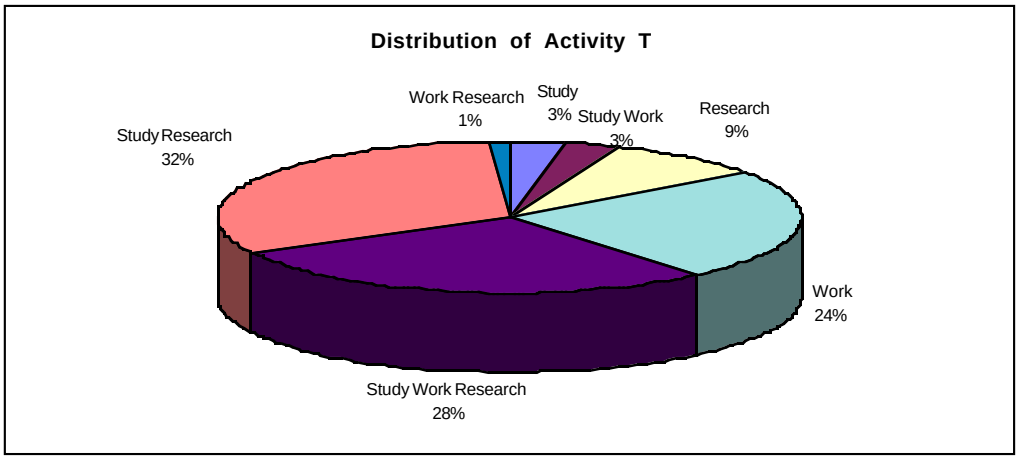


Figure 4. Examples of descriptive analysis

First data analyses by subject (First data classification). Some examples

This step is one of the most complex ones. It is the step when it will be necessary to make new codifications, reports in different ways, etc.

The purpose is to advance by subjects in the analysis. Every subject identified is treated separately. It is not a better analysis that which is made putting together all the variables, among other things, because each subject by itself requires attention and possibly differential treatment. In the case of the survey "Colombia Networks", some analysis were made, among others, which results and most relevant aspects are noted below. Only a few examples have been chosen.

- Socio-demographic analysis

The socio-demographic variables were chosen, with the idea of establishing a classification of the population from that point of view. The result obtained is that the entire population is very homogeneous and with the exception of the family ties that connect some of the people surveyed with the country of residence, the rest of the characteristics are common to most of the population.

- Active relations with Colombia

Fourteen variables were used, with a total of 10 categories. The variables were submitted to a MCA, the first factorial chart of which is shown in figure 5. The analysis seems to show the existence of four large groups. Some that maintain formal relations, others that maintain temporary relationships and others that do not have any relations with Colombia, and a cluster of no answer. The analysis of clusters confirms this suspicion and in addition, the projection of socio-demographic variables helps to find some additional distinctive characteristics of each group. In the next step, each one of the clusters must be analyzed in depth. As of now, it has been sufficient to identify the groups and have a first description of each. The final step in this part is to create a new variable, the Active Relations variable, with four categories, each one associated to one of the resulting clusters. Figure 6 presents the description of one of the clusters. Both results come from the statistical program, then from the SCA and from the classification, respectively.

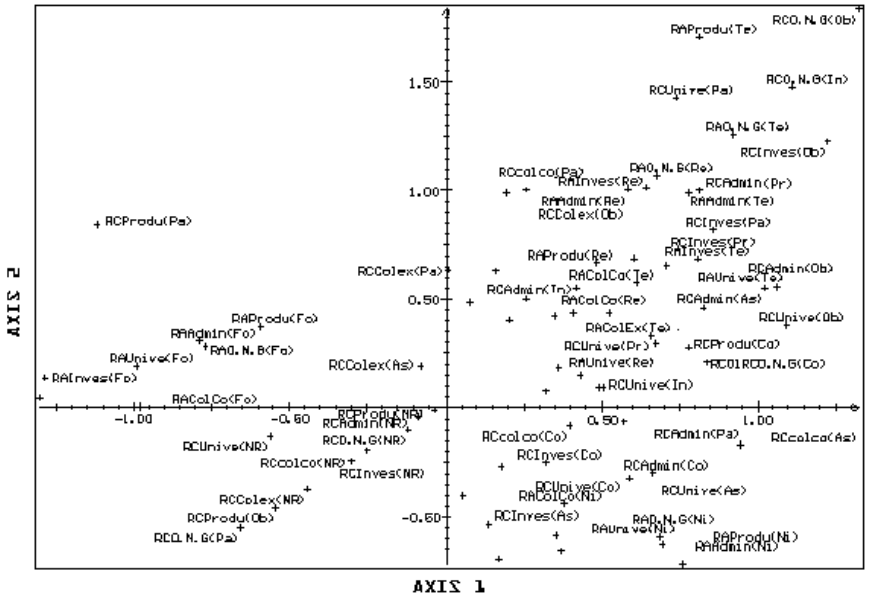


Figure 5. Factorial chart of the subject active relations with Colombia (14) Variables

GLOBAL CHARACTERISTIC					
V.TEST	PROB.	PERCENTS	CLAS/MOD	MOD/CLAS.	MODALITY
8.54	.000	66.67	46.67	9.27	RAColCo(Re)
8.23	.000	84.00	35.00	5.52	RAInves(Re)
7.86	.000	51.67	51.67	13.25	RCUnive(Pr)
7.45	.000	52.83	46.67	11.70	RAUnive(Re)
6.43	.000	68.00	28.33	5.52	RAAdmin(Re)
5.92	.000	36.71	48.33	17.44	RAColEx(Re)
5.10	.000	47.22	28.33	7.95	RCInves(Pr)
4.64	.000	41.46	28.33	9.05	Rccolco(Pr)
4.33	.000	72.73	13.33	2.43	RAO.N.G(Re)
3.86	.000	23.66	51.67	28.92	RCColex(Pr)
2.96	.002	35.71	16.67	6.18	fisica
2.39	.009	45.45	8.33	2.43	RCAdmin(In)
- 2.41	.008	8.96	31.67	46.80	RAAdmin(Fo)
- 2.43	.008	11.66	78.33	88.96	RCAdmin(NR)

Figure 6. Description of Cluster 2 of the subject active relationships with Colombia. The upper list are the dominant characteristics and the lower list are the farthest characteristics.

- Analysis of residential migration

Recent techniques such as the qualitative harmonic analysis have been developed although not fully implemented in a statistical program, for the analysis of migration routes. Within the analysis of the survey “Colombia Networks” a strategy based on the analysis of textual data has been successfully used, which we describe below because it represents a methodological contribution for this type of problems. For the study of longitudinal data, the “traditional” strategy which has been followed is as follows : first time is set at discretion, defining periods (which may be historic or the age of the population surveyed), and grouping the locations of destination in accordance to some criterion. Then the two variables are crossed, time set at discretion by groups of destination obtaining a variable of states where every modality is a time space by a geographic space. To every category is associated the percentage of stay in such states for every individual, thus obtaining a matrix of individuals by states, which is submitted to analysis of correspondence and then to the classification. The customary interpretation techniques are then used.

The technique of textual data analysis has been used in some studies although according to our criterion, in an erroneous manner. What we have done is code again the data as shown in figure 7. That is, we have assumed for every individual a textual answer of the form:

Country1 Area_Continental_1 country1_age1 country1_age2...

Country1 Area_Continental_1 country1_age1 country1_age2...

That is, we have described the entire residential history year by year in the original countries of migration. Every line corresponds to a stay. The objective of placing the name of the country and the continental area is to get neighboring routes closer, that is, we have introduced a neighboring effect. The analysis of correspondence is applied to this file from the point of view of the textual analysis, then the classification and the usual interpretation techniques may be used. Here two individuals are similar because they use approximately the same vocabulary, which is translated in migratory terms in that they follow approximately the same route. The results obtained that are shown in figures 7, 8 and 9 evidence the goodness of the method from the point of view of interpretation. Note particularly that in the case of the study of the survey “Colombia Networks”, the migration is discriminated in particular by the receiving countries and not by a high mobility, as it is the case in urban migrations.

Original Data I.D.	Country	From Date	To Date	Current Age
0001	Germany	1980	1985	34
0001		1986	1989	34
0001	Brazil	1990	1994	34
0002	Argentina			

Coded Data

——0001
COLOMBIA COLOMBIA16 COLOMBIA17 COLOMBIA18 COLOMBIA19
GERMANY EUROPEAN_COM GERMANY20 GERMANY13 GERMANY22 GERMANY 23
BRAZIL SOUTH_AMERICA BRAZIL26 BRAZIL27 BRAZIL28 BRAZIL29
ARGENTINA SOUTH_AMERICA ARGENTINA30 ARGENTINA31 ARGENTINA32
——0002

Figure 7. New coding of the residential migration data for textual analyses. Residential history file

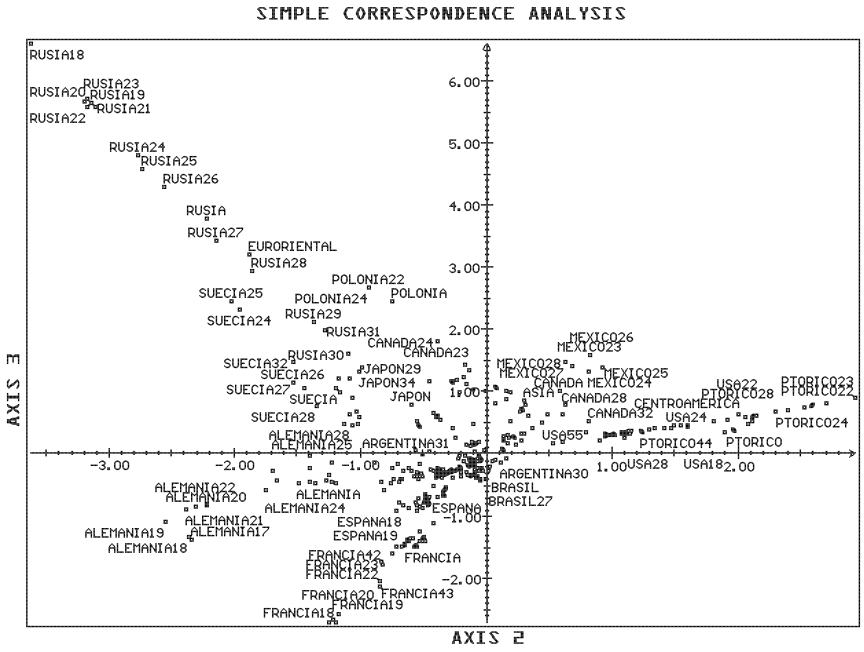


Figure 8. Sample factorial chart of residential analysis output

---

CRITERE DE CLASSIFICATION — RESPONSE OU INDIVIDU CARACTERISTIQUE

6.000 -- 1 COLOMBIA16 COLOMBIA17 COLOMBIA18 COLOMBIA19 COLOMBIA20
-- COLOMBIA21 COLOMBIA22
-- USA NORTAMER, USA23 USA24 USA25 USA26 USA27 USA28 USA29 USA30
-- USA31 USA32 USA33 USA34 USA35

6.000 -- 2 COLOMBIA16 COLOMBIA17 COLOMBIA18 COLOMBIA19 COLOMBIA20
-- COLOMBIA21 COLOMBIA22
-- USA NORTAMER, USA23 USA24 USA25 USA26 USA27 USA28 USA29 USA30
-- USA31 USA32 USA33 USA34 USA35

5.801 -- 3 COLOMBIA16 COLOMBIA17 COLOMBIA18 COLOMBIA19 COLOMBIA20
-- COLOMBIA21
-- USA NORTAMER, USA22 USA23 USA24 USA25 USA26 USA27 USA28 USA29
-- USA30 USA31 USA32 USA33

5.801 -- 4 COLOMBIA16 COLOMBIA17 COLOMBIA18 COLOMBIA19 COLOMBIA20
-- COLOMBIA21
-- USA NORTAMER, USA22 USA23 USA24 USA25 USA26 USA27 USA28 USA29
-- USA30 USA31 USA32 USA33

2.747 -- 1 COLOMBIA16 COLOMBIA17 COLOMBIA18 COLOMBIA19 COLOMBIA20
-- COLOMBIA21 COLOMBIA22 COLOMBIA23 COLOMBIA24 COLOMBIA25
-- ARGENTINA SURAMERICA, ARGENTINA26 ARGENTINA27 ARGENTINA28
-- ARGENTINA29 ARGENTINA30 ARGENTINA31 ARGENTINA32 ARGENTINA33
-- ARGENTINA34 ARGENTINA35 ARGENTINA36 ARGENTINA37 ARGENTINA38
-- ARGENTINA39

2.533 -- 2 COLOMBIA16 COLOMBIA17 COLOMBIA18 COLOMBIA19 COLOMBIA20
-- COLOMBIA21 COLOMBIA22 COLOMBIA23 COLOMBIA24 COLOMBIA25 COLOMBIA26
-- COLOMBIA27 COLOMBIA28
-- ARGENTINA SURAMERICA, ARGENTINA29 ARGENTINA30 ARGENTINA31
-- ARGENTINA32 ARGENTINA33 ARGENTINA34 ARGENTINA35 ARGENTINA36
-- ARGENTINA37 ARGENTINA38 ARGENTINA39 ARGENTINA40 ARGENTINA41
-- ARGENTINA42 ARGENTINA43

2.504 -- 3 COLOMBIA16 COLOMBIA17 COLOMBIA18 COLOMBIA19 COLOMBIA20
-- COLOMBIA21 COLOMBIA22 COLOMBIA23 COLOMBIA24 COLOMBIA25 COLOMBIA26
-- COLOMBIA27 COLOMBIA28
-- ARGENTINA SURAMERICA, ARGENTINA29 ARGENTINA30 ARGENTINA31
-- ARGENTINA32 ARGENTINA33 ARGENTINA34 ARGENTINA35 ARGENTINA36
ARGENTINA37 ARGENTINA38 ARGENTINA39 ARGENTINA40 ARGENTINA41

Figure 9. Some characteristic answers of two clusters

- Analysis of research fields

Figure 10 shows the original source of the data collected in the survey “Colombia Networks”. The purpose of this part of the study is to establish the fields in which the researchers are working. The first attempt of an analysis consisted in making a textual analysis of the key words answered by the people such as they had been adding the fields that they had answered themselves. The technique that used consisted in adding implicit information present in the answers. For this we use the classification system of the Pascal base corresponding to the research field answered. The key words were left just as they arrived. Figure 10 shows how each answer was prepared for the textual analysis (correspondence analysis). Figures 11, 12 and 13 give an output example of the textual analysis, and figures 14 and 15 give the same output for the associate words method.

Research fields

0001	CSocioEconomica	CJuridicas		CIENCIAS_POLITICAS
0003	CSocioEconomica	Economia General	EconInternacional	RELACION_INTERNAL SUDASIA
0004	CMedicaBiologia	Biologia	AnimalProducc	APICULTURA APITOXINA
0005	CSocioEconomica	EconoEnergia	NuclearEnerg	TECNOLOGIA_INDUSTR TECNOLOGIA_AVANZADA
0006	CMedicaBiologia	Biologia	VertebAnatSico	MEDICINA_VETERINARIA

Keywords

0001	AMERICA_LATINA
0001	ASIA__PACIFICO
0001	ECONOMIA_INTEGRACION
0003	CULTURA_IDENTIDAD
0003	ECONOMICO_DESARR
0003	POLITICA
0003	POLITICA_PROCESO
0003	RELACION_INTERNAL
0004	APICULTURA
0004	APITOXINA_EXTRACCION
0004	COLMENA
0004	STRESS
0005	ELECTROMEC_CONTROL
0005	ROBOTICA
0005	TECNOL_INTEGRACION
0005	TECNOLOGIA_BLANDA
0006	ANIMAL
0006	COMPORT_ANORMAL
0006	RESULTADOS_INFLUENC
0006	STRESS

Archivo para análisis

——0001

0001 CSocioEconomica CJuridicas CIENCIAS_POLITICAS
AMERICA_LATINA ASIA_PACIFICO ECONOMIA_INTEGRACION.

——0003

0003 CSocioEconomica EconomiaGeneral EconInternacional RELACION_INTERNAL SUDASIA
CULTURA_IDENTIDAD ECONOMICO_DESARR POLITICA POLITICA_PROCESO
RELACION_INTERNAL.

——0004

0004 CMedicaBiologia Biologia Animal Producc APICULTURA APITOXINA
APICULTURA APITOXINA_EXTRACCION COLMENA STRESS.

——0005

0005 CSocioEconomica EconoEnergia NuclearEnerg TECNOLOGIA_INDUSTR TECNOLOGIA_AVANZADA
ELECTROMEC_CONTROL ROBOTICA TECNOL_INTEGRACION
TECNOLOGIA_BLANDA.

——0006

0006 CMedicaBiologia Biologia VertebAnatSico MEDICINA_VETERINARIA
ANIMAL COMPORT_ANORMAL RESULTADOS_INFLUENC STRESS

Figure 10. New coding of data for analysis of research fields

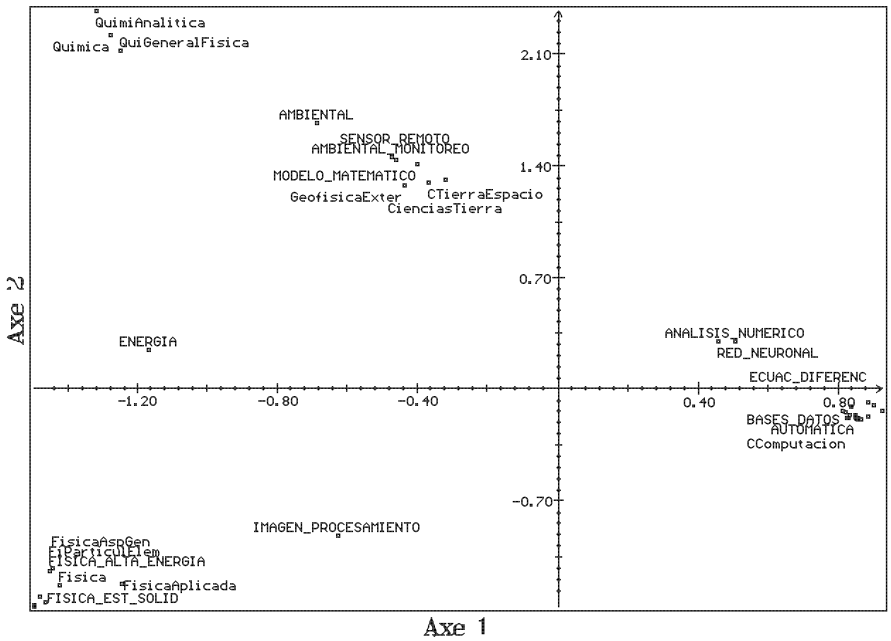


Figure 11. Example of factorial chart in the analysis of the subject of research fields Exact and Technology Sciences group

TEXTE NUMERO 6 a06a = CLASSE 6 / 13

	LIBELLE DE LA FORME GRAPHIQUE	---POURCENTAGE--- INTERNE	GLOBAL	INTERNE
1	BASES	14.29	.53	4.
2	DATOS	14.29	.80	4.
3	BASES_DATOS	10.71	.40	3.
4	MULTIMEDIA	10.71	.53	3.
5	CACienciasCompu	14.29	4.38	4.
6	MODELO	3.57	.53	1.
7	INFORMATICA	3.57	.53	1.
8	CienAplicadas	14.29	9.96	4.
2	Fisica	.00	3.19	0.
1	CExactasTecnologia	14.29	18.59	4.

Figure 12. Internal description of a research field cluster

TEXTE NUMERO 6 a06a = CLASSE 6 / 13

CRITERE DE CLASSIFICATION	REPONSE OU INDIVIDU CARACTERISTIQUE
2.168 --	1 CExactasTecnologia CienAplicadas
--	BASES_DATOS MULTIMEDIA
--	ANIMACION, BASES, DATOS, MEMORIA, MULTIMEDIA,
2.078 --	2 CExactasTecnologia CienAplicadas
--	BASES_DATOS
--	BASES, DATOS, PROTOTIPO, .
1.302 --	3 CExactasTecnologia CienAplicadas
--	BASES_DATOS
--	DEDUCTIVOS, DEPENDENCIA, MULTIVARIADA, AXIOMAS
--	INFERENCIA, BASES, DATOS, MODELO, .
1.273 --	4 CExactasTecnologia CienAplicadas
--	INFORMATICA

Figure 13. Some answers characteristics of a research cluster

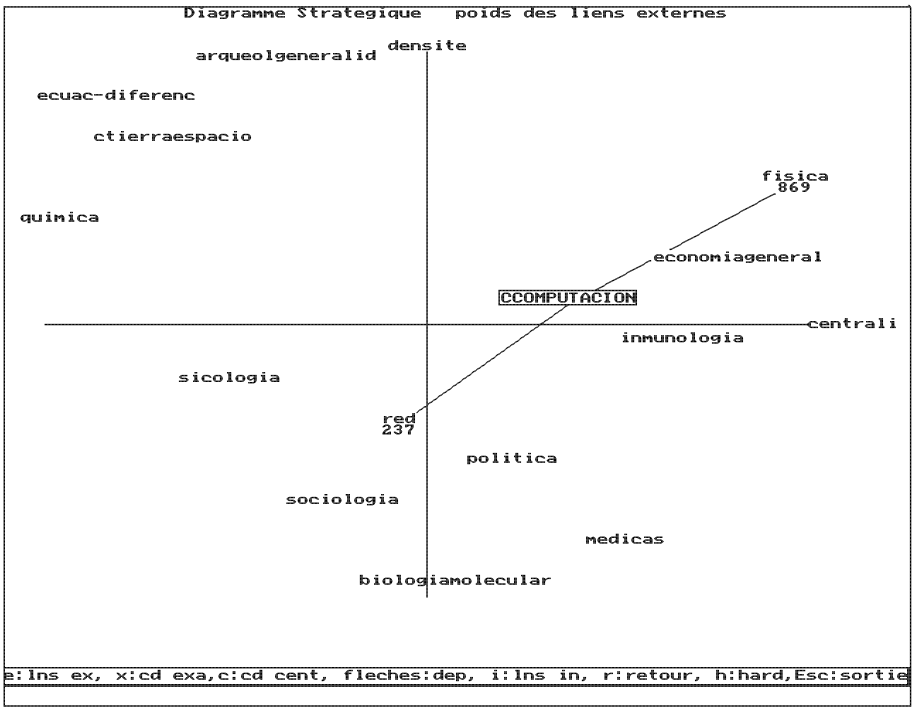


Figure 14. Example of associate words method for the research subject

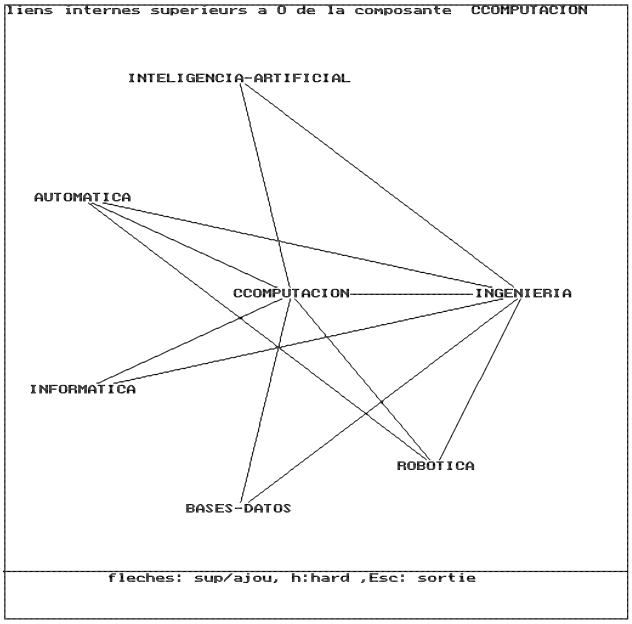


Figure 15. Example of a cluster of associate words method for research subject

Internal analysis of clusters

To discover the composition of the population by subjects is not sufficient for the researcher. The next step is to make the internal description of each cluster. From the analysis of the cluster already some distinctive features of each one may have been possibly discovered, but only the internal descriptive analysis may reveal the characteristics of each group. The graph of figure 16, for example, shows how the study within a cluster may be made, how the internal

movement is, that is, the individuals belonging to each cluster are internally typified. Other descriptive analysis will reveal for example, where they are located, what do they do, what it their idea about returning to the country, etc., of, for instance, the researchers in the area of robotics.

Conclusion

One of the great gaps that this research group has found in the review of studies in the area of scientific migrations is the non use of the modern techniques of statistical analysis, basically coming from the French school. May be the absence of an adequate bibliography which without disregarding a minimum mathematical treatment will deal with the concrete applications in the study of social sciences may be one of the main causes. The study project "The brain gain revisited through the Colombian case. Study of the Caldas Network", of which the survey "Colombia Networks" forms a part, has had a multi-disciplinary team including statisticians which has permitted to advance in a significant manner in the utilization of such tools in this type of studies, even presenting methodological proposals. The purpose of these notes has been to present an advance of the methodology which has been successfully applied in the project.

References

Barbary, O. (1994). Análisis de Datos Biográficos. Simposio de Estadística, Departamento de Matemáticas y Estadística, Universidad Nacional de Colombia, Santafé de Bogotá.

Bécue, M. (1991). Análisis de Datos Textuales. Métodos Estadísticos y Algoritmos. CISIA, París.

Benzecri, J. P. (1976). L'Analyse des Données, Tome 1 : La Taxonomie, Tome 2 : L'Analyse des Correspondence. Dunod, París (2da ed. 1976).

Centre de la Sociologie de l´Innovation (1994). LEXIMAPPE - DOC, INIST, Paris.

Courtial, J. P. (1994). Science Cognitive et sociologie des sciences, Presses Universitaires de France, Colection sous la direction de J. P. Courtial, Paris.

Crivisqui, E. (1993). Análisis factorial de correspondencias un instrumento de investigación en ciencias sociales. Laboratorio de Informática Social Universidad Católica de Asunción, Asunción.

Hausler L. (1993). Des phrases et des itinéraires, en « Actes des secondes journées internationales d'analyse statistique de données textuelles », Montpellier 21-22 de octubre de 1993, Ecole Nationale Supérieure des Télécommunications, Paris.

Houzel Y. y LE Vaillant M : [1994] : Analyse statistique de données textuelles et traitement des données de calendriers : application à l'analyse de l'insertion professionnelle des élèves issus des écoles d'art, en « Actes des journées CEJEE-CEREQ sur les données longitudinales dans l'analyse du marché du travail », Toulouse, octubre de 1993, Ecole Nationale Supérieure des Télécommunications, Paris.

Lebart, L., Morineau, A., and Warwick, K. (1984). Multivariate Descriptive Statistical Analysis. Correspondence Analysis and Related Techniques for Large Matrices. John Wiley & Sons, USA.

Lebart, L., Morineau, A., and Fenelon J. P. (1979) Traitement des Données Statistiques, Méthodes et Programmes. Dunod, París.

Lebart, L., Salem A. (1994). Statistique Textuelle. Dunod, París.

Montenegro A. (1996). Encuesta Redes Colombia. Memorias del Primer Coloquio sobre Ciencia Tecnología y Cultura, Academia Colombiana de Ciencias Exactas Físicas y Naturales, Colección Memorias, No.6. Pags 101-114, Santafé de Bogotá.

Pardo, C.E. (1992). Análisis de la Aplicación del Método de Ward de Clasificación al caso de Variables Cualitativas. Tesis para optar al título de M. Sc. Estadística. Universidad Nacional de Colombia. Santafé de Bogotá.

Montenegro A., Pardo C. E. (1996), Introducción al Análisis Estadístico de Datos Textuales, Departamento de Matemática y Estadística, Unidad de Extensión, Universidad Nacional de Colombia, Bogotá.

Montenegro A., Pardo C. E. (1996), Los itinerarios individuales interpretados como frases, una aplicación de la estadística textual a la tipología de trayectorias, Memorias del Seminario Internacional de capacitación e investigación en Recolección y Análisis de datos longitudinales, Universidad Nacional de Colombia. Bogotá, diciembre de 1996.

[1] This research project is developed by agreement between the institutions Universidad Nacional de Colombia, Universidad del Valle and Institut Français de Recherche pour le Developpement en Coopération-ORSTOM with the support from Colciencias and Icfes.

[2] The Caldas Network was established by Colciencias with the purpose of being the Colombian network of researchers and professionals abroad. The objective of Colciencias by creating it was to encourage the active cooperation of the Colombian scientific and professional community living abroad in the specialized work of the groups of national researchers, opening the possibility for them to join research projects underway or promoting their participation in new ones, as well as to the country's restructuring process which was initiated by President Gaviria. The network forms a part of the special strategies formulated within the national science and technology plan formulated during the Gaviria government, within the item "promotion of groups and networks", leading to the formation of a scientific community.

Perspectivas de análisis estadístico en el estudio de migraciones de científicos y profesionales

Alvaro Montenegro

Introducción

Se presenta el proceso metodológico seguido para llevar a cabo la encuesta “Redes Colombia”, dentro del marco del proyecto de investigación “El brain gain revisited a través del caso colombiano. Estudio de la Red Caldas ” [1]. El objetivo del proyecto es hacer un análisis profundo de la población de investigadores colombianos expatriados, organizados principalmente alrededor de la red Caldas [2]. La encuesta fue enviada a los investigadores colombianos expatriados que este grupo de investigación localizó en 24 países. Se emplearon diferentes medios de envío de la encuesta, entre ellos principalmente el correo aéreo y el correo electrónico.

Las herramientas

El análisis estadístico de la información recolectada, requiere al menos de la utilización de las siguientes herramientas.

Sistema de base de datos

El costos del lanzamiento y recolección de información en una encuesta internacional es bastante elevado, por lo que la información recolectada debe ser suficientemente amplia, de manera que se aproveche al máximo el esfuerzo que se hace. Esto implica que la encuesta tendrá información de muy variado tipo y para los análisis se requerirán reportes complejos. Entonces es indispensable construir de una manera técnica una base de datos normalizada para almacenar la información. La información debe entrar de una vez codificada a la base de datos, por lo que será necesario utilizar tablas de clasificación estándar. En el caso de la encuesta “Redes Colombia”, se utilizaron tablas de disciplinas de la UNESCO, y el sistema de clasificación de la base Pascal, para la codificación de datos de escolaridad, campos de investigación y actividades laborales.

Lenguaje de programación

La experiencia nos ha demostrado que a pesar de contar con una sofisticada base de datos, y de tener experiencia en su uso, la preparación de los datos para los análisis estadísticos, des-

pués de que salen de la base de datos es en muchos casos muy compleja. Según la clase de análisis que se requiera, los datos deben reorganizarse y recodificarse, antes de ingresar a la herramienta estadística.

Hoja electrónica

La hoja electrónica moderna es de invaluable ayuda para la manipulación de la información. Los mejores gráficos estadísticos pueden obtenerse de esta herramienta. Adicionalmente buena parte del trabajo de recodificación y reorganización de los datos puede hacerse con la herramienta.

Herramientas estadísticas

La elección de la herramientas estadísticas depende del tipo de análisis que requieran. En los estudios sociales, lo mas recomendado es utilizar las herramientas basadas en las técnicas del análisis de datos. Como se discute en la sección 2 el propósito del enfoque francés del análisis de datos es descubrir y describir las características presentes en una población, en contraposición con los métodos paramétricos, cuyo propósito es modelar y luego probar.

Las técnicas

Además de las técnicas del análisis descriptivo, las técnicas del análisis de datos son las más indicadas para los estudios sociales. Para el análisis de la encuesta “Redes Colombia se han utilizado las siguientes técnicas del análisis de datos.

Análisis de Correspondencia Simple (ACS)

Mediante esta técnica es posible comparar 2 variables categóricas. En el ACS se construyen perfiles de respuesta a partir de la matriz de cruce de las dos variables, y su propósito es encontrar un espacio en el cual la inercia total de los datos pueda ser descompuesta a o lo largo de los nuevos ejes únicamente. La matriz de datos inicial es de tamaño n*m, en donde n y m son respectivamente el número de categorías de cada variable, y el elemento ij de la matriz es el número de individuos que presentan simultáneamente las categorías i de la primera variable y j de la segunda variable. En los cálculos se utiliza la distancia chi-cuadrado, que tiene la característica de darle menos peso a las categorías mas frecuentes, y mas peso a las menos frecuentes, las cuales precisamente por su baja frecuencia son las que determinan el análisis. El resultado final es la obtención de planos (planos factoriales) en los cuales se pueden apreciar las relaciones entre unas categorías y otras, teniendo en cuenta los siguientes principios de interpretación.

Las dos variables pueden ser representadas simultáneamente en el mismo plano factorial. Las modalidades mas frecuentes quedan representadas cerca del nuevo origen coordenado. Estas son las características comunes de la población. Las modalidades menos frecuentes aparecen alejadas del origen. Estas son las características que diferencian a la población.
Las modalidades que aparecen relativamente cerca entre sí son características de un mismo grupo de individuos y por tanto lo caracterizan.

La posición de una modalidad cualquiera i (respectivamente j) en un plano factorial es el baricentro de todas las modalidades j (respectivamente i) de la otra variable que fueron seleccionadas simultáneamente con la modalidad i (respectivamente j) en las respuestas de la encuesta. Modalidades adicionales (que serán explicativas) pueden proyectarse en el plano factorial con el objeto de completar la caracterización de los grupos presentes en la población.

Análisis de Correspondencia Múltiple (ACM)

El ACM es la extensión natural de ACS, para analizar simultáneamente múltiples variables. La matriz inicial para los análisis es ahora una matriz n*p, en donde n es el tamaño de la muestra y p el número total de categorías presentes, incluyendo todas las variables para las cuales se hará el análisis. El objetivo del ACM es el mismo que en el ACS, es decir, se busca encontrar un espacio en el cual la inercia de la nube de puntos se descomponga totalmente a lo largo de los nuevos ejes coordenados. La distancia chi - cuadrado es también utilizada y tiene los mismos efectos que antes. Los principios de interpretación son similares al ACS, teniendo en cuenta que en este caso el origen de los planos factoriales es el baricentro de todas las categorías.

Análisis de datos textuales

Es la técnica de análisis de datos mas reciente. Se deriva de ACS, con la característica de que se adapta para el manejo de enormes matrices muy dispersas. Esta técnica fue desarrollada para el análisis de textos literarios y en particular para el análisis a preguntas abiertas de encuestas. La base del método esta en la creación de la variable léxica cuyas categorías son cada una de las palabras diferentes presentes en los textos. La matriz para los análisis es por lo general una matriz n*p en donde n es el número de textos y p el número de categorías de la variable léxica, y el análisis es un ACS. En el estudio de la encuesta “Redes Colombia” se ha encontrado que la herramienta puede ser utilizada con éxito en el análisis de información biográfica y en la obtención de cartas científicas a partir de palabras clave.

Análisis de clases

.Las tres técnicas anteriores son lo bastante descriptivas como para que con alguna experiencia el investigador pueda forjarse una idea de lo que sucede en al población, a partir de la observación directa de diferentes planos factoriales. Sin embargo para tener una descripción rigurosa de la población estudiada es necesario hacer procesos de clasificación (cluster analysis), sobre los datos a partir de sus ubicaciones en los planos factoriales. Este proceso permite al final describir la población tal y como es sin perder la realidad multivariante presente en los datos.

Análisis de palabras asociadas

Esta técnica proviene de la cienciometría y es utilizada para el análisis de los contenidos de un corpus de datos documentales que son construidos a partir de palabras que son usadas para indexar los documentos originales.

La base del método es el cálculo del coeficiente de asociación entre dos palabras. Sean i, j dos palabras del corpus.

El coeficiente de asociación entre i y j se define por

$$E_{ij} = \left(\frac{c_{ij}}{c_i}\right)\left(\frac{c_{ij}}{c_j}\right)$$

en donde c_i y c_j son las frecuencias de las palabras i y j respectivamente en todo el corpus, y c_{ij} es la frecuencia de coocurrencia de las palabras i y j en un mismo texto.

El coeficiente de asociación es calculado para todas las parejas de palabras que tienen una frecuencia mayor que un umbral. A partir de los coeficientes de asociación se efectua una clasificación, y se obtienen grupos de palabras. Cada grupo determina una temática presente en el corpus.

A continuación se calculan dos índices para cada grupo. El índice de centralidad y el índice de densidad. El índice de centralidad de un grupo de palabras es la media de los índices de asociación entre las palabras del grupo y las palabras de otros grupos. Es decir, es in índice de relación entre grupos. Una temática es mas central si está mas relacionada con las demás. El índice de densidad es la media de los coeficientes de asociación dentro de un grupo. Una temática es más densa si tiene un mayor desarrollo. Estos índices permiten construir un diagrama estratégico en dos dimensiones, cruzando los índices de centralidad (primer eje) y densidad (segundo eje) de cada uno de los grupos. El origen del diagrama es la mediana de los índices de centralidad y densidad respectivamente. El diagrama se compone de cuatro cuadrantes que pueden ser interpretados.

- El primer cuadrante (mayor centralidad, mayor densidad) presenta las temáticas desarrolladas, la temáticas de referencia.
- El segundo cuadrante (menor centralidad, mayor densidad) presenta las temáticas desarrolladas con poca influencia global.
- El tercer cuadrante (menor centralidad, menor densidad) presenta las nuevas temáticas, las temáticas nacientes.
- El cuarto cuadrante (mayor centralidad, menor densidad) presenta las temáticas en desarrollo, las temáticas puente, las temáticas prometedoras.

Los tipos de datos

Los tipos de datos de la encuesta deben ser cuidadosamente definidos antes de diseñar la encuesta y deben corresponder a los propósitos de la investigación. En el caso de la encuesta “Redes Colombia”, se busca en general describir la situación de la diáspora identificada de investigadores colombianos expatriados, en términos de :

Datos sociodemográficos

Edad, sexo, nacionalidad (es) del encuestado y sus familiares cercanos, escolaridad, actividad actual.

Datos biográficos

Trayectoria residencial
Trayectoria académica
Trayectoria laboral.

Datos de campos de investigación

Campos y palabras clave que describen los contenidos de las investigaciones.

Datos factuales

Tipo de relaciones activas que se tienen con Colombia
Tipo de entidad en la que se trabaja
Medios de comunicación que se utilizan
Pertenencia a redes de asociación de colombianos.

Datos de opinión

Nivel de satisfacción y expectativas frente al trabajo actual
Beneficios y aportes esperados respecto a la red Caldas
Condiciones de favorabilidad y dificultad para establecer relaciones activas con Colombia
Aciertos y problemas de la red Caldas, evolución de la ciencia y tecnología en Colombia (preguntas abiertas).

Datos de publicaciones

Referencias de las publicaciones.

La metodología

La figura 1. muestra el proceso metodológico que se ha seguido en la encuesta "redes Colombia". La metodología es aplicable a cualquier estudio de este tipo. Cada uno de los pasos se describe a continuación.

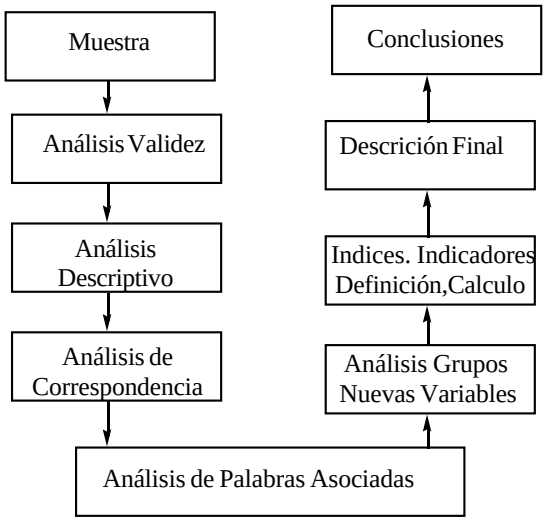


Figura 1. Esquema de la metodología para el análisis estadístico de migraciones científicas y profesionales

La muestra

En un tipo de encuesta como la encuesta "Redes Colombia", es muy improbable que pueda hacerse un muestreo siguiendo alguno de los modelos de muestreo tradicionales por las siguientes razones :

No se tiene una marco muestral completo.
No es posible entrevistar directamente a los encuestados.
No se tiene certeza de ubicar a cada individuo, debido a su movilidad.
La probabilidad de respuesta no es la misma en cada país.
El objeto de estudio es muy dinámico.

Las encuestas para el estudio de las migraciones científicas y profesionales deben ser enviadas a diferentes partes del mundo. La fuente primaria de información para ubicar a los individuos a ser estudiados se encuentra en los archivos gubernamentales los cuales son por lo general incompletos, y no están siempre actualizados. Entonces nuevas fuentes deben ubicarse, de tal manera que sea posible encontrar otros individuos y así aumentar el marco. Para la encuesta "Redes Colombia", se partió de la base de datos de Colciencias, entidad que orienta la políticas de ciencia y tecnología en Colombia, en la cual se encontraban registrados y totalmente identificados 826 investigadores y profesionales colombianos en exterior, en el año de 1994. Nuevas personas fueron ubicadas gracias a información recibida de algunos encuestados, y otros a partir de listas construidas por coordinadores de la red Caldas en los diferentes países.

La metodología para la obtención de la muestra, debe tener en cuenta, las dificultades mencionadas arriba, con el objeto de garantizar hasta donde sea posible la confiabilidad de los resultados obtenidos. Los pasos que se proponen son :

Decidir el número de formularios a enviar

Con la información disponible sobre ubicaciones de las personas, se debe decidir, de acuerdo al presupuesto disponible, cuál será el número de formularios enviados, y cuál la forma de envío. Debe utilizarse toda la información disponible, con el objeto de obtener una muestra que sea realmente significativa de la población estudiada.

Una estimación previa de algunos parámetros resulta de utilidad para la toma de decisiones antes del envío y después de la recepción de los formularios. La hipótesis que puede manejarse por ejemplo es que la base datos gubernamental constituye una buena muestra de la población estudiada. Si esto se asume, entonces es posible hacer para el caso de las migraciones científicas y profesionales las siguientes estimaciones :

Se considera la población distribuida geográficamente, por lo que se puede suponer un modelo multinomial, en donde cada categoría es una país receptor. Entonces es posible estimar la proporción de individuos en cada país. Nótese, que no es posible establecer intervalos de confianza, debido a que se desconoce el tamaño de la población.
Si se tiene alguna información adicional, como por ejemplo el área de trabajo, investigación, escolaridad, etc. puede usarse adicionalmente. Por ejemplo en el caso de estudio "Redes Colombia", se disponía de la clasificación a partir de los programas de investigación de Colciencias. Esta información adicional permite estimar parámetros de interés para el estudio. Para el caso de las migraciones científicas es más importante que la muestra sea muy representativa de las área científicas, aunque se aleje un poco (aunque no mucho) de la distribución geográfica.

Recibir la encuestas y almacenar los datos

Toda la información proveniente de las encuestas debe almacenarse en la base de datos construida para tal fin. La figura 2. Presenta el menú principal de la base de datos de la encuesta "Redes Colombia".

Análisis de la validez y representatividad de la encuesta

Es el primer paso importante del análisis estadístico y no puede ser omitido. Del resultado de éste análisis dependerá si debe tomarse una submuestra o dar pesos diferentes a algunos individuos. Además debe indicar si las conclusiones son realmente aplicables a toda la población bajo estudio. Dentro del estudio de la encuesta "Redes Colombia" se hicieron las pruebas de bondad de ajuste chi-cuadrado que se describen a continuación.

BASE DE DATOS

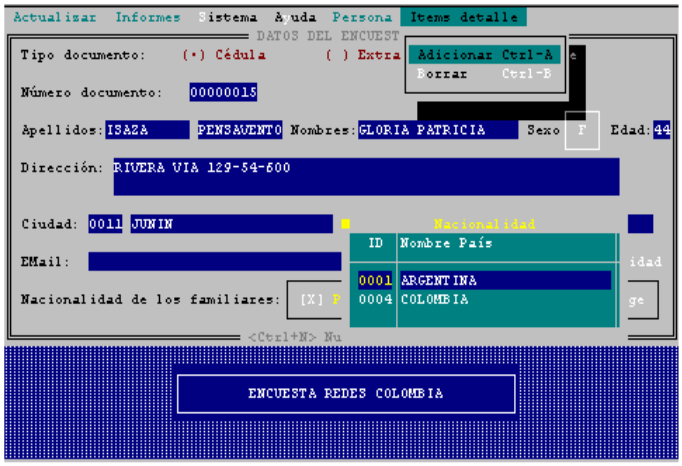


Figura 2. Menú principal de encuesta “Redes Colombia”

Se comparó la distribución de envio por países y la distribución de recepción. La distribución de la muestra se ajusta a la de envío. La figura 3 muestra la comparación de distribuciones.

Se comparó la distribución por países de la base de datos de Colciencias y la submuestra resultante de las encuestas obtenidas a partir de la base de datos de Colciencias. La submuestra ajustó perfectamente. La figura 3 muestra la comparación.

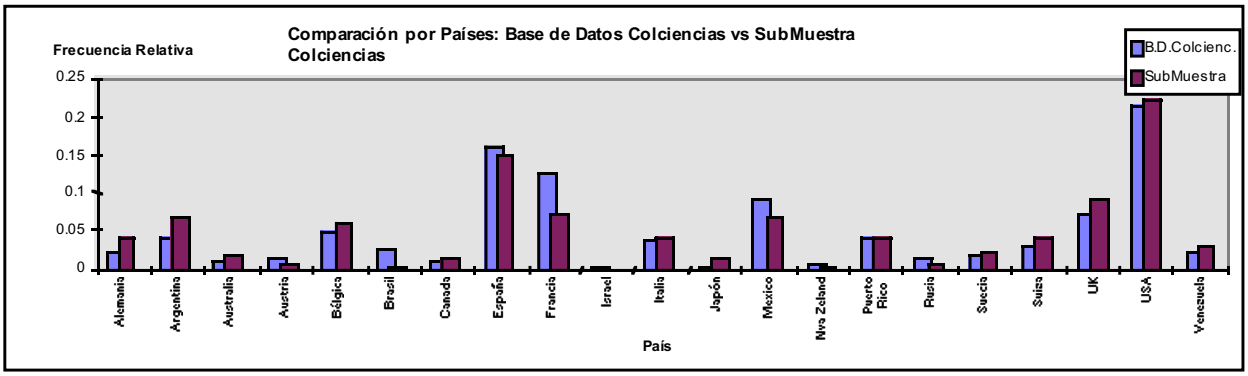
Se comparó la distribución por programas de investigación de la submuestra resultante de las encuestas obtenidas a partir de la base de datos de Colciencias, con los parámetros de Colciencias con resultado positivo.

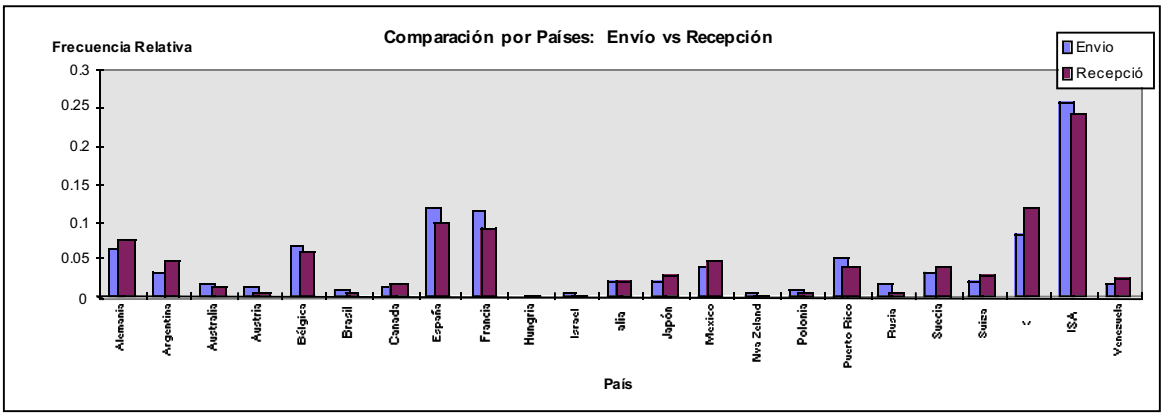
Se verificó la cuasi-independencia de respuesta dentro de cada país, comparando la distribución, por programas de investigación para la muestra y submuestra respectivamente. En el caso de la muestra total solo en dos países, los parámetros no ajustaron bien. En estos caso tenemos mas individuos ubicados, provenientes de fuentes diferentes a Colciencias.

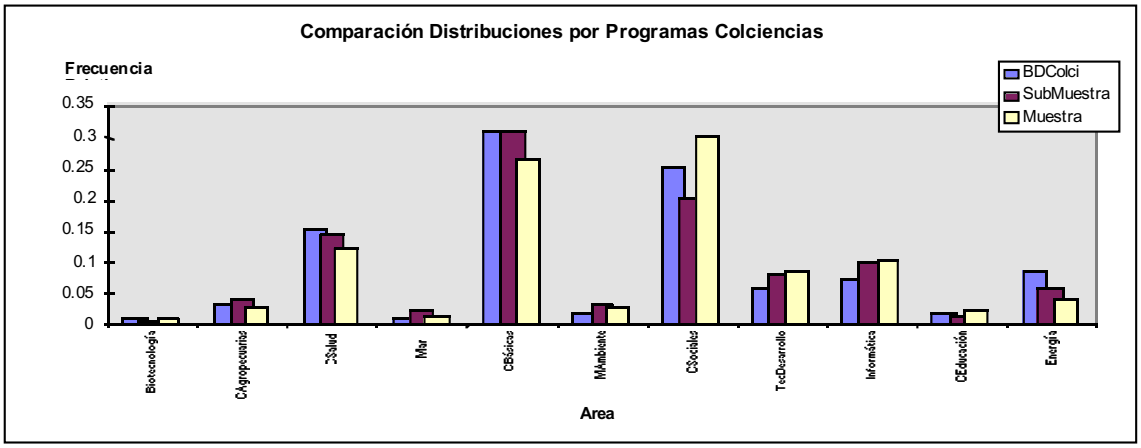
La conclusión es que la muestra es muy representativa de la población estudiada, y como el estudio es principalmente descriptivo (exploratorio) es posible continuar con el siguiente paso del análisis sin ninguna modificación. La figura 3. muestra la comparación de las distribuciones por programas de investigación de la muestra total y de la base de datos de Colciencias de 1994, en la encuesta “Redes Colombia”, en general, y la comparación interna dentro de Estados Unidos. Finalmente es necesario entender que los resultados obtenidos deben ser interpretados como tendencias presentes en la población y no como parámetros.

Primer análisis descriptivo

Este primer paso es necesario para tener una primera impresión de las características globales de la población así sea de forma aislada. Algunas primeras conclusiones pueden obtenerse desde ya. Por ejemplo es posible decir que la población examinada en la encuesta “Redes Colombia” esta conformada en 69 % por hombres, que la mayor parte de la población distribuye su tiempo entre el trabajo y la investigación (35 %) y entre el estudio y la investigación (25 %). Que el 87 % de la población encuestada tiene que ver con procesos de investigación, ya sea como estudiantes de maestría y doctorado o como investigadores de profesión, que el 83 % tiene un nivel de escolaridad entre maestría y doctorado, que las migraciones laborales mas frecuentes ocurren entre los países de Estados Unidos, Francia, España, Inglaterra, México, Alemania, Argentina, Brasil,







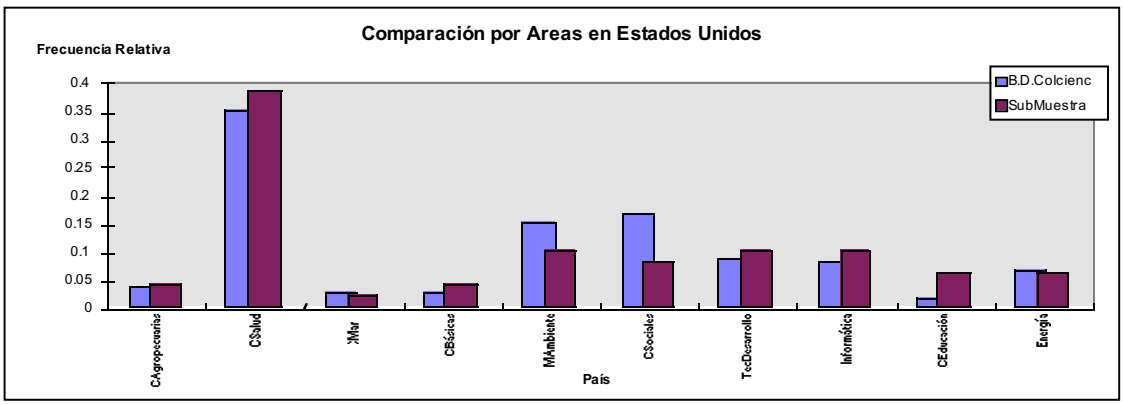
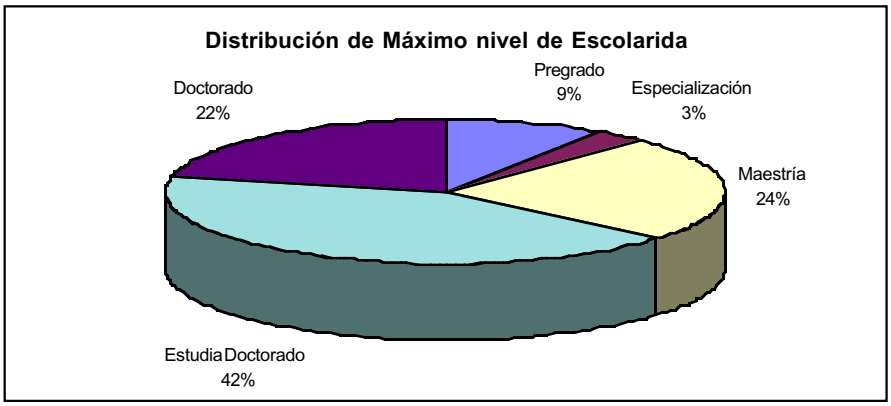


Figura 3. Análisis de la muestra de la encuesta “Redes Colombia”. Comparación de distribuciones

Venezuela, y obviamente Colombia. Que la mayor parte de las personas migran por primera vez por razones escolares, etc.... Esta información de hecho es útil para algunas tomas de decisión. De hecho puede ser información estratégica para algunas entidades. Para nosotros constituye en realidad el primer contacto con los datos. La figura 4 muestra dos ejemplos.



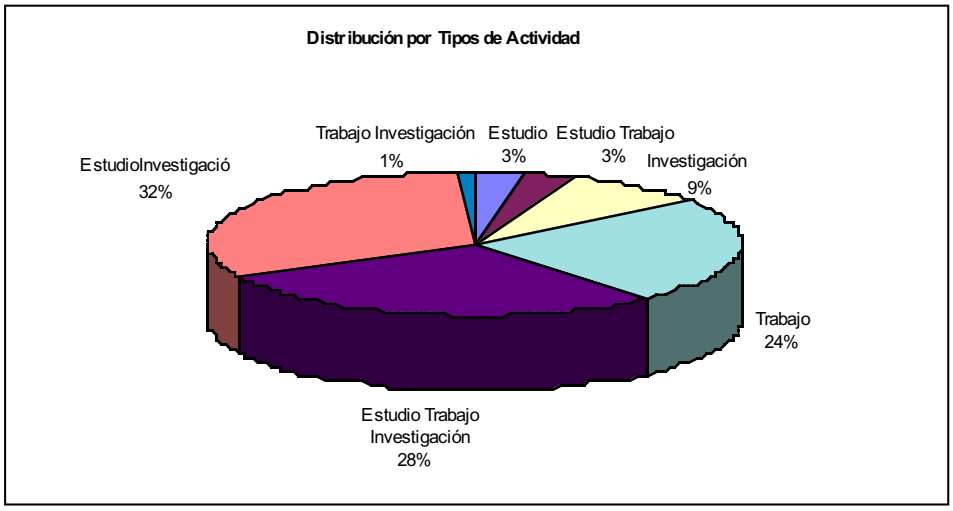


Figura 4. Ejemplos de análisis descriptivos

Primeros análisis de datos por temas (Primera clasificación de los datos), algunos ejemplos

Este paso es de los más complejos. Es el paso en donde será necesario hacer recodificaciones, reportes de diferente forma, etc.

El propósito es avanzar por temas en el análisis. Cada tema identificado es tratado en forma separada, No es mejor análisis aquel que se hace juntando todas las variables, entre otras cosas, por que cada tema por sí mismo requiere de atención y posiblemente tratamiento diferenciado. En el caso de la encuesta “Redes Colombia”, se llevaron a cabo entre otros algunos de los análisis cuyos resultados y aspectos más importantes se anotan a continuación. Se han escogido solo unos pocos ejemplos.

Análisis sociodemográfico

Se incluyeron las variables sociodemográficas, con la idea de establecer una clasificación de la población desde este punto de vista. El resultado obtenido es que toda la población es muy homogénea y con la excepción de los lazos familiares que ligan a algunos de los encuestado con el país de residencia, las demás características son comunes a la mayoría de la población.

Relaciones activas con Colombia

Se utilizaron 14 variables, con un total de 10 categorías. Las variables se sometieron a un ACM, cuyo primer plano factorial es mostrado en la figura 5. El análisis parece mostrar la existencia de cuatro grandes grupos. Unos que mantienen relaciones formales, otros que mantienen relaciones temporales, y otros que no mantienen relaciones con Colombia, y una clase de no respuesta. El análisis de clases confirma esta sospecha y además la proyección de las variables sociodemográficas ayuda a encontrar algunas características distintivas adicionales de cada grupo. En el próximo paso cada uno de las clases debe ser analizado a fondo. Por lo pronto ha bastado con identificar los grupos y tener una primera descripción de cada uno. El paso final en esta parte es crear una nueva variable, la variable Relaciones Activas, con cuatro categorías, cada una asociada a uno de las clases resultantes. La figura 6 presenta la descripción de uno de las clases. Ambas salidas provienen del programa estadístico, luego del ACM y de la clasificación respectivamente.

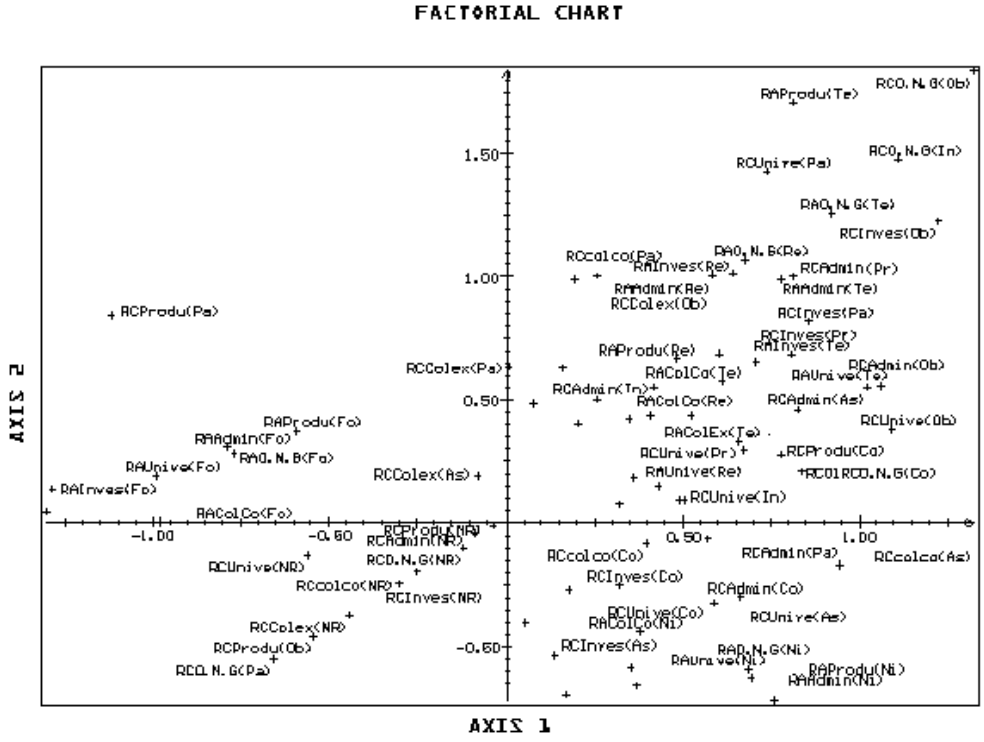


Figure 5. Plano factorial del tema relaciones activas s con Colombia (14) Variables

GLOBAL CHARACTERISTIC					
V.TEST	PROB.	PERCENTS	CLAS/MOD	MOD/CLAS.	MODALITY
8.54	.000	66.67	46.67	9.27	RAColCo(Re)
8.23	.000	84.00	35.00	5.52	RAInves(Re)
7.86	.000	51.67	51.67	13.25	RCUnive(Pr)
7.45	.000	52.83	46.67	11.70	RAUnive(Re)
6.43	.000	68.00	28.33	5.52	RAAdmin(Re)
5.92	.000	36.71	48.33	17.44	RAColEx(Re)
5.10	.000	47.22	28.33	7.95	RCInves(Pr)
4.64	.000	41.46	28.33	9.05	Rccolco(Pr)
4.33	.000	72.73	13.33	2.43	RAO.N.G(Re)
3.86	.000	23.66	51.67	28.92	RCColex(Pr)
2.96	.002	35.71	16.67	6.18	fisica
2.39	.009	45.45	8.33	2.43	RCAdmin(In)
- 2.41	.008	8.96	31.67	46.80	RAAdmin(Fo)
- 2.43	.008	11.66	78.33	88.96	RCAdmin(NR)
- 2.53	.006	2.04	1.67	10.82	pregrado

Figura 6. Descripción de la clase 2 del tema relaciones activas con Colombia La lista superior son las características dominantes y la lista inferior son las características mas alejadas

Análisis de migración residencial

Técnicas recientes como el análisis armónico cualitativo han sido desarrolladas aunque no implementadas completamente en un programa estadístico, para el análisis de trayectorias migratorias. Dentro del análisis de la encuesta "Redes Colombia" se ha utilizado con éxito una estrategia basada en el análisis de datos textuales, la cual describimos a continuación por constituir un aporte metodológico para este tipo de problemas. Para el estudio de datos longitudinales, la estrategia "tradicional" que se ha seguido es la siguiente : primero se discretiza el tiempo, definiendo periodos (que pueden ser históricos o de edad de la población encuestada), y agrupando los sitios de destino de acuerdo a algún criterio. En seguida se cruzan las dos variables tiempo discretizado por grupos destino obteniendo una variable de estado en donde cada modalidad es un espacio temporal por una espacio geográfico. A cada categoría se asocia el porcentaje de permanencia en tal estado para cada individuo, obteniendose una matriz de individuos por estados, la cual es sometida al análisis de correspondencia, y luego a la clasificación. Las técnicas de interpretación habituales son entonces utilizadas.

La técnica de análisis de datos textuales ha sido utilizada en algunos estudios aunque según nuestro criterio en forma errónea. Lo que nosotros hemos hecho es recodificar los datos como se muestra en la figura 7. Es decir hemos supuesto para cada individuo una respuesta textual de la forma

País1 Area_Continental_1 pais1_edad1 pais1_edad2...
País2 Area_Continental_2 pasi2_edad1 pais2_edad_2...

Datos originales I.D.	País	Fecha	Fecha	Edad
0001 0001 0002	Alemania Brasil Argentina	1980 1986 1990	1985 1989 1994	34 34 34
Datos codificados				
——0001 COLOMBIA COLOMBIA16 COLOMBIA17 COLOMBIA18 COLOMBIA19 ALEMANIA COM_EUROPEA ALEMANIA20 ALEMANIA13 ALEMANIA22 ALEMANIA 23 BRASIL SUR_AMÉRICA BRASIL26 BRASIL27 BRASIL28 BRASIL29 ARGENTINA SUR_AMÉRICA ARGENTINA30 ARGENTINA31 ARGENTINA32 ——0002				

Figura 7. Recodificación de los datos de migración residencial para los análisis textuales
Archivo de historia residencial

Es decir hemos escrito toda la historia residencial año por año en los países originales de migración. Cada línea corresponde a una estadía. El objetivo de colocar el nombre del país y el área continental es el de acercar trayectorias vecinas, es decir hemos introducido un efecto de vecindad. A este archivo se aplica el análisis de correspondencia desde el punto de vista del análisis textual, luego la clasificación y las técnicas del interpretación usuales pueden emplearse.

Aquí dos individuos se parecen porque usan aproximadamente el mismo vocabulario, lo que se traduce en término migratorios como que siguen aproximadamente las misma trayectoria. Los resultados obtenidos que se muestran en las figuras 7, 8 y 9 demuestran las bondades del método desde el punto de vista de la interpretación. Nótese en particular que en el caso de estudio de la encuesta "Redes Colombia", la migración se discrimina principalmente por los países receptores y no por una alta movilidad, como ocurre en las migraciones urbanas.

Análisis de biografías laborales y académicas

Se usado una mezcla de la técnica anterior y de la próxima. Por problemas de tiempo no alcanzamos a describir en detalle.

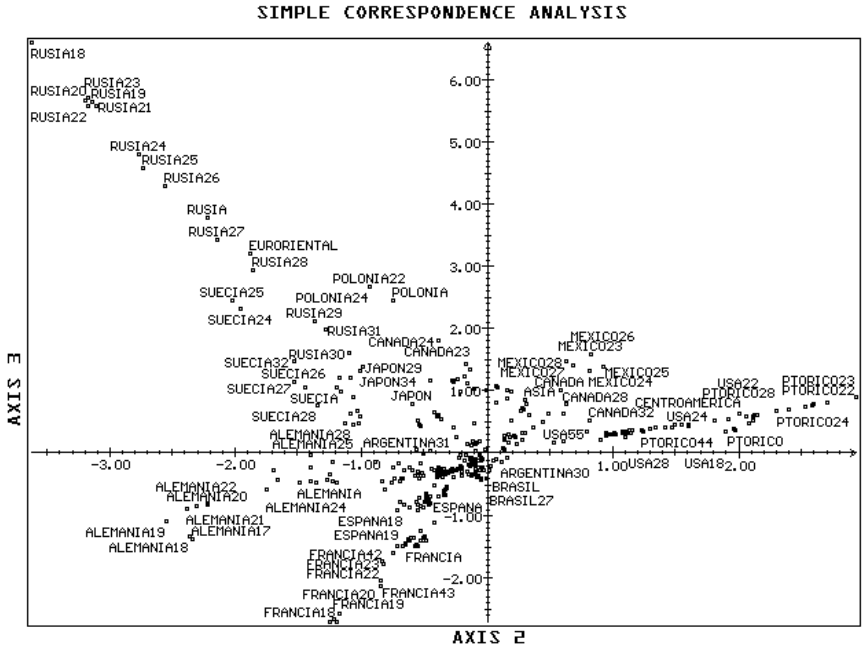


Figura 8. Plano factorial ejemplo de salida de análisis residencial

---

CRITERIO DE CLASIFICACÓN — RESPUESTA O INDIVIDUO CARACTERÍSTICO

6.000 -- 1 COLOMBIA16 COLOMBIA17 COLOMBIA18 COLOMBIA19 COLOMBIA20
-- COLOMBIA21 COLOMBIA22
-- USA NORTAMER, USA23 USA24 USA25 USA26 USA27 USA28 USA29 USA30
-- USA31 USA32 USA33 USA34 USA35
6.000 -- 2 COLOMBIA16 COLOMBIA17 COLOMBIA18 COLOMBIA19 COLOMBIA20
-- COLOMBIA21 COLOMBIA22
-- USA NORTAMER, USA23 USA24 USA25 USA26 USA27 USA28 USA29 USA30
-- USA31 USA32 USA33 USA34 USA35
5.801 -- 3 COLOMBIA16 COLOMBIA17 COLOMBIA18 COLOMBIA19 COLOMBIA20
-- COLOMBIA21
-- USA NORTAMER, USA22 USA23 USA24 USA25 USA26 USA27 USA28 USA29
-- USA30 USA31 USA32 USA33
5.801 -- 4 COLOMBIA16 COLOMBIA17 COLOMBIA18 COLOMBIA19 COLOMBIA20
-- COLOMBIA21
-- USA NORTAMER, USA22 USA23 USA24 USA25 USA26 USA27 USA28 USA29
-- USA30 USA31 USA32 USA33
2.747 -- 1 COLOMBIA16 COLOMBIA17 COLOMBIA18 COLOMBIA19 COLOMBIA20
-- COLOMBIA21 COLOMBIA22 COLOMBIA23 COLOMBIA24 COLOMBIA25
-- ARGENTINA SURAMERICA, ARGENTINA26 ARGENTINA27 ARGENTINA28
-- ARGENTINA29 ARGENTINA30 ARGENTINA31 ARGENTINA32 ARGENTINA33
-- ARGENTINA34 ARGENTINA35 ARGENTINA36 ARGENTINA37 ARGENTINA38
-- ARGENTINA39
2.533 -- 2 COLOMBIA16 COLOMBIA17 COLOMBIA18 COLOMBIA19 COLOMBIA20
-- COLOMBIA21 COLOMBIA22 COLOMBIA23 COLOMBIA24 COLOMBIA25 COLOMBIA26
-- COLOMBIA27 COLOMBIA28
-- ARGENTINA SURAMERICA, ARGENTINA29 ARGENTINA30 ARGENTINA31
-- ARGENTINA32 ARGENTINA33 ARGENTINA34 ARGENTINA35 ARGENTINA36
-- ARGENTINA37 ARGENTINA38 ARGENTINA39 ARGENTINA40 ARGENTINA41
-- ARGENTINA42 ARGENTINA43
2.504 -- 3 COLOMBIA16 COLOMBIA17 COLOMBIA18 COLOMBIA19 COLOMBIA20
-- COLOMBIA21 COLOMBIA22 COLOMBIA23 COLOMBIA24 COLOMBIA25 COLOMBIA26
-- COLOMBIA27 COLOMBIA28
-- ARGENTINA SURAMERICA, ARGENTINA29 ARGENTINA30 ARGENTINA31
-- ARGENTINA32 ARGENTINA33 ARGENTINA34 ARGENTINA35 ARGENTINA36
ARGENTINA37 ARGENTINA38 ARGENTINA39 ARGENTINA40 ARGENTINA41

Figura 9. Algunas respuestas características en dos clases

Análisis de campos de investigación

La figura 10 muestra la fuente original de los datos recolectados en la encuesta “Redes Colombia”. El propósito de esta parte del estudio es establecer los campos en los cuales los investigadores se encuentran trabajando. La técnica que empleada consiste en agregar información implícita presente en la respuestas. Para ello utilizamos el sistema de clasificación de la base Pascal. Concretamente los que hicimos fue agregar a la respuesta los tres primeros campos de la base Pascal correspondientes al campo de investigación respondido. Las palabras clave se dejaron tal como llegaron. La figura 10 muestra como se preparó cada respuesta para el análisis textual (análisis de correspondencia) y para el análisis de contenido (palabras asociadas). Las figuras 11, 12 y 13 muestra un ejemplo de salida del análisis textual, y las figuras 14 y 15 muestran las mismas salidas por el método de palabras asociadas.

Campos de Investigación

0001	CSocioEconomica	CJuridicas		CIENCIAS_POLITICAS
0003	CSocioEconomica	Economia General	EconInternacional	RELACION_INTERNAL SUDASIA
0004	CMedicaBiologia	Biologia	AnimalProducc	APICULTURA APITOXINA
0005	CSocioEconomica	EconoEnergia	NuclearEnerg	TECNOLOGIA_INDUSTR TECNOLOGIA_AVANZADA
0006	CMedicaBiologia	Biologia	VertebAnatSico	MEDICINA_VETERINARIA

Palabras clave

0001	AMERICA_LATINA
0001	ASIA__PACIFICO
0001	ECONOMIA_INTEGRACION
0003	CULTURA_IDENTIDAD
0003	ECONOMICO_DESARR
0003	POLITICA
0003	POLITICA_PROCESO
0003	RELACION_INTERNAL
0004	APICULTURA
0004	APITOXINA_EXTRACCION
0004	COLMENA
0004	STRESS
0005	ELECTROMEC_CONTROL
0005	ROBOTICA
0005	TECNOL_INTEGRACION
0005	TECNOLOGIA_BLANDA
0006	ANIMAL
0006	COMPORT_ANORMAL
0006	RESULTADOS_INFLUENC
0006	STRESS

Archivo para análisis

——0001
0001 CSocioEconomica CJuridicas CIENCIAS_POLITICAS
AMERICA_LATINA ASIA_PACIFICO ECONOMIA_INTEGRACION.
——0003
0003 CSocioEconomica EconomiaGeneral EconInternacional RELACION_INTERNAL SUDASIA
CULTURA_IDENTIDAD ECONOMICO_DESARR POLITICA POLITICA_PROCESO
RELACION_INTERNAL.
——0004
0004 CMedicaBiologia Biologia Animal Producc APICULTURA APITOXINA
APICULTURA APITOXINA_EXTRACCION COLMENA STRESS.
——0005
0005 CSocioEconomica EconoEnergia NuclearEnerg TECNOLOGIA_INDUSTR
TECNOLOGIA_AVANZADA
ELECTROMEC_CONTROL ROBOTICA TECNOL_INTEGRACION
TECNOLOGIA_BLANDA.
——0006
0006 CMedicaBiologia Biologia VertebAnatSico MEDICINA_VETERINARIA
ANIMAL COMPORT_ANORMAL RESULTADOS_INFLUENC STRESS

Figura 10. Recodificación de los datos para el análisis de campos de investigación

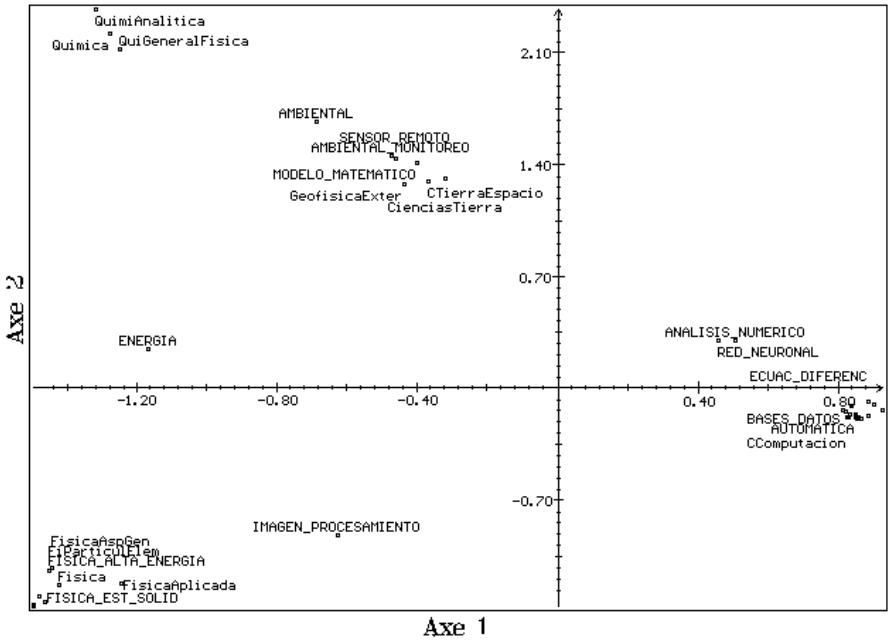


Figura 11. Ejemplo de plano factorial en el análisis del tema campos de investigación. Grupo Ciencias Exáctas y tecnológicas

TEXTE NUMERO 6 a06a = CLASSE 6 / 13

LIBELLE DE LA FORME GRAPHIQUE	---POURCENTAGE--- INTERNE	GLOBAL	INTERNE
1 BASES	14.29	.53	4.
2 DATOS	14.29	.80	4.
3 BASES_DATOS	10.71	.40	3.
4 MULTIMEDIA	10.71	.53	3.
5 CACienciasCompu	14.29	4.38	4.
6 MODELO	3.57	.53	1.
7 INFORMATICA	3.57	.53	1.
8 CienAplicadas	14.29	9.96	4.
2 Fisica	.00	3.19	0.
1 CExactasTecnologia	14.29	18.59	4.

Figura 12. Descripción interna de una clase de campo de investigación

TEXTE NUMERO 6 a06a = CLASSE 6 / 13

CRITERE DE CLASSIFICATION	REPONSE OU INDIVIDU CARACTERISTIQUE
2.168 --	1 CExactasTecnologia CienAplicadas
--	BASES_DATOS MULTIMEDIA
--	ANIMACION, BASES, DATOS, MEMORIA, MULTIMEDIA,
2.078 --	2 CExactasTecnologia CienAplicadas
--	BASES_DATOS
--	BASES, DATOS, PROTOTIPO, .
1.302 --	3 CExactasTecnologia CienAplicadas
--	BASES_DATOS
--	DEDUCTIVOS, DEPENDENCIA, MULTIVARIADA, AXIOMAS
--	INFERENCIA, BASES, DATOS, MODELO, .
1.273 --	4 CExactasTecnologia CienAplicadas
--	INFORMATICA

Figura 13. Algunas respuestas características de una clase de investigación

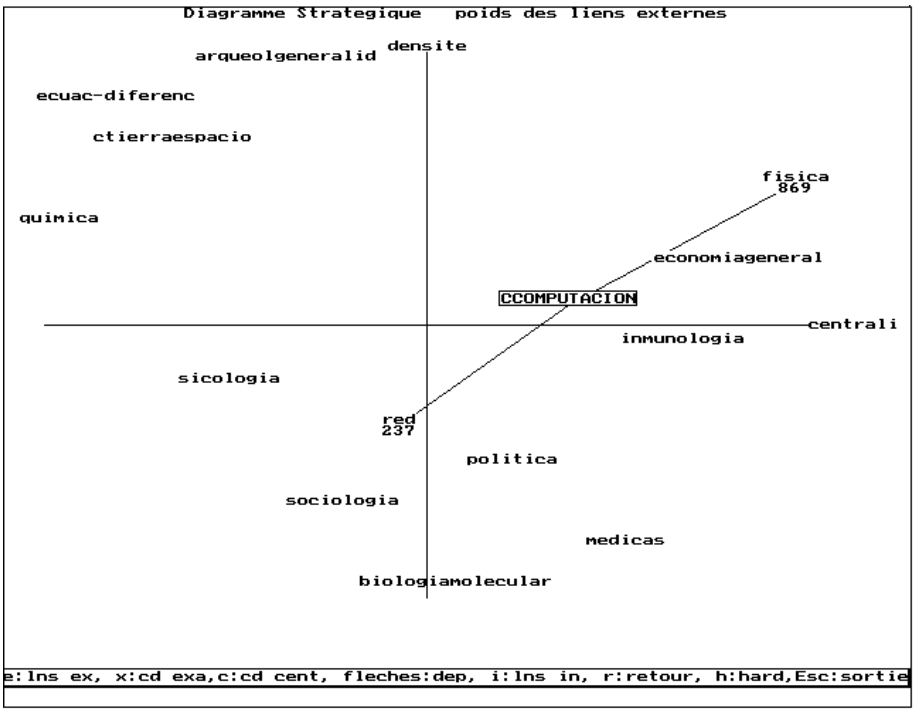


Figura 14. Ejemplo del método de palabras asociadas para el tema de investigación

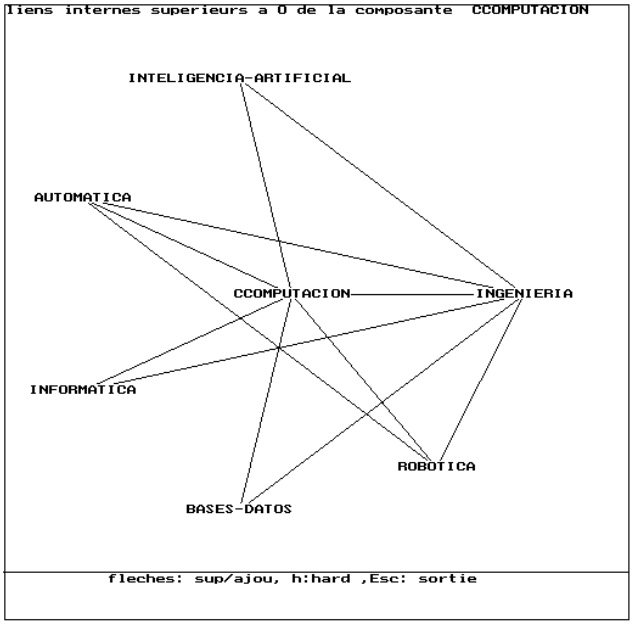


Figura 15. Ejemplo del método de palabras asociadas para el tema de investigación

Análisis Interno de las clases

Descubrir la composición de la población por temas no es suficiente para el investigador. El siguiente paso es hacer la descripción interna de cada clase. Desde el análisis de clases ya se han descubierto posiblemente algunos rasgos distintivos de cada uno, pero solo un análisis descriptivo interno puede revelar la características de cada grupo. El gráfico de la figura 16 por ejemplo muestra como puede estudiarse dentro de un clase de migración como es el movi-

miento interno, es decir como se tipifican internamente los individuos pertenecientes a cada clase. Otros análisis descriptivos revelaran por ejemplo como en donde están ubicados, que hacen, cual es su pensamiento de retorno al país etc., de por ejemplo los investigadores

El análisis final

El propósito final es hacer una descripción macro de la población. La propuesta es crear una variable por cada tema investigado, en donde cada categoría es una clase. Como cada clase ha sido ampliamente estudiado, es claro lo que significa pertenecer a uno cualquiera de ellos. Lo que nos proponemos es dar una descripción en términos de estas clases. Una análisis de correspondencia múltiple se aplicará, y su interpretará posiblemente la estructura general de la población estudiada.

Conclusión

Una de los grandes vacíos que este grupo de investigación ha encontrado en la revisión de trabajos en el área de las migraciones científicos es la no utilización de las técnicas modernas de análisis estadístico, fundamentalmente provenientes de la escuela francesa. Quizá la ausencia de una bibliografía adecuada que sin abandonar un tratamiento matemático mínimo trate de las aplicaciones concretas en el estudio de las ciencias sociales sea una de las principales causas. El proyecto de estudio “El brain gain revisited a través de caso colombiano. Estudio de la red Caldas”, del cual hace parte la encuesta “Redes Colombia” ha contado con un equipo multidisciplinario incluidos estadísticos que ha permitido avanzar de manera significativa en la utilización de tales herramientas en éste tipo de estudios, llegando incluso a presentar propuestas metodológicas. El propósito de estas notas ha sido presentar un avance de la metodología que con éxito ha sido aplicada en el proyecto.

Bibliografía

Barbary, O. (1994). Análisis de Datos Biográficos. Simposio de Estadística, Departamento de Matemáticas y Estadística, Universidad Nacional de Colombia, Santafé de Bogotá.

Bécue, M. (1991). Análisis de Datos Textuales. Métodos Estadísticos y Algoritmos. CISIA, París.

Benzecri, J. P. (1976). L’Analyse des Données, Tome 1 : La Taxonomie, Tome 2 : L’Analyse des Correspondence. Dunod, París (2da ed. 1976).

Centre de la Sociologie de L´Innovation (1994). LEXIMAPPE - DOC, INIST, Paris.

Courtial, J. P. (1994). Science Cognitive et sociologie des sciences, Presses Universitaires de France, Colection sous la direction de J. P. Courtial, Paris.

Crivisqui, E. (1993). Análisis factorial de correspondencias un instrumento de investigación en ciencias sociales. Laboratorio de Informática Social Universidad Católica de Asunción, Asunción.

Hausler L. (1993). Des phrases et des itinéraires, en « Actes des secondes journées internationales d’analyse statistique de données textuelles », Montpellier 21-22 de octubre de 1993, Ecole Nationale Supérieure des Télécommunications, Paris.

Houzel Y. y LE Vaillant M : [1994] : Analyse statistique de données textuelles et traitement des données de calendriers : application à l'analyse de l'insertion professionnelle des élèves issus des écoles d'art, en « Actes des journées CEJEE-CEREQ sur les données longitudinales dans l'analyse du marché du travail », Toulouse, octubre de 1993, Ecole Nationale Supérieure des Télécommunications, Paris.

Lebart, L., Morineau, A., and Warwick, K. (1984). Multivariate Descriptive Statistical Analysis. Correspondence Analysis and Related Techniques for Large Matrices. John Wiley & Sons, USA.

Lebart, L., Morineau, A., and Fenelon J. P. (1979) Traitement des Données Statistiques, Méthodes et Programmes. Dunod, París.

Lebart, L., Salem A. (1994). Statistique Textuelle. Dunod, París.

Montenegro A. (1996). Encuesta Redes Colombia. Memorias del Primer Coloquio sobre Ciencia Tecnología y Cultura, Academia Colombiana de Ciencias Exactas Físicas y Naturales, Colección Memorias, No.6. Pags 101-114, Santafé de Bogotá.

Pardo, C.E. (1992). Análisis de la Aplicación del Método de Ward de Clasificación al caso de Variables Cualitativas. Tesis para optar al título de M. Sc. Estadística. Universidad Nacional de Colombia. Santafé de Bogotá.

Montenegro A., Pardo C. E. (1996), Introducción al Análisis Estadístico de Datos Textuales, Departamento de Matemática y Estadística, Unidad de Extensión, Universidad Nacional de Colombia, Bogotá.

Montenegro A., Pardo C. E. (1996), Los itinerarios individuales interpretados como frases, una aplicación de la estadística textual a la tipología de trayectorias, Memorias del Seminario Internacional de capacitación e investigación en Recolección y Análisis de datos longitudinales, Universidad Nacional de Colombia. Bogotá, diciembre de 1996.

[1] Este proyecto de investigación es desarrollado por convenio entre las instituciones Universidad Nacional de Colombia la Universidad del Valle, y el Institut Français de Recherche pour le Developpement en Coopération con el apoyo de Colciencias e Icfes.

[2] La red Caldas fue establecida por Colciencias con el propósito de ser la red colombiana de investigadores y profesionales en el exterior. El objetivo de Colciencias al crear la fue incentivar la cooperación activa de la comunidad científica y profesional colombiana radicada en el exterior en el trabajo especializado de los grupos de investigación nacionales, abriendo la posibilidad de que se vincularan a proyectos de investigación en ejecución o promoviendo su participación en otros nuevos, así como al proceso de reestructuración del país que inició el presidente Gaviria. La red hace parte de las estrategias especiales formuladas dentro del plan nacional de ciencia y tecnología formulado en el gobierno Gaviria, dentro del ítem "fomento a grupos y redes", conducente a la formación de una comunidad Científica.