UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

WP.8
English only

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Data Editing
(Madrid, Spain, 20-22 October 2003)

Topic (iii): Data editing processes within survey processing

TREATMENT OF THE ECONOMIC ACTIVITY AND THE OCCUPATION IN THE CENSUS OF POPULATION: SPANISH EXPERIENCE

Invited Paper

Submitted by the National Statistical Institute, Spain[1]

Abstract

The coding of the economic activity and occupation variables, which are gathered by self-filling out, with help from the census agent when needed, in order to an exhaustive exploitation (16 million occupied people), is a difficult task.

Noting that people find some difficulties in describing these two variables in a way that allows a right coding, the NSI of Spain decided to use a mixed method: To offer a list including the main industries and occupations, and in case of not finding in the list or not being totally satisfied with the list description, describe it in a free text box.

The result, after a previous analysis of the collected descriptions and their treatment by an automatic coding system, was that around the 95% of the occupied population was codified, being the rest processed by assisted coding and imputation.

This article describes the sequence of the works.

Key words: Census of population, text categorization, coding of industry, coding of occupation

1. Introduction

With reference to 1st November 2001, a population census was carried out in Spain, which was unusual for the following reasons:

a) Pre-printed on the questionnaires were the name and address of the household members.

b) The questionnaires were scanned after the agents completed and collected them.

Most questions were multiple-choice questions, providing a set of categories in which people could be classified. However, for two questions it should be possible to write a

[1] Prepared by Francisco Hernández Jiménez (fhernan@ine.es), Francisco Fernández Serra (franfer@ine.es), Ascensión Alvarez (aalvarez@ine.es) and Asunción Piñán Gaviria (apinan@ine.es).

free text, given their difficulty, and this would entail a problem in the subsequent treatment of the questions. We are referring to the variables occupation and economic activity for which approximately 16 million people had to give an answer. In a Spanish census, said variables had never been exhaustively processed at such broken down levels.

As the Recommendations from UNECE-EUROSTAT for the European round of 2000 Censuses of Population and Housing suggested, regarding economic activity, the national version of the three digit NACE Rev.1[2] (CNAE-93[3]) was applied, what meant its classification into 222 categories. The variable occupation was also classified at the three digit level of CNO-94[4] (national version of ISCO-88[5]), i.e. 206 categories.

2. Difficulty to define correctly occupation and economic activity

Information was collected directly from the questionnaires filled in by people included in the census, which meant that they had to be able to describe correctly in their own words both their occupation and their industry, thus allowing their coding at a three digit level.

First pilot survey

The first pilot survey showed that it is rather tricky for a person to adequately describe his situation answering only one question, for his coding at the required level. For example, a secondary education teacher may describe his activity as "education", "English teacher", "teacher", "secondary" … and in most cases this is not an accurate enough information for his appropriate classification.

The *first pilot survey* included a short list of economic activities (commerce, manufacture…) together with a free text to be completed by all the people. Regarding occupations, a greater number of descriptions were presented and those who could not find a satisfactory one, had the possibility to reply in their own words.

The result was that the question on economic activity brought about great confusion and was hard to answer correctly. On the contrary, regarding occupation, the respondents usually found it easier to answer and the quality of the information was better.

On the other hand, there seemed to be some confusion on the concepts economic activity and occupation.

Second pilot survey

The *second pilot survey* included longer lists of pre-codes for the variable economic activity, whereas those for occupation did not vary. This led to an easier filling in of the questionnaire and an improved quality of the answers.

3. Strategy for the treatment of the variables industry and occupation

After analysing the pilot surveys, the following was done:

[2] Statistical Classification of Economic Activities in the European Community
[3] Clasificación Nacional de Actividades Económicas 1993: Spanish classification of economic activities
[4] Clasificación Nacional de Ocupaciones 1994: Spanish classification of occupations
[5] International Standard Classification of Occupations

a) These variables would be researched upon by means of a question, allowing people to classify themselves with the help of a predefined list, or to describe themselves in their own words. Questions would be expressed as follows:

INDUSTRY	**What is the main activity of the establishment or venue where you worked?** Find it in the TABLE OF ACTIVITIES (on the white sheet with a red title) and note the corresponding number: If you could not find the activity or have any doubts, describe it below:
OCCUPATION	**What was your occupation?** **BEWARE:** We are **NOT** asking your degree (bachelor, doctor…) nor your occupational situation (civil servant, employer…) nor you labour category (skilled labourer, apprentice…) but the **type of work.** Find it in the TABLE OF OCCUPATIONS (on the white sheet with a yellow title) and note the corresponding number. If you could not find your occupation or have any doubts, describe it below.

b) The list of pre-codes did not go beyond one side of a DIN-A4. There was one list for economic activities and one for occupations (Annexe II shows an example of these lists).

c) In the list of economic activities pre-codes, the categories were identified by three numbers (same codes as those defined in CNAE-93).

d) In the list of occupations pre-codes, the categories were identified by a letter and a number.

The purpose of coding these two variables in a different way was to avoid as much as possible mistaking one for the other.

e) People who were entered in the census were given the possibility to define themselves in their own words if they are missing in the pre-codes lists.

f) The replies regarding the variables economic activity and occupation given in the own words of the respondents, would undergo a first treatment by automatic coding.

g) Non coded texts would undergo manual coding.

h) Non response would be imputed

4. Specification of pre-codes lists

On the basis of the data from the Spanish Labour Force Survey, an analysis was made of how the variables economic activity and occupation of the Spanish population were geographically distributed.

Economic activity

It was clear that the economic activities varied according to territories and that if a list for each province (Spain is divided in 52 provinces) was established, 80% of the

population could adequately be classified by means of pre-codes. Later, these lists were grouped into eight different types of lists, which were enough to cover the whole of Spain.

Two conditions were required to determine which activities would be part of the list:

- The activity had to be representative of the number of workers.

- The activity had to be difficult to describe for a correct coding.

The titles for each of these pre-codes should be as specific as possible, avoiding terms such as “other”, “similar” or “etc.”. If it was to be expected that any of the categories would be used frequently as an answer, a greater precision was sought in its wording, even to the point of dividing into two a given category of the classification.

Also, as far as possible, an attempt was made to include in the list the activities referred to the new economy.

Occupation

It had become clear that this variable was related to the variables level of education of the respondents, their socio-economic status and economic activity. These connections were taken into account for the elaboration of the list, in order to simplify the number of pre-codes to be considered. It was also a fact that for some occupations people found it hard to describe them, this being why special emphasis was laid on these ambiguous categories.

This simplification achieved that with 130 pre-codes, almost all the 206 categories of the classification were covered.

For the elaboration and distribution of the different types of lists of occupations, it was born in mind that there was a strong correlation between the educational profile of the household members and their occupations (a criterion that differs from the one used for the variable industry).

Therefore, depending on the family’s education (a variable derived from the Population Register), the appropriate list was sent to each family. Finally, four different lists were created.

Automatic coding

Automatic coding of the variables economic activity and occupation is part of the Text Categorization, defined in Sebastiani (2002) as “the task of assigning a Boolean value to each pair $(d_j,c_i) \in DxC$, where D is a domain of documents and $C=\{c_1,\ldots,c_{|C|}\}$ is a set of predefined categories. A value of T assigned to (d_j,c_i) indicates a decision to file d_j under c_i, while a value of F indicates a decision not to file d_j under c_i “.

The purpose of this type of categorization is that each response text for these variables be assigned a single code. Furthermore, the structure of the categories remains fixed, it being necessary to assign each text to one of these categories. Some characteristics of these texts make them special:

- the texts are usually made up of 2 or 3 words, those of more than 4 words being rare.

- the descriptions are usually very similar, since each sector has a jargon used by the people belonging to it

- the lists of pre-codes in some way had an influence on the language used for descriptions

a) Analysis of the vocabulary: elaboration of the initial corpora

For the elaboration of the corpora, account was taken of the labour force survey, which codes the variables economic activity and occupation at the same level that is required in the census.

The labour force survey is carried out by means of a laptop computer and a personal interview of a household member by the interviewer. The latter, besides coding the variables occupation and economic activity, completes a text with the description of said activity or occupation.

The analysis of these texts, after them being cleaned through the elimination of semantically irrelevant words (included in a preliminary list of 59 words such as "the", "of"…) and duplicated words, and by reducing the number of words with the help of synonyms, yielded the targeted corpora, which included:

	INDUSTRIES	**OCCUPATIONS**
Number of different words	7.452	4.223
Number of different texts	16.577	9.401
Number of synonyms	3.544	2.821

Although we thought that the corpora were rather limited, the pilot surveys showed that their coverage was acceptable. Anyway, recognising that the coverage could not be the best possible, it was decided, as will be pointed out below, that when a text contained more than 1/3 of the words not appearing in the corpus, it was directly sent to the coding queue.

These initial corpora were improved by *train-and-test approaches* (Sebastiani, 2002), that were made after analysing the experimental results (Mitchell, 1996). After the first and the second pilot surveys, the corpora were very slightly increased. Before coding the census, an analysis was made of the words used in the pilot surveys answers, not belonging to the vocabulary of the corpora. Those that were bad spelling or had print faults (very few), or were synonyms of already existing ones, were included in the corpus to be used.

On the other hand, before starting the final coding, once all the census replies were recorded, the most repeated texts were analysed and the conclusion was reached that many of them could not be coded correctly because they did not contain enough information. Although they were not included in the corpus of texts, they were identified through fictitious codes for their later treatment (see paragraph h).

b) Coding method

Some tests were made with the coding of data derived from the first and the second pilot surveys. The outcome was that **selecting the corpus texts that had as many as possible words in common with the text to be codified,** in most cases at least one text was liable to code the text correctly. Consequently, it was decided that the Nearest Neighbour (Winkler, 2003) method would be used, for which it was necessary to assign

a number of weights to the words and to define some indicators to measure the distance among the texts. The selected corpus texts are called *neighbour potential donors* (*NPDs*) in this document due to the used method.

c) Weightings to be used

An essential element of the automatic coding was to select the *NPDs* as close as possible to the text to be coded. To this end, it was important to chose the word with most discriminating power (*filter word*) among those making up the text to be coded. Two different weights, therefore, were assigned to the words in the corpus:

c.1) Weight of the word associated with the corpus

Two corpora exist, one made up of m texts (derived from surveys) and one made up of k words stemming from earlier texts. Each text t_i $(i=1..m)$ of the corpus of texts, made up of words belonging to the corpus of words, may be represented by an array $t_i=(X_{i1},\ldots,X_{ik})$, where:

$X_{ij}=1$ if the jth word of the corpus of words is included in the text t_i, and
$X_{ij}=0$ otherwise.

The first weight associated to each word of the corpus of words was defined as:

$$w_j = \sum_{i=1}^{m} X_{ij} \qquad \forall \quad j = 1 \ldots k$$

c.2) Weight of the word associated with the category

It was deemed desirable to give a weight to the words for their discriminating power in relation to the classification categories.

Let $c_1,\ldots,c_l$ be the categories of the classification. Taking into account that each text t_i is associated with a single category c_h, the words p_j that belong to text t_i can be associated with category c_h.

As in the above case, an array $c_h=(X_{h11},\ldots,X_{hij},\ldots,X_{hmk})$ may be defined, where:

$X_{hij}=1$ if the jth word of the corpus of words is included in text t_i and the latter is associated with category c_h, and
$X_{hij}=0$ otherwise.

With each word p_j a weight per category c_h is associated, which is defined by

$$w_{jh} = \sum_{i=1}^{m} X_{hij} \qquad \forall \quad j = 1 \ldots k, \quad h = 1 \ldots l$$

To give an index comparable among words, the index

$$wc_j = \max_h \frac{w_{jh}}{w_j} \qquad \forall \quad j = 1 \ldots k$$

is defined.

d) Selection of *NPDs*

In the first place, the text *u* to be coded is cleaned:

a) Words without semantic content are eliminated.
b) Duplicated words are eliminated.
c) Words are replaced by synonyms.

It is also assumed that words are independent and that their order has no influence (naïve Bayes).

In the first place the *filter word* of the text *u* to be coded is selected according to the system explained below:

From the words of the text to be coded *u* that are in the corpus, the above defined two weights *(w_j, wc_j)* are considered, the *filter word* being determined according to the following conditions:

1. From the words of the text to be coded *u* with $wc_j>0,3$ and $w_j>4$, the word of maximum w_j is selected as the *filter word*, the aim being the achievement of a *filter word* that is representative in some code and appears very frequently.

2. If the above condition is not fulfilled, we have one of the following cases:
   2.1. $wc_j£0,3$ y $w_j>4$
   2.2. $wc_j>0,3$ y $w_j£4$
   2.3. $wc_j£0,3$ y $w_j£4$

From any of these:

2.a. The word of maximum wj is selected, provided that wj <100.
2.b. If all the words have wj≥100, any of them is selected.

Once the *filter word* was selected, we obtained a first batch of *NPDs* made up of those texts of the corpus that contained the *filter word* in question. From this batch of texts, a selection was then made of those that had the greater amount of words in common with the text *u* to be coded.

e) Indicators for measuring the distance between a *NPD* and the text to be coded

After this process, a set of *z NPDs* has been selected. It is possible to define several different indicators. Two of them measure the distance from the *NPD* t_i *(i=1...z)* to the text *u* to be coded. The other two indicators are one measuring the distance between the text *u* to be coded and every category in the classification c_h *(h= 1…l)*, and another one measuring the distance between a pre-code and a code in the classification.

e.1) Indicators of the distance between texts

In order to measure the distance from a *NPD* t_i to the text to be coded *u*, three different weights are considered: the weight assigned to the *NPD* , the weight assigned to the text to be coded and the weight assigned to the words used in the *NPDs* selection.

The **weight assigned to the *NPD* t_i** can be defined as the addition of the weights assigned to the words of this text in the corpus of words (for a definition of the used values, see paragraph C.1):

$$w_{ti} = \sum_{j=1}^{k} w_j X_{ij} \qquad \forall \quad t_i \quad liable \quad neighbour$$

When considering the text to be coded *u*, the corresponding array can be defined as $u=(X_{u1},\ldots,X_{uk})$ where:

$X_{uj}=1$ if j^{th} word of the corpus of words is included in text t_i, and
$X_{uj}=0$ otherwise.

The **weight assigned to the text *u* to be coded** can be defined as:

$$w_u = \sum_{j=1}^{k} w_j X_{uj}$$

In order to assign a weight to the set of common words used in the selection, a new array is defined as $(Y_{1i},\ldots,Y_{\#(u)i})$, where:

$Y_{gi}=1$ if g^{th} word of the text *u* to be coded is included in the *NPD* t_i, and
$Y_{gi}=0$ otherwise.

The **weight assigned to the common words** between the text *u*[6] to be coded and the *NPD* t_i can be defined as:

$$w_A = \sum_{g=1}^{\#(u)} \sum_{j=1}^{k} w_j X_{ij} Y_{gi}$$

The **indicator of the distance between a *NPD* t_i and the text to be coded *u*** is defined as:

$$IND3 = \frac{w_A}{w_u} * \frac{w_A}{w_{ti}} = \frac{\left[\sum_{g=1}^{\#(u)} \sum_{j=1}^{k} w_j X_{ij} Y_{gi}\right]}{\left[\sum_{j=1}^{k} w_j X_{uj}\right]} * \frac{\left[\sum_{g=1}^{\#(u)} \sum_{j=1}^{k} w_j X_{ij} Y_{gi}\right]}{\left[\sum_{j=1}^{k} w_j X_{ij}\right]}$$

This indicator varies from 0 to 1. An indicator value closer to 1 represents a shorter distance between texts.

This indicator entails a problem generated by the words with a high frequency in the corpus, which have a great influence on the indicator value. In this sense, it is possible to define an analogous indicator for measuring the distance from one text to another, considering the weights assigned to the words of the corpus in relation to the category wc_i:

[6] The common words between the text *u* to be coded and the liable neighbour t_i are the same words for all the liable neighbours selected in each reiteration.

Let wc_A be the **weight assigned to the common words** between the *NPD* and the text to be coded:

$$wc_A = \sum_{g=1}^{\#(u)} \sum_{j=1}^{k} wc_j X_{ij} Y_{gi}$$

Let wc_{ti} be the **weight assigned to the *NPD* t_i**:

$$wc_{ti} = \sum_{j=1}^{k} wc_j X_{ij}$$

Let wc_u be the **weight assigned to the text to be coded *u***:

$$wc_u = \sum_{j=1}^{k} wc_j X_{uj}$$

The new **indicator of the distance between a *NPD* t_i and the text to be coded *u*** is defined as:

$$IND4 = \frac{wc_A}{wc_u} * \frac{wc_A}{wc_{ti}} = \frac{\left[\sum_{g=1}^{\#(u)} \sum_{j=1}^{k} wc_j X_{ij} Y_{gi}\right]}{\left[\sum_{j=1}^{k} wc_j X_{uj}\right]} * \frac{\left[\sum_{g=1}^{\#(u)} \sum_{j=1}^{k} wc_j X_{ij} Y_{gi}\right]}{\left[\sum_{j=1}^{k} wc_j X_{ij}\right]}$$

This indicator varies from 0 to 1. An indicator value closer to 1 represents a shorter distance between texts.

e.2) Indicator of the distance between a text and a category

Let *n* be the number of *NPDs* selected from the corpus of texts. Every *NPD* will have assigned an only category of the classification.

It is possible to define the array $c_h=(Z_{1h}\ldots Z_{nh})$, where:

$Z_{jh}=1$ if the j^{th} selected *NPD* have assigned the category c_h, and
$Z_{jh}=0$ otherwise.

Let a_h be the number of *NPDs* selected in the category c_h:

$$a_h = \sum_{i=1}^{n} Z_{ih}$$

On the other hand, let b be the total number of selected *NPDs*:

$$b = \sum_{h=1}^{l} \sum_{i=1}^{n} Z_{ih}$$

The indicator is defined as:

$$IND2_h = \frac{a_h}{b} = \frac{\sum_{i=1}^{n} Z_{ih}}{\sum_{h=1}^{l}\sum_{i=1}^{n} Z_{ih}}$$

This indicator varies from 0 to 1. An indicator value closer to 1 represents a shorter distance between the text to be coded and the category.

e.3) Indicator of the distance between pre-codes and codes

A rather frequent situation is that of the respondent who, besides the pre-code, gives some sort of description. There were different cases:

a) The respondent had a second job and wished to describe it (for example ”and also works as a bricklayer”).

b) The respondent thought the activity or occupation was insufficiently defined and wished to specify (for example “Spanish teacher”).

c) Although the respondent had chosen a pre-code, he did not feel sure and describes the industry or occupation.

Since option a) was not very frequent, it was decided to treat the texts with pre-codes like the remaining texts. However, when the final code had to be assigned, if there was a strong discrepancy between the code derived from the automatic coding and the selected pre-code, the cases were studied for the assignment of the best code.

In this case, the treatment of the variables activity and occupation differed:

- For the variable activity, the structure of the classification (CNAE-93) is the natural way of ordering industries and their hierarchic structure may be used to determine the distance between 2 codes. Since the census pre-codes correspond to the classification codes, on the basis of levels that are common to the *NPD* code and the pre-code, the following indicator is defined:

  IND1=0.25 if pre-code and *NPD* code coincide only at the section level.
  IND1=0.50 if pre-code and *NPD's code* coincide at subsection level.
  IND1=0.75 if pre-code and *NPD's code* coincide at division level
  IND1=1.00 if pre-code and *NPD's code* coincide at group level.

- For the occupation variable, the structure of the classification did not warrant that the codes were near (for example, a doctor would never define himself as near to a mathematician, even though in the structure of the classification they are very much so). That is why a table was created with pairs of pre-codes and codes; if the pair resulting from the process appeared in the table, the code was assigned. If the pair was not in the table, it was included in a batch for later consideration.

e.4) Over-weighting of categories not included in the lists of pre-codes

The lists of pre-codes entailed the problem that the codes that were not included were going to be infra-represented in the final results. Therefore, when the *NPDs* corresponded to these codes, they were going to receive an additional weighting, i.e.

$$p = \frac{IND2 + IND4}{2}$$

f) Reiteration process: the text has not been coded and not all its words were used

As explained earlier, the incoming texts were short and it was, therefore, necessary to warrant that their entire information had been taken into account. To this end, if the text to be coded was not resolved in a coding process and if any of its words had not been considered, there would be a reiteration of the process and the resulting information would be integrated later.

f.1) Determination of the *filter word* in the reiteration process

The reiteration process follows the same steps as the initial process, except that the *filter word* is obtained from those that were not taken into account in the preliminary coding processes. If there are several words:

1. If there are words with $4<w_j£100$, the filter word is that of them with maximum wc_j.

2. If there are only words with $w_j£4$ or $w_j>100$, the *filter word* is the one that having $w_j£4$ has maximum wcj.

3. If there only are words with $w_j>100$, the reiteration would not be carried out because there would be much noise in the results.

f.2) Obtaining indicators

The same indicators are obtained as in the initial coding process.

f.3) Integration of the information derived from reiteration processes

After a reiteration process, two different files have to be joined, one with the information available before reiteration (*FILE1)* and one with the information obtained from reiteration *(FILE2).*

Those categories that comply with condition *IND3>0.3* or *IND4>0.3* or $a_h>1$ in *FILE1* or in *FILE2* are selected as *NPD's categories*, the result being a new file *F* of *NPDs*.

For *F* the indexes are redefined naturally. In the case of the indicators of the distance between texts (*IND3, IND4*), the highest value of the values assigned in *FILE1* and in *FILE2* is taken.

Annexe III shows a practical example of the calculation of all these values in the reiteration process.

g) Coding algorithm (see annexe I)

Once the information is available and the results (mainly based on the pilot surveys) are analysed, the following algorithm was created for decision making in the coding process.

g.1) Number of words in the text *u* to be coded, in the corpus of words

For a starter and since there were doubts as to the coverage of the corpus, it was deemed desirable to code only those texts whose majority of words were in the corpus.

It is possible to represent the array $u=(U_1,\ldots,U_{\#(u)})$ where:

$U_i=1$ if the i^{th} word of text *u* is included in the corpus and
$U_i=0$ otherwise.

It is possible to consider two integer values:

a) The total number of words of the text to be coded *u*:

$$D_1=\#(u)$$

b) The number of words of the text to be coded *u* that are in the corpus of words:

$$D_2 = \sum_{i=1}^{\#(u)} U_i$$

If $D_2 \pounds D_1/2$, then the text would be discarded and passed on to the coding queue. This means that a text of 2 words would only be coded if both words are in the corpus; if a text has 3 words, it would be coded if at least 2 of them are in the corpus and if the text has 4 words, at least 3 of them should belong to the corpus.

g.2) Number of words in the text *u* to be coded, used in the selection of *NPDs*

For the texts chosen for coding, the first important fact was the number of words considered for the selection of *NPDs* (*A*) as compared to the number of words in the text to be coded (*B*). If it was relatively low ($A/B<0.34$) the process was reiterated. On the contrary, if $A/B \geq 0.34$, the coding was attempted.

g.3) Number of *NPD's categories*

The next step regarded the number of different classification categories presented by the *NPDs* was tackled.

a_h $(h=1..l)$ is the number of *NPDs* selected in category c_h. If we consider the set

$$N = \{a_h / a_h > 0\}$$

It is possible to define the number *h* of *NPD's categories*, as

$$h=\#(N)$$

The treatment varied to the number of categories obtained.

If all the words were used for coding (*A/B=1*), whether in the first process or in later reiterations, the result obtained could be coded or not coded, being part in this last case of the coding queue.

g.4) Usage of the distance indicators

In some cases, when taking a decision, account was taken not only of the best *NPD* having a given value but also of it being much better than the second *NPD*.

Although the algorithm may seem complicate and needs many calculations, in fact the process is rather fast, since the type of initial selection of *NPDs* allowed a relatively small set of them. And thanks to the algorithm, the process was often successful without a need of reiteration.

h) Fictitious codes

After the entire census information was scanned and before launching the final process that would lead to attainment of queues, it was noticed that quite a number of very high frequency texts were not going to be coded.

The problem lay in the fact that these texts, although perfectly defined, did not contain enough information for them to be coded at the established levels (sub-specification). An example of such texts for activities is ”construction”, which may be classified into 5 different CNAE-93 categories (*451 Site preparation, 452 Building of complete constructions or parts thereof; civil engineering, 453 Building installation, 454 Building completion, 455 Renting of construction or demolition equipment with operator*), it being necessary to give a more thorough of the activity for its coding at the required level.

Before launching the final process, it was decided to study the texts with a frequency over 25 and to assign a fictitious code to those that could not be coded at the required level but that could receive the same treatment.

Sometimes, new texts were also incorporated in the corpus or synonyms were created to improve coverage.

The texts coded with a fictitious code did not be incorporated to the corpus of texts, but they were an independent batch. If the cleaned text to be coded were identical to one coded with a fictitious code, it was treated as was set for this fictitious code, without being involved in the automatic coding process.

These fictitious codes were analysed one by one and solved using ancillary variables (occupation when coding economic activities; level of education, socio-economic status and economic activity when coding occupations), being sometimes necessary a probabilistic imputation based on external information.

	INDUSTRY	**OCCUPATION**
Number of fictitious codes	1.470	988

i) Results of the coding process

The following table shows a numeric summary of the results obtained after the coding process:

			INDUSTRY			OCCUPATION	
		Coded	Non coded	Total	Coded	Non coded	Total
Without text	With pre-code	12.293.295	-	12.293.295	12.409.947[7]	-	12.409.947
With text	Without pre-code	1.555.499	500.590	2.056.090	1.501.914	398.153	1.900.067
	With pre-code	473.596	130.932	604.527	582.593	137.813	720.406
	Total	14.322.390	631.522	14.953.912[8]	14.494.454	535.966	15.030.420[7]

The coded texts were distributed in the following way:

		INDUSTRY	OCCUPATION
Coded texts	With a code from the classification	1.261.150	1.356.317
	With a fictitious code	726.110	683.250
	Blank cleaned text	41.835	44.940
	Total	2.029.095	2.084.507

The percentages corresponding to coded texts were:

	INDUSTRY	OCCUPATION
Responses using a text	2.660.617	2.620.473
Responses using a text that has been coded	2.029.095	2.084.507
Percentage of coded texts	76.26%	79.55%

The coding percentages corresponding to the whole census are:

	INDUSTRY	OCCUPATION
Responses to this question	14.953.912	15.030.420
Responses to this question that have been coded	14.322.390	14.494.454
Percentage of coded responses	95.78%	96.43%

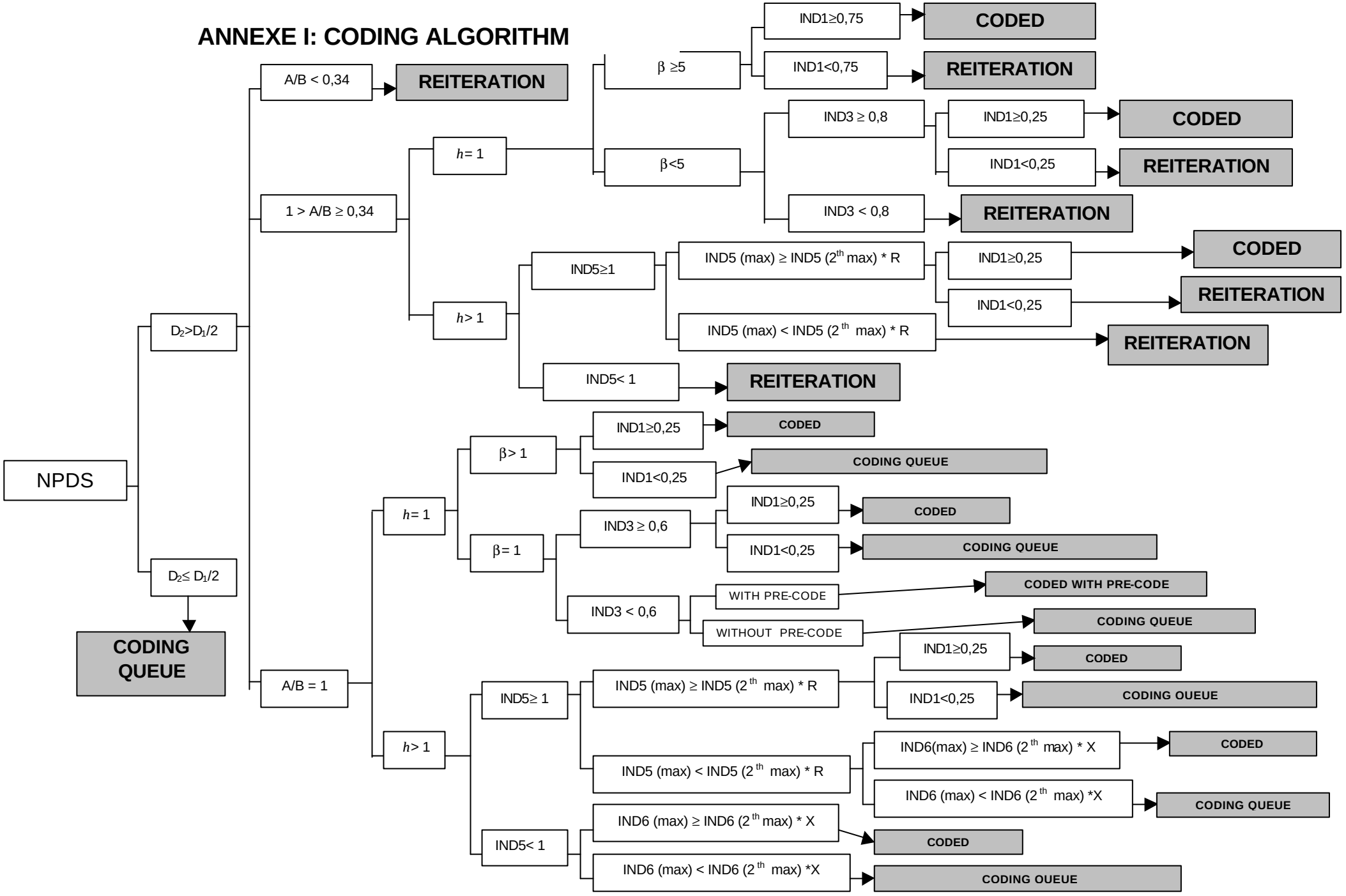
It is still needed an evaluation of the quality of this coding. Although the quality was quite high in the pilot test, it has not being possible to evaluate it after the census, because the process of the information is still running.

[7] In order to transform these pre-codes to real codes existing in the classification, ancillary information was needed. Because of the non response, this ancillary information was not always available.

[8] The occupied population in Spain is higher than this amount. This is due to the non reponse to the census questionnaire or to the corresponding questions.

Madrid, 15th September 2003

ANNEXE I: CODING ALGORITHM



VALUES USED IN THE ALGORITHM

D_1.- Number of words of the text *u* to be coded

D_2.- Number of words of the text to be coded *u* that are included in the corpus of words

A.- Number of words of the text to be coded *u* used for the selection of *NPDs*

$B = D_2$.- Number of words of the text to be coded *u* that are included in the corpus of words

h.- Total number of *NPD's categories*

b.- Total number of selected *NPDs*

IND1.- Indicator of distance between a pre-code and a code

IND2.- Indicator of distance between a text and a category

IND3.- Indicator of the distance between a *NPD* t_i and the text to be coded *u* considering the weights w_j assigned to the words in the corpus.

IND4.- Indicator of the distance between a *NPD* t_i and the text to be coded *u* considering the weights wc_j assigned to the words in relation to the categories.

IND5=IND2+IND3

IND6=IND1+IND4+IND5+p= IND1+IND2+IND3+IND4+p

p.- Over-weighting of categories not included in the lists of pre-codes

R=1.6 - (IND5(max)-1)/4

X=1.6 - (IND6(max)-1)/4

ANNEXE II: EXAMPLES OF PRE-CODES LISTS

EXAMPLE OF PRE-CODE LIST OF OCCUPATIONS

Bricklayers and Other Building or Mining Workers
U1 Construction or mining labourer
U2 Bricklayer, miner
U3 Superintendent of works, foreman, chargehand
U4 Painter, paperhanger
U5 Plumber, heating engineer
U6 Joiner (wood, aluminium)
U7 Electrician
U8 Decorator, plasterer, formworker, structural metalworker
U9 Parquet layer, tile layer, glazier, roofer

Deliverymen, Lorry, Taxi and Other Drivers
O1 Lorry driver
O2 Taxi driver, car or van driver
O3 Bus driver
O4 Motorcycle deliveryman, courier
O5 Tractor driver
O6 Engine driver
O7 Heavy machinery driver/operator

Medical Personnel
C1 Orderly, stretcher-bearer
C2 Nursing assistant (hospital or home)
C3 SRN, qualified nurse
C4 Physician (any branch), dentist
C5 Veterinary surgeon
C6 Pharmacist
C7 Assistant pharmacist, veterinary surgeon, dentist
C8 Optician, physiotherapist, chiropodist, speech therapist

Teaching Personnel
D1 Infant or primary teacher
D2 Secondary teacher
D3 University teacher
D4 Special education teacher
D5 Technical vocational training teacher
D6 Private teacher; educational inspector

Domestic Service or Cleaning; Cooks and Waiters
M1 Domestic service, cleaning lady
M2 Office, hotel cleaning personnel
M3 Waiter
M4 Cook
M5 Road sweeper, refuse collector

Proprietors or managers of small establishments (fewer than 10 employees)
A1 The company is the actual establishment or the company has fewer than 10 employes
A2 The company has 10 employees or more (e.g. a bank branch manager)

Shop Assistants, Salesmen and Commercial Agents
N1 Shop assistant
N2 Cashier, ticket clerk; lottery, charity organisation draw ticket salesperson ...
N3 Door-to-door salesman
N4 Telephone salesman
N5 Representative, travelling salesman, medical salesman
N6 Insurance agent, travel agent, purchasing agent, stockbroker

Officials who deal directly with the public
K1 Telephonist, receptionist, travel agency clerk
K2 Postman, library assistant, public opinion surveyor
K4 Other clerical officer dealing directly with the public

Other officials
L1 Office secretary, clerical worker, legal clerk
L2 Bank assistant, accounts clerk
L3 Storeman, station manager
L4 Other clerical officer whose main task is not dealing directly with the public

Farmers, Stockbreeders, Fishermen and their Assistants
T1 Farmhand, stockbreeding or fishing worker
T2 Farmer, gardener, nurseryman
T5 Fisherman, fish farmer
T6 Stockbreeder, shepherd; forestry worker

Defence and Security
RO Armed Forces
R4 Member of the national, regional or municipal police force
R5 Civil Guard
R6 Licensed security officer; private security guard
R7 Fireman, forest ranger

Skilled Industrial Workers; Tradesmen
Mechanic, Service Engineer, Welder...
W1 Mechanic, machine fitter
W2 Electrical equipment service engineer
W3 Shop supervisor, works team foreman
W4 Panel beater, welder, moulder
W5 Locksmith, foundry worker, diemaker, polisher

Mechanised Industrial Production Worker; Fitter
Z1 Industrial product fitter
Z2 Industrial robot operator
Z3 Fixed machinery operator: oven, press, saw, milling machine, weaving machine, packing machine...
Z4 Production line worker

Craftsmen; Traditional Craft Industry Worker
X1 Maker of food, beverages and tobacco products
X2 Tailor, shoemaker, embroiderer, upholsterer
X3 Printing: film developer, bookbinder
X4 Pottery or glassware craftsman
X5 Wood, leather, textile industry craftsman
X6 Cabinet maker, turner, basketmaker

Government Administrative Officers or Managers of Companies or more than 10 Employees
B1 Executive or legislative power; government office administrator (up to deputy director)
B2 Chairman or general manager
B3 Head of department of the company's actual business activity
B4 Other head of department (accounting...)

Law, Social Science and Arts Professionals
F1 Junior contracted accountant; qualified social worker
F2 Senior contracted accountant
F3 Lawyer, public prosecutor
F4 Tax or employment consultant, solicitor/notary, registrar
F6 Psychologist, sociologist, interpreter, translator
F7 Writer, journalist; actor, painter, musician.....
F8 Welfare worker; social worker

Computer Technicians and Scientific Officers
H1 Sytems analyst or equivalent
H2 Applications analyst or equivalent
H3 Programmer or computer operator
H4 Keyboarder
H5 Draughtsman, technical designer
H6 Laboratory, electronic, chemical technician
H7 Quality control, safety officer
H8 Photographer, cameraman, sound technician

Other Occupations typical of Further or Advanced Education
J1 Ordinary/honours degree engineer or equivalent
J2 Architect, quantity surveyor
J6 Tax Inspector or other occupation belonging solely to the PA, group A
J7 Assistant Tax Inspector or other PA occupation, group B

EXAMPLE OF PRE-CODE LIST OF INDUSTRIES

Hotels and Catering
551 Hotel, boarding house, guest house
552 Camping site, holiday apartments
553 Bar that serves meals, restaurant
554 Bar that does not serve meals, pub

Health and Social Services
851 Healthcare activities (hospital, clinic, doctor's surgery...)
853 Day nursery; old people's home; treatment centre for drug addicts; treatment centre for the handicapped
854 NGO

Food Industry
158 Manufacture of bread, cakes and buns, biscuits, and pasta; confectionery
151 Meat industry
159 Manufacture of beverages (wine, mineral water...)

555 Catering company
Building Trade
451 Demolition and ground clearing
459 Construction of Public Works (bridges, roads...)
452 Construction of buildings; bricklaying and masonry work in general, minor alterations
453 Company engaged in electrical installations, plumbing, insulation
454 Company engaged in installation of doors and windows, glazing, painting, plastering or tiling

Retailing
522 Greengrocer's, butcher's, fish shop, cake shop, frozen foods shop, grocer's or other food shop
521 Hypermaket, supermarket or department store
523 Pharmacy, toiletries
524 Household goods, hardware, do-it-yourself; household electrical appliance or furniture shop, shoe shop, boutique; optician's
529 Jeweller's, watchmaker's; gift shop, bargain shop; toy shop, sports shop; stationer's, bookshop, newspaper shop/stand
526 Street market, door-to-door
528 Telephone or Internet sales

Wholesaling
511 Commercial agent; commodity market
513 Food, beverages or tobacco products
514 Clothes, household electrical appliances or furniture
515 Building materials, scrap metal, chemical products
516 Machinery, industrial equipment or electrical supplies
Motor Vehicle Services
501 Motor vehicle dealer/distributor or sales
502 Motor vehicle repair shop
503 Sale of motor vehicle spares
504 Motorcycle sale and repair
505 Petrol station

Transport
601 Rail transport
602 Road transport; taxi
611 Sea transport
621 Air transport
631 Goods storage and warehousing
632 Bus or train station, harbours and airports
647 Urban courier service

Education
801 Infant or primary school
802 Secondary school
803 University or college
804 Tuition centre, driving school or other educational institution

Government Services
641 Mail
752 Defence, Justice, Law and Order, Civil Defence, Foreign Affairs
753 Social Security
751 Other Ministries, Departments, Town Council, Local Authority or other Government (central, regional or local) Agency

Domestic or Cleaning Service
950 Of households or communities (domestic help, janitor...)
900 Road sweeping and refuse collection
747 Cleaning company

Banking and Insurance
651 Bank or Savings Bank
660 Insurance company
671 Portfolio management company

Other Services
930 Hairdresser's or beauty parlour; drycleaner's
746 Security company
527 Repair of clocks and watches, household electrical appliances, shoes, clothing
741 Tax or accounting consultancy; lawyer's office; solicitor's/notary's office
742 Technical engineering and architecture services
720 Computer services company
922 Broadcasting activities
642 Telecommunications
730 R&D (Research and Development)
748 Reprography services, photographic studios
633 Travel agency
744 Advertising agency
703 Estate agency; property management
401 Electricity utility
410 Water utility
402 Gas utility

155 Dairy industry

Motor and Electrical or Electronic Machinery Industry
341 Manufacture of motor vehicles
343 Manufacture of shock absorbers, exhaust pipes, steering wheels or other non-electrical parts for motor vehicles
353 Aircraft construction
300 Manufacture of computers and other office machines
322 Manufacture of telephones, fax machines, and radio and television sets
334 Manufacture of optical instruments and photographic equipment
316 Manufacture of electrical components (generators, electrodes, electrical insulators, burglar alarms....)
291 Manufacture of tap fittings, pumps, compressors, valves, transmission components and engines for boats
292 Manufacture of general purpose industrial machinery (lifts, packing, ovens, ventilation....)

Chemical Industry
244 Manufacture of pharmaceuticals
245 Manufacture of perfumres, detergents or cleaning products
246 Maufacture of chemicals (lubricants, for photography, cassettes and CDs, explosives...)
252 Manufacture of plastic products

Other Industries
222 Printing, printing shop
221 Publishing
182 Clothes making
361 Furniture industry
212 Manufacture of paper and cardboard articles
281 Manufacture of structures and metal joinery
287 Manufacture of metal containers and nuts and bolts

Agriculture, Stockbreeding, Gardening...
013 Agricultural production combined with stockbreeding (each represents at least 1/3 of the total)
011 Agriculture
012 Stockbreeding
014 Gardening, pruning, harvesting, etc. services
020 Silviculture

ANNEXE III: AN EXAMPLE OF THE CODING PROCESS

This example is made for the case of being coding economic activities. The following text has to be coded:

FRUITS, PULSES AND VEGETABLES STORE

This text is accompanied by a pre-code in the census questionnaire, which is **513**.

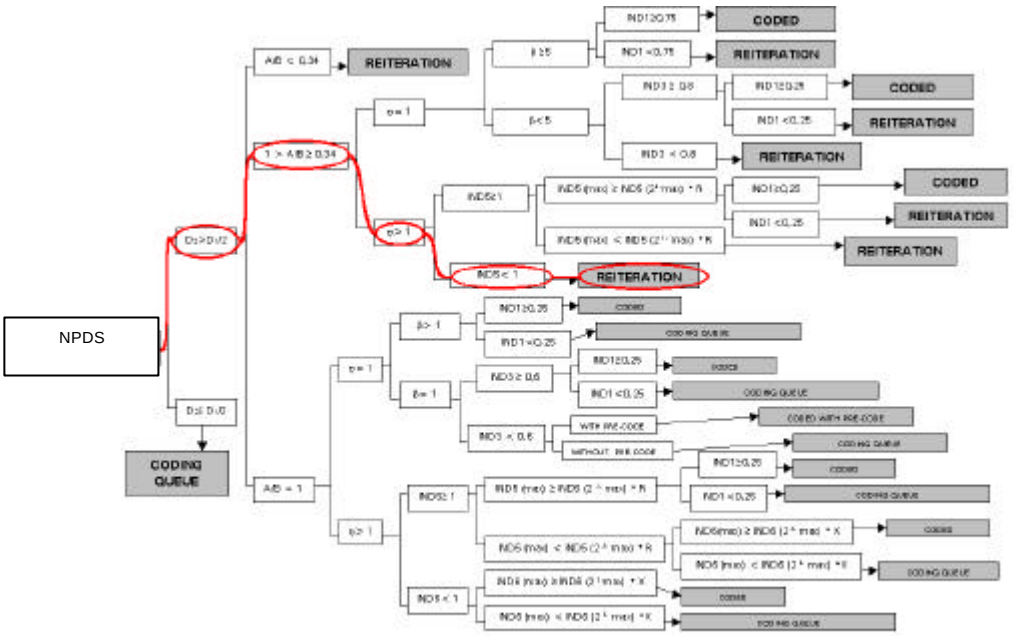
Separating the words in this text, and searching them in the corpus of words, the result is:

WORDS	W_j	WC_j
FRUITS	47	0.43
STORE	1	1
PULSES	6	0.33
VEGETABLES	10	0.40

1. **FIRST SELECTION OF *NPDS***: The process considers the words FRUITS (*filter word*) and VEGETABLES, selecting the following *NPDs*:

File1															
c_h	*TEXT*	D_1	D_2	*A*	*B*	h	a_h	b	*IND1*	*IND2*	*IND3*	*IND4*	*p*	*IND5*	*IND6*
014	PACKING OF FRUITS, PULSES AND VEGETABLES FOR THE PRIMARY MARKET	4	4	3	4	3	1	3	0	0,33	0,07	0,05	0	0,40	0,45
513	WHOLESALE OF FRUITS, POTATOS AND VEGETABLES, INCLUDING PULSES	4	4	3	4	3	1	3	1	0,33	0,02	0,30	0	0,35	1,65
522	RETAIL SALE OF PULSES, FRUITS AND VEGETABLES	4	4	3	4	3	1	3	0,50	0,33	0,32	0,12	0	0,65	1,27

Applying the algorithm, the result is a new **REITERATION** of the process:



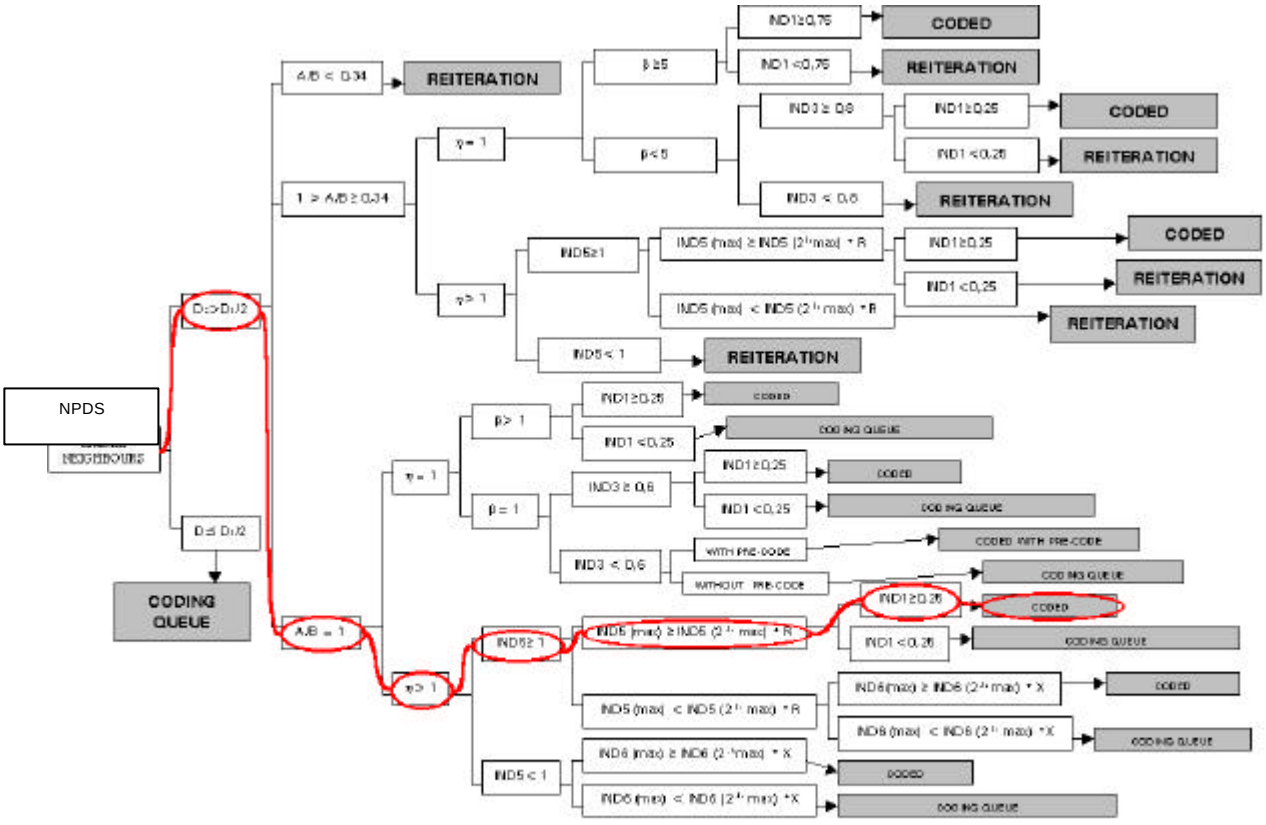
2. **SECOND SELECTION OF *NPDS***: The process considers the words STORE (*filter word*) and FRUITS, selecting only the following *NPDs*:

File2															
c_h	***TEXT***	***D_1***	***D_2***	***A***	***B***	h	a_h	b	***IND1***	***IND2***	***IND3***	***IND4***	***p***	***IND5***	***IND6***
513	WHOLESALE STORE OF CITRUS FRUITS	4	4	2	4	1	1	1	1	1	0,35	0,45	0	1,35	2,80

3. **CALCULATING THE VALUES IN FILE *F***: The next step is the integration of the information obtained from the two processes:

File *F*															
c_h	***TEXT***	***D_1***	***D_2***	***A***	***B***	h	a_h	b	***IND1***	***IND2***	***IND3***	***IND4***	***p***	***IND5***	***IND6***
513	WHOLESALE STORE OF CITRUS FRUITS	4	4	4	4	2	2	3	1	0,67	0,35	0,45	0	1,02	2,47
522	RETAIL SALE OF PULSES, FRUITS AND VEGETABLES	4	4	4	4	2	1	3	0,50	0,33	0,32	0,12	0	0,65	1,27

Applying the algorithm for the values in file *F*, it results that the text is **CODED** with code 513:



REFERENCES

CNAE-93, Clasificación Nacional de Actividades Económicas 1993, ISBN 84-260-2747-4, National Statistics Institute of Spain, 1993.
http://www.ine.es/inebase/cgi/um?M=%2Ft40%2Fclasnac%2F&O=inebase&N=&L=

CNO-94, Clasificación Nacional de Ocupaciones 1994, ISBN 84-260-2895-0, National Statistics Institute of Spain, 1994.
http://www.ine.es/inebase/cgi/um?M=%2Ft40%2Fclasnac%2F&O=inebase&N=&L=

ISCO-88, International Standard Classification of Occupations, ISBN 92-2-106438-7, International Labour Office (ILO), Geneva, 1990.
http://www.ilo.org/public/english/bureau/stat/class/isco.htm

Mitchell, T.M. (1996), "Machine Learning", McGraw Hill, New York, NY.

NACE Rev.1, Statistical Classification of Economic Activities in the European Community, Eurostat, ISBN 92-826-8767-8, Office for Official Publications of the European Communities, Luxembourg,1996. http://europa.eu.int/comm/eurostat/ramon/

Recommendations from UNECE-EUROSTAT for the European round of 2000 Censuses of Population and Housing.

Sebastiani, F. (2002), "Machine Learning in Automated Text Categorization", *Association of Computing Machinery Computing Surveys*, 34, to appear.

Winkler, W. E. (2003), "Machine Learning Methods for Text Classification", talk given at the Italian National Statistical Institute, January 2003.