# ECONOMIC UNIT ADMINISTRATIVE RECORDS LINKAGE USING PROBABILISTIC TECHNIQUES

**Laura Otero Franco**

**EUSKAL ESTATISTIKA ERAKUNDEA**
BASQUE STATISTICS OFFICE

Donostia-San Sebastián, 1
01010 VITORIA-GASTEIZ
Tel.: 945 01 75 00
Fax.: 945 01 75 01
E-mail: eustat@eustat.es
www.eustat.es

# PRESENTATION

An administrative record is a document that contains information related to a unit, that may be an individual, an establishment or other entity, that an administrative department collects for its own purposes. In recent years, the potential of statistics based on using administrative records has progressed at an extremely fast rate.

Administrative records have become a powerful tool to measure and analyse the situation in different spheres. One of the most frequent statistical uses of the administrative records is for the construction and upkeep of directories, due to their efficiency and cost rationality.

However, using administrative information in the statistics environment is not always straightforward. For example, administrative records are frequently gathered in previous periods or for units that are very different to the ones to be used in the statistical operation, and mainly (as a characteristic intrinsic to any administrative records) they have not been conceived or designed for statistical purposes. The information provided by these records therefore needs prior processing before it is used in the statistics environment.

One of these prior processes is data linkage or record linkage. These techniques enable the information referring to a single unit contained in different administrative records to be linked.

Vitoria-Gasteiz, November 2010

JAVIER FORCADA SAINZ

General Director

# CONTENTS

PRESENTATION ..... 1

CONTENTS ..... 3

INTRODUCTION ..... 4

- INTRODUCTION AND OBJECTIVES ..... 4
- DESCRIPTION OF THE PROJECT ..... 4
- BACKGROUND ..... 5

ADMINISTRATIVE RECORDS ..... 6

- CHARACTERISTICS ..... 6
- ADVANTAGES AND DISADVANTAGES ..... 7
- COMPANY RECORDS AND OTHER LEGAL UNITS OF INTEREST FOR ECONOMIC STATISTICS ..... 8

RECORD LINKAGE ..... 11

- METHODOLOGY ..... 11

PROGRAMMING ..... 15

- GENERAL PROGRAM ..... 15
- INITIAL PARAMETERS. ..... 18
- ANALYSIS ..... 19
- STANDARDISATION AND HOMOGENISATION. ..... 19
- PROBABILITY CALCULATION. ..... 25
- BLOCKING. ..... 26
- LINKS ..... 27

ANALYSIS OF THE RESULTS ..... 29

- DESCRIPTION OF THE FILES ..... 29
- ANALYSIS OF THE FILES ..... 30
- RESULTS OF THE LINKAGE ..... 32

CONCLUSIONS ..... 33

BIBLIOGRAPHY ..... 34

Chapter 1

# INTRODUCTION

## Introduction and objectives

Data linkage or record linkage is defined as the procedure to find pairs of elements in two different administrative records (one element per record) that represent the same unit. If the administrative records were the same, it would be said that duplicate elements had been located.

There are two general techniques within data linkage:

- **Deterministic linkage:** The elements are related when they exactly coincide in all fields or in a pre-determined number of fields.
- **Probabilistic linkage:** A certain probabilistic weighting is allocated to each pair of elements and the pairs whose weighting is sufficiently high.

This technical notebook seeks to set out the study of the administrative records linkage that contains data on economic units by means of probabilistic methods. More specifically, the study is focused on the linkage of the following files: the EUSTAT Economic Activities Directory (DIRAE) and the Social Security file.

## Description of the Project

The project is divided into three phases:

I. Studying and adapting the probabilistic linkage methodology proposed by Fellegi & Sunter to process the administrative records of economic units.

II. Programming an application in SAS that automatically links two administrative records of economic units.

III. Analysis of the results obtained.

This technical notebook has been divided as follows to present the different stages of the project:

Chapter 1 sets out the objectives of the technical notebook, the project stages and EUSTAT's background in record linkage.

Chapter 2 introduces the administrative record concept and sets out its characteristics, advantages and disadvantages and considers the case of administrative records with economic unit data.

Chapter 3 describes the record linkage methodology using the probabilistic techniques presented by Fellegi and Sunter in their 1969 article entitled “A theory for Record Linkage”.

Chapter 4 considers the structure of the economic unit administrative record linkage program using probabilistic techniques developed by EUSTAT:

Chapter 5 analyses the results obtained when using the linkage program with the files leading to the study, that is, the EUSTAT Economic Activities Directory (DIRAE) and the Social Security file.

Finally, Chapter 6 contains the conclusions reached by studying these techniques and the application of the linkage program developed.

## Background

When the last review of the Municipal Population Register was performed in 1996, EUSTAT decided to start working with administrative records in order to produce Population Statistics Records. The information contained in the full Municipal Population Register as of 31 December each year and the Statistics on the Natural Population Movement (Births, Deaths and Marriages) was therefore used.

Determinist linkage techniques began to be introduced to process those files and they used common identification variables (name, surname, ID/tax number, date of birth, postal address, etc.). Subsequently, a study began of the probabilistic techniques based on the theoretical model presented by Fellegi and Sunter in 1969 in their article "A theory for Record Linkage”, funded by a EUSTAT mathematical-statistical methodology and research grant.

Given the need to expand the use of both the determinist and probabilistic procedures, work began on programming an independent linkage application, known as the Linkage Model (MDF).

Driven by the success obtained by the individual record linkage, it was decided to study the application of the same probabilistic linkage methods to merge economic unit files. In principle, the idea was to adapt the existing programme for individual linkage, but it was soon found that that was no trivial matter given the specific nature of the economic unit records.

Chapter 2

# ADMINISTRATIVE RECORDS

Administrative records are a source of statistical data, along with surveys or censuses. The main difference with the other two sources is basically that the administrative record complies with legislative criteria established by the administrative department that owns the record, instead of with the methodological criteria established by the statistical authority.

There is a very large number of administrative records in the possession of the different public administrations likely to be used for statistical purposes, which collect a great variety of data with information relating to individuals, companies, institutions and other populations, data that a priori would not be necessary to collect again from the reporters, thus reducing the economic cost of the data collection.

The statistics institutions have prepared lists of records likely to be used statistically more frequently in recent years and have made great progress in the supply of information from the major data holders, along with the social security, health, education and tax authorities. Other records such as the Civil Registry have been used for statistical purposes for some time now.

## Characteristics

Some of the characteristics of the administrative records are listed below:

1) They are legislative documents that record an administrative event, that is, they are not designed for statistical purposes.

2) They do not follow a set of statistical steps as their fundamental use is for administrative control, coordination, monitoring, records and/or planning.

3) They have tools that are characteristic of information collection (templates, forms, cards, record cards, notebooks or minute books, etc.) that can be used to meet the administrative control and monitoring needs.

4) Their frequency may or may not be established in the relevant legislation. For example, it may be registered once (such as university degree), a frequency may be established (such as transit and licence records) or it may be random (such as health records).

5) They provide as broad coverage as that of the authority that owns the administrative record.

6) They define different units to be recorded, individuals, buildings or establishments, facts (events or activities), services, resources, commercial transitions, etc, and therefore, multiple variables.

7) The units recorded and the variables measures can be converted to statistical variables by means of relatively simple methodologies.

8) They are stored in different format archives that range from hard copy, files or dossiers to digital media.

## Advantages and disadvantages

Administrative records offer the following advantages over other sources:

1) Low cost data production. In general, administrative records follow a legislative procedure of an institution that needs to control an administrative action. Therefore, the information is generally collected at the institution offices and there are no field operating costs.

2) Lower burden regarding completing forms for the people providing the information. In general, national surveys are usually long, even though it has to be remembered that there are complicated administrative records such as tax returns.

3) They enable information to be circulated between the government bodies and thus avoid doubling up the work of the public administration. This theoretical advantage requires the public administration entities to be synchronised in terms of data requests. This process is very complex as the templates of each entity are independent from those of the others or the populations targeted by the instruments can vary.

4) A comprehensive coverage of the target population is achieved. In many cases, the administrative records achieve comprehensive coverage, for example in areas of health, justice or education. However, there are cases such as victimization where the administrative records do not achieve coverage as they may not be sufficient cases reported.

5) The non-response errors are fewer than in other sources and there are no sample errors. It is true that there are no sample errors as a sample is not taken strictly speaking, but this error is replaced by the coverage error.

6) Desegregation into sub-populations is possible. This advantage is very important as the administration record can contain an interesting variety of data that can be used to obtain sub-populations. It is therefore important to analyse the administrative record to assess its relevance.

7) The information systems are strengthened in all territorial spheres of a country. It is obvious when the State has a policy to reinforce its information systems which means this advantage is a greater opportunity.

8) The quality of the information increases as forms can be produced using the details required by the area being studied. However, amending forms is complex, even when there is an opportunity to do so.

9) It is a sound base for constructing data series. The administrative records provide a history of the administrative process, which helps to construct a time series.

On the other hand, the administrative records also have certain disadvantages, such as the following:

1) The lack of correspondence between the administrative and statistical units. Obviously, the analysis units cannot coincide as the administrative record may refer to an individual and not necessarily to the household or to a commercial establishment.

2) Differences in the definitions of the variables. In general, the templates and formats do not contain methodologies that define the operational variables or can be identification or description variables without going further as a descriptive study would require.

3) Lack of conversion between the statistical and administrative codes.

4) Reference periods and data not coinciding with the statistical purpose. This can occur as the administrative records do not have a de facto statistical purpose, but there are processes to convert administrative records into statistical data.

5) Impact of political changes on the continuity of the administrative records. The administrative procedures and templates can certainly vary with the political toing and froing. This aspect is very important and shows a weakness of the administrative record.

6) Lack of a common identifier in the records to reconcile the data and lack of stable staff for the task.

7) Lack of a long-term view that leads to developing the statistical system and which favours economic and social trends.

8) Lack of a cooperation policy among the entities that supply the records, the lack of an agreement among all the participating stakeholders and institutions and the lack of statistical laws for all territorial desegregations.

## Company records and other legal units of interest for economic statistics

Some of the most important records to detect and update the economic production statistical units are the Business Tax (IAE), Social Security contribution units (SS) and the Trade Registries (RM).

The IAE is a municipal tax that is charged annually on business, professional and artistic activities. However, its centralised management is performed in our case by the provincial tax offices and by the Spanish Tax Authorities (AEAT) in the “Common Territory” (territorio comun) as the rest of Spain is referred to. Each activity in a specific place requires a licence, which means that a taxpayer may hold several licences at one or more registered addresses. Along with other identification details of the tax payer and

the address, the licence contains registration and termination dates, a business code or epigraph (based on the former CNAE-74) and a literal description of the activity.

The IAE covers nearly all the business activities carried out by companies, individuals and bodies corporate, the self-employed and artists (in our case, in the sphere of the provinces of the Basque Country). All the tax payers subject to Value Added Tax (VAT) are required to so declare when registering for the IAE.

However, current legislation establishes that all individuals and bodies corporate under a certain turnover are exempt from paying the tax. That has had the immediate consequence that those tax payers often do not terminate their registration when they stop trading, which provides an inflated image of the economic situation.

The IAE also does not provide information about farmers, stock breeders and other primary sector activities, or is it clear about the non-trading activities of the public authorities and other non-profit institutions. The files that the provincial tax authorities provide frequently do not include the data of the large incorporations with national business licences, with which they have special pro ratio agreements regarding the part of the tax for the Basque municipalities.

The Social Security records of interest for the records of companies and bodies corporate consist of the Contribution Account Codes (CCC), the Self-Employed Special System records (RETA) and for the workers include under the Agricultural Special System (REA, which no longer exists as it has been incorporated to the RETA) and the Maritime Special System (REA). There are currently quarterly files with the registered units as of a specific date, which in our case are always restricted to the geographical sphere of our autonomous community.

The CCC contains the basic data of the individual or corporate entrepreneur, just the tax identification and name, and details of each contribution account relating to the employees contracted by the entrepreneur. Each entrepreneur may have one or more CCC in a specific place if there are groups of individuals contracted under different Social Security contribution methods. Each CCC contains details of the management of the contribution unit, CNAE code of the activity, status and registration date, type of labour relation and number of workers under that CCC.

The RETA records contain descriptive details of the individuals subject to that system, along with the management details of the business and, less frequently, the CNAE code. No distinction is made between the entrepreneurs as such and other workers, main partners, cooperative member, dependent self-employed and others, where the Social Security system requires them to be included in this system instead of in the general system.

It can thus be assumed that we do not have any details from the SS regarding the large number of companies, particularly limited companies, whose workers (usually not more than two or three) are also the owners or main shareholders and who contribute through the RETA instead of by means of a CCC of the general system.

The SS data are usually more up-to-date, as the entrepreneurs are fined merely for the late filing of a return and late payments of the contributions or even for not terminating its contribution units if they cease trading. The open yet obsolete contribution units nearly exclusively refer to companies fraudulently abandoned by its owners until they are automatically deregistered by the authorities.

The Trade Registries are the main tool to provide legal security to the companies and other parties interested in trading. All companies are required to register and declare certain events, while that is only a voluntary requirement for the individuals and other non-corporate entities that trade. The companies are also required to deposit their accounts with the trade registry as public disclosure. The company can be fined if it fails to file the accounts one year.

The available RM files enables us to learn the fundamental data of the companies whose registered office is in any of the three provinces of our autonomous community: Tax and registration identification numbers, company name, address of the registered office, corporate purpose and other less important data. The filing of the financial statements is also of great interest for economic surveys and statistics.

In the same way as with the SS, there may be "abandoned" companies in the RM that have not been wound up by their owners or that, even though they are still trading, have not filed their financial statements for several years.

In all the cases of the aforementioned three administrative sources, IAE, SS and RM, the contents of the respective variables are totally disparate, with each one having their own reasoning, omissions, errors, etc., which is the result of very different administrative services and which are highly disperse throughout the geographical area.

Chapter 3

# RECORD LINKAGE

As has been seen, administrative records are not a true statistic source (origin), as are the censuses and surveys, as, on the one hand, their purpose is administrative and, on the other hand, they act as legislative control, that is, they record an individual event or act referring to an entity and which affects it directly.

Furthermore, a statistical operation requires definitions and classifications in line with the targets of the research, while the administrative records do not necessarily coincide with those methodological aspects.

Record linkage is one of the tasks that require the processing of administrative records as a statistical source. By means of this technique, information from different administrative sources can be used in an appropriate way for the statistical use. This chapter describes the methodology that EUSTAT has used to develop the probabilistic linkage techniques.

## Methodology

The theoretical model proposed by Fellegi and Sunter in their article "A theory for record linkage" in 1969 was followed to prepare the automatic linkage program. The basis of this model is as follows:

### Theoretical Model

$A$ and $B$ are used to denote the administrative record to be linked and $a$ and $b$ are the generic members of the administrative records, respectively.

Both records are assumed to have common members, and therefore, the objective of the linkage is to recognise, from among all the pairs of *AxB* that can be formed, those that refer, in our case, to the same economic unit. That is, the objective is to divide the set

$$AxB = \{(a,b) \mid a \in A, b \in B\}$$

into the combination of the disjointed sets

$$M = \{(a,b) \mid a = b, a \in A, b \in B\}$$

and

$$U = \{(a,b) \mid a \neq b, a \in A, b \in B\}$$

which are called sets of **matches** and **no-matches** respectively.

Each unit of the population in question has associated characteristics, such as, name, postal address, job, etc. Those members that refer to a single economic unit have to be identified. However, the process to create the administrative records may introduce errors or inaccuracies (code, transcription and typing errors, phonetic or typographic variations, data loss, etc.) in the elements generated. As a result of these errors, two members of A and B that do not refer to the same economic unit may generate identical elements and, more frequently, two identical members of A and B may produce different elements. The elements corresponding to the members of A and B are denoted by $\alpha(a)$ and $\beta(b)$, respectively.

The first step to try to match elements from two administrative records is to compare them. The result of the comparison is a set of codes, that are codified by statements such as: "the name coincides in both elements", "the name coincides and it is Almacenes Garrido", "the name does not coincide", "the name is absent in one of the elements" or "part but not all of the name matches". Formally, the **comparison vector** is defined as a function vector of the elements $\alpha(a)$ and $\beta(b)$ as:

$$\gamma[\alpha(a),\beta(b)] = \{\gamma^1[\alpha(a),\beta(b)],\ldots,\gamma^k[\alpha(a),\beta(b)]\}$$

It can be seen that $\gamma$ is a function defined on $AxB$. It can be written as $\gamma(a,b)$, $\gamma(\alpha,\beta)$ or just as $\gamma$. The set of all the possible realizations of $\gamma$ is called ***comparison space*** and is denoted by $\Gamma$.

During the linkage operation, $\gamma(a,b)$ is observed and it has to be decided whether:

- $(a,b)$ is a match, $(a,b) \in M$ (this decision is called ***link*** and is denoted by $A_1$)
- $(a,b)$ is a no-match, $(a,b) \in U$ (this decision is called **no-link** and is denoted by $A_3$)

However, there are situations where it would be impossible to take one of these two decisions for specific error levels, which enables a third decision, denoted by $A_2$, which is called ***possible link***.

Under these conditions, an $L$ ***linkage rule*** as an application of the $\Gamma$ comparison space regarding the set of random decision functions $D = \{d(\gamma)\}$ where:

$$d(\gamma) = \{P(A_1 \mid \gamma),\ P(A_2 \mid \gamma),\ P(A_3 \mid \gamma)\};\quad \gamma \in \Gamma$$

and

$$\sum_{i=1}^{3} P(A_i \mid \gamma) = 1$$

In other words, for each observed value of $\gamma$, the linkage rule assigns the probabilities of taking each one of the three possible decisions.

The error levels associated to each linkage rule have to be considered. It is assumed that a pair of elements $[\alpha(a), \beta(b)]$ is selected at random to be compared using a probabilistic process. The resulting comparison vector $\gamma[\alpha(a), \beta(b)]$ is therefore a random variable. The ***conditional probability of*** $\gamma$ ***given that*** $(a,b) \in M$ is denoted as $m(\gamma)$, and will be:

$$m(\gamma) = P\{\gamma[\alpha(a), \beta(b)] \mid (a,b) \in M\} = \sum_{(a,b)\in M} P\{\gamma[\alpha(a), \beta(b)]\}.P[(a,b) \mid M]$$

In a similar way, the ***conditional probability of*** $\gamma$ ***given that*** $(a,b) \in U$ is denoted by $u(\gamma)$. Therefore,

$$u(\gamma) = P\{\gamma[\alpha(a), \beta(b)] \mid (a,b) \in U\} = \sum_{(a,b)\in U} P\{\gamma[\alpha(a), \beta(b)]\}.P[(a,b) \mid U]$$

There are two types of errors associated to this linkage rule. The first occurs when it is decided to assign them as *links* when comparing pairs of elements that do not correspond with *matches* and has as probability:

$$P(A_1 \mid U) = \sum_{\gamma\in\Gamma} u(\gamma).P(A_1 \mid \gamma)$$

The second type of error occurs when a *match* is compared and is considered as *no-link* and has as probability:

$$P(A_3 \mid M) = \sum_{\gamma\in\Gamma} m(\gamma).P(A_3 \mid \gamma)$$

A linkage rule in space $\Gamma$ is said to be a ***linkage rule at error levels*** $\mu$, $\lambda$ $(0 < \mu < 1, \quad 0 < \lambda < 1)$ and is denoted by $L(\mu, \lambda, \Gamma)$ if

$$P(A_1 \mid U) = \mu \quad y \quad P(A_3 \mid M) = \lambda$$

The linkage rule $L(\mu, \lambda, \Gamma)$ is said to be optimum if the relationship

$$P(A_2 \mid L) \leq P(A_2 \mid L')$$

is kept for any $L'(\mu, \lambda, \Gamma')$ between all the linkage rules that verify the previous relationships.

It can be said that, according to this definition, an optimum decision rule is the one that maximises the probabilities of adopting positive comparison provision (that is, decisions $A_1$ and $A_3$) subject to fixed error levels. This seems to be a reasonable decision, given that adopting decision $A_2$ requires expensive manual linkage operations. Furthermore, on the other hand, it seems that if the probability of $A_2$ is not small, the linkage process is of doubtful usefulness.

The authors proposed an optimum linkage rule at error levels $(\mu, \lambda)$ that is expressed as follows:

$$d(\gamma) = \begin{cases} (1,0,0) & if\ T_\mu \le \dfrac{m(\gamma)}{u(\gamma)} \\ (0,1,0) & if\ T_\lambda < \dfrac{m(\gamma)}{u(\gamma)} < T_\mu \\ (0,0,1) & if\ \dfrac{m(\gamma)}{u(\gamma)} \le T_\lambda \end{cases}$$

where $T_\mu = \dfrac{m(\gamma_n)}{u(\gamma_n)}$, $T_\lambda = \dfrac{m(\gamma_{n'})}{u(\gamma_{n'})}$ and $n, n'$ two whole numbers so that $0 < n \le n' < N_\Gamma$.

It would be possible to tolerate sufficiently high error levels to eliminate the possibility of action $A_2$ in many applications. In this case, $n$ and $n'$, or even $T_\mu$ and $T_\lambda$, are considered so that the average set of $\gamma$ in the previous expression would be empty. In other words, each pair $(a,b)$ is located either in $M$ or in $U$. In fact, this is the decision that has been adopted in EUSTAT when developing the automatic linkage program, so that a unique limit $T_\mu = T_\lambda$ is established

Further details of constructing the optimum linkage rule, calculating the weighting $m(\gamma)$ and $u(\gamma)$ and other details of the theoretical model can be consulted in the aforementioned article by Fellegi and Sunter [1] or the "Automatic Methods of Record Linkage and their Use in EUSTAT" technical notebook [2].

Chapter 4

# PROGRAMMING

A program has been prepared in SAS to carry out the linkage between two economic unit files in accordance with the methodology set out in the previous chapter. This program is mainly aimed at the linkage of two specific files, i.e. the EUSTAT Economic Activities Directory (DIRAE) and the Social Security file.

Due to the specific features of both files, it has not been possible to build a generic program for the linkage of any two files, but ad hoc programming has been necessary for those specific files. However, the linkage program is a sequential and modular program, in such a way that many macros used by the program can be used in subsequent linkages with different files.[1]

This chapter describes the structure of the linkage program. In the following chapter, emphasis will be placed on the files being studied and on the necessary auxiliary procedures to implement the linkage between them.

## General program

The linkage program consists of the following SAS files:

→ Main program where the user defines the necessary arguments for its implementation.

→ Program that contains the macros used by the main program and which perform the different stages of the linkage.

Apart from containing the macros of the linkage based on the theoretical model of Fellegi and Sunter described in the previous section, it also contains macros to carry out a blocking procedure. This procedure is not essential in theory, but it is so at a practical level. The comparison of all the elements of an administrative record with all the elements of another administrative record cannot be assumed at computational level and, therefore, a blocking criterion needs to be used that selects a sub-set of pairs of elements likely to be linked, and thus avoid the massive comparison of all the elements.

Apart of those files, an auxiliary program has been created that builds two external tables necessary for the correct implementation of the linkage program. The tables constructed are the following:

- A set of data that contains some particles (prepositions, articles, etc.) that are eliminated from the alphabetic linkage variables in the standardisation and homogenization step.

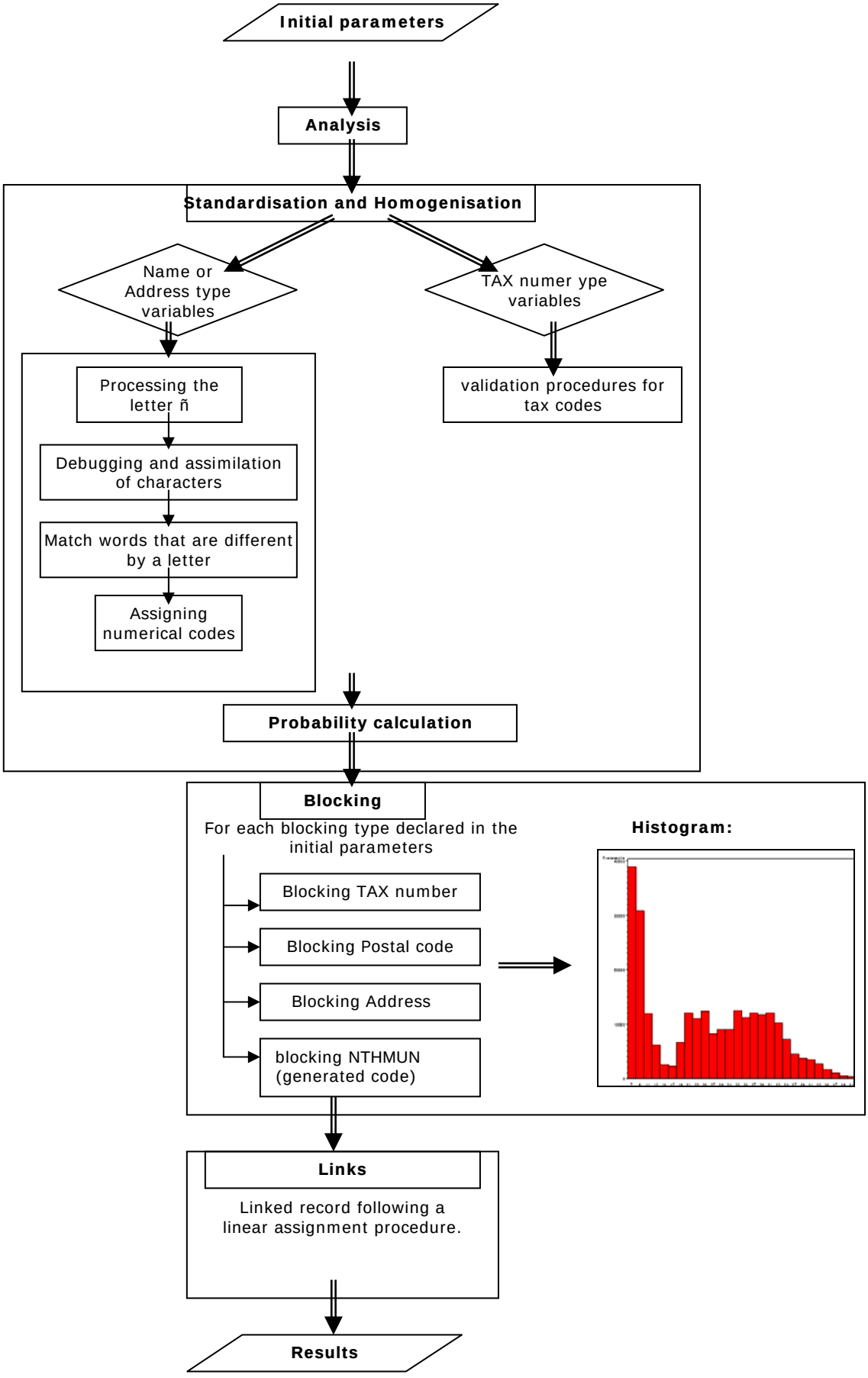
[1] The SAS macros developed in this project are available to anyone who is interested and are used for administrative record linkage for statistics production.

- A set of data that contains some initial acronyms of the company name that are eliminated from the alphabetic linkage variables in the standardisation and homogenization step.

This auxiliary program must be implemented at least once prior to the linkage program being run in order for the tables to exist when the program is run.

These tables have been built with a specific number of records, but records should be added to those tables as the results are obtained, so that they contain the greatest number of particles and acronyms of the company name possible, and the standardisation and homogenization step will be more efficient.

A flow diagram is included below that graphically depicts the algorithm of the linkage program, which is set out below:


Initial parameters
Analysis
Standardisation and Homogenisation
Name or Address type variables
TAX numer ype variables
Processing the letter ñ
Debugging and assimilation of characters
Match words that are different by a letter
Assigning numerical codes
validation procedures for tax codes
Probability calculation
Blocking
For each blocking type declared in the initial parameters
Blocking TAX number
Blocking Postal code
Blocking Address
blocking NTHMUN (generated code)
Histogram:
Links
Linked record following a linear assignment procedure.
Results

## Initial parameters

Some parameters have to be initialised prior to running the linkage program for it to work correctly. The input data that the user must enter are set out below:

- The user has to indicate what the linkage variables are, i.e., the name that the variable uses in each of the two files and what type of variable it is. The possible types of variables that can be used in the linkage and the relevant code are as follows:

| CODE | LINKAGE VARIABLE |
|---|---|
| TYPE 0 | Record IDENTIFIER KEY |
| TYPE 1 | NAME |
| TYPE 2 | STREET |
| TYPE 3 | MUNICIPALITY |
| TYPE 4 | PROVINCE |
| TYPE 5 | TELEPHONE |
| TYPE 6 | POST CODE |
| TYPE 7 | TAX NUMBER |

It should be stressed that the TYPE 0 (Record IDENTIFIER KEY) is not a linkage variable, but rather a code that univocally identifies each element within its administrative record. However, it is an essential variable as the pairs of linked elements will be identified by means of their relevant keys in the administrative records.

- At least one blocking criterion has to be indicated. As has been previously commented, even though the blocking is not an implicit linkage technique, it is essential in practice when working with files of a certain size. The possible blocking criteria that can be used and the relevant code are set out below.

| CODE | BLOCKING TYPE | DESCRIPTION |
|---|---|---|
| Criterion 1 | Tax No Blocking | Tax number coincidence |
| Criterion 2 | Post Code Blocking | Coincidence of the post code or post codes (if there is more than one) |
| Criterion 3 | Street Blocking | Coincidence of the code of the literal of the street or streets (if there is more than one) |
| Criterion 4 | NTHMUN Blocking | Coincidence of the code of the first word of the name of the company in addition to the province and the municipality of the company |

- Some theoretical initial parameters have to be established (see Bibliography [1]):

| | |
|---|---|
| $e_A$ , $e_B$ | Probability that the value of the linkage variable has been erroneously record in each of the files, respectively. It is assumed that the probability of being badly recorded is independent of each of the specific values. |
| $e_T$ | Probability that the value of the linkage variable appears differently in both files despite being correctly recorded in both. This can occur, for example, if the files were created at different times and the economic unit changes name. |

It is assumed that the values of the $e_A$ and $e_B$ parameters are sufficiently small for the probability of a coincidence between two identical entries, even though wrong, to be insignificant.

- At a given moment, the program asks the user to enter a value for the limit to determine from which weighting the pairs of economic units are to be linked.

## Analysis

The first step of the program is the analysis of the previously declared linkage key variable and linkage variables.

Certain checks are therefore performed to see that at least one linkage variable and one blocking criterion are established.

If there is any error or any datum has not been correctly introduced, the program finalises its run and a message identifying the error appears on the screen in the log that the user must solve.

If, on the other hand, all the necessary information has been entered, the program continues to run.

## Standardisation and Homogenisation

The standardisation and homogenisation tasks are highly important in the linkage processes as correct prior processing of the variables can considerably increase the number of linked pairs, and what is more important, link pairs of great quality, that is, provide greater security that they really correspond to the same unit.

There are two types of linkage variables that must be standardised: the alphabetic linkage variables (name of the economic unit and street) and the tax number. Each one of these classes has its own standardisation. The standardisation and homogenisation process of each of these linkage variable classes is described below.

***Alphabetic linkage variables***

The standardisation and homogenisation of the alphabetic linkage variables consists of different stages. The end target is to obtain two new fields: *est_var** and *cod_var**, assuming that the alphabetic linkage variable to standardise is *var**.

The new *est_var** field contains the standardised literal of the original linkage variable *var**, while the new *cod_var** code contains a numerical code that identifies the value of the linkage variable. Therefore, a group of names of economic units or streets whose original value are different but have been standardised are represented by the same code *cod_var**.

The different standardisation and homogenisation processes developed in the linkage program are outline below:

- ***Processing the letter ñ***

  As the administrative record files may have been generated using different character codes in the text editors or programs, certain characters such as the accented vowels or the letter ñ usually appear badly coded. This macro corrects the appearance of the symbols /, # and ¥ instead of Ñ.

  The macro studies all the words that contain the symbols /, # and ¥ and analyses if they may be typing errors that in reality correspond to the letter Ñ. Care has to be taken as the symbols /, # and ¥ may occur and must not be corrected by an Ñ.

  This type of errors are avoided by only correcting those words were the new word formed by replacing the symbol /, # and ¥ by Ñ in the original word exists in any other element.

- ***Debugging and assimilation of characters***

  This stage aims to homogenise the linkage variables as far as possible so that the two texts that correspond to a single value do not appear different due to typing errors.

  The following debugging tasks are therefore carried out:

  - Punctuation marks, such as the full stop, comma or hyphen, are eliminated.
  - If the alphabetic linkage variable is the name of a company or establishment, the company acronym characters are eliminated.
  - Accents are eliminated from the vowels.
  - Non alpha-numerical characters, such as parenthesis, asterisks, pound signs, etc, are eliminated.
  - Articles and prepositions that do not provide relevant information to the alphabetic linkage variable are eliminated.
  - The following characters or groups of characters liable to error due to written or phonetic similarities are considered as a single grapheme.

| Characters or groups of characters that are represented by the same grapheme | | |
|---|---|---|
| Y | I | |
| TX, TS, CH | TZ | |
| K | C | If they are before the vowels A, O, U |
| K | QU | If they are before the vowels E, I |
| N | M | If they proceed the constants B, P |
| V | B | |
| Ñ | N | |
| GU | G | |
| Z | C | If they proceed vowels E, I except when the Z is proceeded by a T |

- ***Match words that are different by a letter***

At this stage, pairs of words are studied that are only differentiated by a letter, with this being a specific letter. The following table contains the pairs of letters that are analysed in the program.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| C | K | K | C | K | QU | QU | K |
| C | TZ | TZ | C | L | LL | LL | L |
| C | Z | Z | C | M | N | N | M |
| C | X | X | C | Q | QU | QU | Q |
| C | Q | Q | C | R | RR | RR | R |
| C | QU | QU | C | S | X | X | S |
| G | J | J | G | S | TZ | TZ | S |
| I | J | J | I | S | Z | Z | S |
| I | LL | LL | I | TZ | X | X | TZ |
| J | X | X | J | TZ | Z | Z | TZ |
| K | Q | Q | K | X | Z | Z | X |

If when changing in a word a letter for its relevant pair, it coincides with another word existing in the set of the files, they are considered to be the same and are therefore assigned the same numerical code.

- ***Assigning numerical codes***

  There is a list of all the words of the alphabet linkage variable processed together with their standardised form and their numerical code. In this phase, the relevant standardised and numerical codes are assigned to the original value of the alphabetic linkage variable in each of the files to be linked.

***TAX NUMBER***

This standardisation phase perfectly classifies and validates all the tax codes used in Spain. A field of 9 alpha-numerical characters is analysed and a numerical value is returned for each type of code analysed where all the positive values (greater than zero) indicate that the tax code is correct.

It is assumed that the tax code is correct when the control character is the corresponding one. The control character is a function of the figures or letters that make up the tax identification and possibly the positions that they occupy. The purpose of including it is to detect and so avoid transcription and typing errors of that identification. Therefore, the control character that corresponds to the code is calculated and it is checked it if matches.

The values that can be generated by the programmed function to validate the Tax Number control character are shown below:

| **Type** | **Unknown** | **Tax Number (individual)** | **Tax Number (Company)** | **Tax Number (Foreigners)** | **Temporary tax number (Company)** |
|---|---|---|---|---|---|
| Correct: | | 1 | 2 | 3 | 4 |
| Incorrect: | 0 | -1 | -2 | -3 | |

The framework complies with the following specifications of Spanish legislation:

Decree 2423/1975, of 25 September

Royal Decree 338/1990, of 9 March

Royal Decree 1624/1992, of 29 December that amends 338/1990

Royal Decree 155/1996, of 2 February

Order of 3 July 1998, which amends the Annex of Decree 2423/1975

Royal Decree 1065/2007, of 27 July

Order EHA/451/2008, 20 February 2008

Order INT/2058/2008, 14 July 2008

The validation procedures for each tax code are described below:

- ***Tax number of individuals***

  In general, the number of the Spanish Identification Document (DNI) is used as the tax identification number for Spanish individuals, while the Foreigner Identification Number (NIE) is used for foreign individuals. Both are assigned by the Ministry of the Interior.

  The identification number begins with X, Y or Z in the case of foreign individuals.

  Furthermore, the tax office assigns an identification key that begins with the letters K, L, or M, depending on the case, for Spaniards under 14 years old without DNI or residents abroad that do not have an NIE:

  The DNI number has eight figures, while the numbers for the other cases of Spanish or foreign individuals have seven (once the initial letter K, L, M, X, Y or Z has been eliminated). The control character is always a letter in all cases.

  The following table describes the different tax numbers for individuals:

| **Type** | **Format** | **Comment** |
|---|---|---|
| DNI | Eight numbers + control digit | Spanish with national identity document assigned by the Ministry of Interior |
| NIF K | K + 7 numbers + control digit | Spaniards under 14 |
| NIF L | L + 7 numbers + control digit | Spaniards resident abroad without DNI |
| NIF M | M + 7 numbers + control digit | NIF that the Tax Office grants to foreigners that do not have a NIE. |
| NIF X | X + 7 numbers + control digit | Foreigners identified by the Police with a foreigner identification number (NIE) assigned before 15 July 2008. |
| NIF Y | Y + 7 numbers + control digit | Foreigners identified by the Police with a NIE assigned after 16 July 2008 (Order INT/2058/2008, Spanish Official Gazette of 15 July) |
| NIF Z | Z + 7 numbers + control digit | Letter reserved for when the "Y" number have been exhausted for foreigners identified by the Police with a NIE. |

The method used to calculate the control character is described below:

*DNI*

The result of dividing the number made up of the 8 numbers of the DNI by 23 and the relevant letter is assigned according to the following table:

| | | | | | |
|---|---|---|---|---|---|
| 0 | T | 8 | P | 16 | Q |
| 1 | R | 9 | D | 17 | V |
| 2 | W | 10 | X | 18 | H |
| 3 | A | 11 | B | 19 | L |
| 4 | G | 12 | N | 20 | C |
| 5 | M | 13 | J | 21 | K |
| 6 | Y | 14 | Z | 22 | E |
| 7 | F | 15 | S | 23 | T |

*NIF X, Y, Z*

The X is replaced by 0, the Y by 1 and the Z by a 2 and then the same procedure is used as for a standard DNI.

*NIF K, L, M*

Its control character is calculated as if it were a NIF of an individual (described below).

- ***NIF of bodies corporate and entities in general***

The Tax Identification Number is assigned by the tax office to bodies corporate and to the entities without legal capacity – commercial association, institutions, ventures, etc. – and consists of nine characters, with the ninth being the control character (a digit or a letter).

The first character reflects the legal status, that can be A, B, C, D, E, F, G, H, J, N, U, V and W for companies and commercial entities and P, Q, R and S for religious congregations and public administration entities and institutes.

The following table describes the values that can be taken depending on their legal status:

| **Letter** | **Legal status** | **Control character** |
|---|---|---|
| A | Limited companies | Numerical |
| B | Limited liability companies | Numerical |
| C | General partnership | Numerical |
| D | Partnerships | Numerical |
| E | Joint ownership arrangements | Numerical |
| F | Cooperatives | Numerical |
| G | Associations | Numerical |
| H | Homeowners associations | Numerical |
| J | Civil corporations, with or without legal status | Numerical |
| P | Local corporations | Alphabetic |
| Q | Public entities | Alphabetic |

| | | |
|---|---|---|
| R | Religious congregations and institutions | Alphabetic |
| S | Entities of the Administration of the State and of the Autonomous Communities | Alphabetic |
| U | Joint ventures | Numerical |
| V | Other types not defined by the other keys | Numerical |
| N | Foreigner companies | Alphabetic |
| W | Permanent establishments of non-resident entities in Spain | Alphabetic |

The following steps are performed to calculate the control letter or digit (remember that this calculation is used to calculate the control character of a NIF of individuals and entities in general and for the NIF of individuals that begin with the letters K, L and M). The seven figures are then placed in positions 2 to 8, in the order that they are written. Let $a_1, a_2, \ldots, a_7$, then

1. Let $A = a_2 + a_4 + a_6$ be the sum of the digits with pair subscript

2. Let $B = b_1 + b_3 + b_5 + b_7$ where $b_i =$ sum of the digits $2xa_i$, $i = 1,3,5,7$

3. Let $C = A + B$, $E =$ last digit of $C$ and $D = 10 - E$ (if $E = 0$, then $D = 0$)

4. If the control character is a digit, then it is *D*. However, if the control character is alphabetic, the letter corresponding to the value of D in the following table:

| *D* value | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 0 |
|---|---|---|---|---|---|---|---|---|---|---|
| Control character | A | B | C | D | E | F | G | H | I | J |

## Probability calculation.

When comparing the linkage values of a pair of records, there can be three different types of indicators:

$$\gamma = \begin{cases} \gamma_1 & \equiv & the values coincide and are\ j-th \\ \gamma_2 & \equiv & the values don't coincide \\ \gamma_3 & \equiv & a value is absent \end{cases}$$

where $j = 1, \ldots, m$ indicates a specific value of the $m$ different values that the linkage variable in question can take.

The following probabilities are calculated at this stage:

- Probability that the values of the linkage variable coincide and are equal to $j$ -th possible value of the linkage variable, given that the pair of records represents the same economic unit ($m(\gamma_1)$).

- Probability that the values of the linkage variable coincide and are equal to $j$ -th possible value of the linkage variable, given that the pair of records does not represent the same economic unit ($u(\gamma_1)$).

- Probability that the values of the linkage variable do not coincide, given that the pair of records represents the same economic unit $m(\gamma_2)$).

- Probability that the values of the linkage variable do not coincide, given that the pair of records does not represent the same economic unit $u(\gamma_2)$).

- Probability that some of the values of the linkage variable are absent, given that the pair of records represents the same economic unit $m(\gamma_3)$).

- Probability that some of the values of the linkage variable are absent, given that the pair of records does not represent the same economic unit $u(\gamma_3)$).

The detailed calculation of these probabilities is available in the documentation quoted in the bibliography.

## Blocking.

The blocking criteria stated by the user in the program are then performed. Furthermore, the final weight is calculated for each of the pairs of elements for the set that select the blocking criterion.

According to the blocking criteria that the user has stated, the relevant macro is fun to finally obtain a set of data with those pairs of elements that comply with some of the blocking criteria.

- **%blockingNIF**. Selects the pairs of elements for which the NIF of the company coincides.

- **%blockingCP**. Selects the pairs of elements for which the post code coincides. In this case, there may be more than one post code. Therefore, if there is more than one post code linkage variable, the blocking is performed for the coincidence of all the variables of that type. That is, the pairs of elements are listed where each and every one of the post code variables coincide.

- **%blockingCalle**. Selects the pairs of elements for which the text of the street coincides. In this case, there may be more than one street text linkage variable. Therefore, if there is more than one street text linkage variable, the blocking is performed for the coincidence of all the variables of that type. That is, the pairs of elements are listed where each and every one of the street text variables coincide.

- **%blockingNTHMUN**. Selects the pairs of elements where the first word of the name of the codified companies coincides, along with the province and the municipality of the company.

As many blocking criteria as wanted can be set up. Setting up more than one is recommended to ensure that a pair of elements that represent the same economic unit are not linked due to errors in the blocking variable.

Once the set of pairs of elements that meet a blocking criterion has been constructed, their total weighting is calculated for each of them as follows:

$$w = \sum_{k=1}^{K} w_k \quad \text{where } w_k = \log \frac{m(\cdot)}{u(\cdot)} = \log m(\cdot) - \log u(\cdot), \quad k = 1, \ldots, K$$

where $K$ is the number of linkage variables stated by the user in the program.

Once the total weighting for all the pairs of elements that met a blocking criterion has been calculated, the user is shown a histogram with the frequency of those pairs to determine the limit weighting.

## Links

Two things can occur for a given pair of elements: that the elements represent the same economic unit or that they are different economic units.

In an ideal theoretical situation, if the two elements represent the same economic unit, all its linkage variables should coincide and therefore have a very high total weighting. However, this does not occur in all cases as there can be differences in some linkage variables due to typing errors or status changes or because some value of some linkage variable is absent.

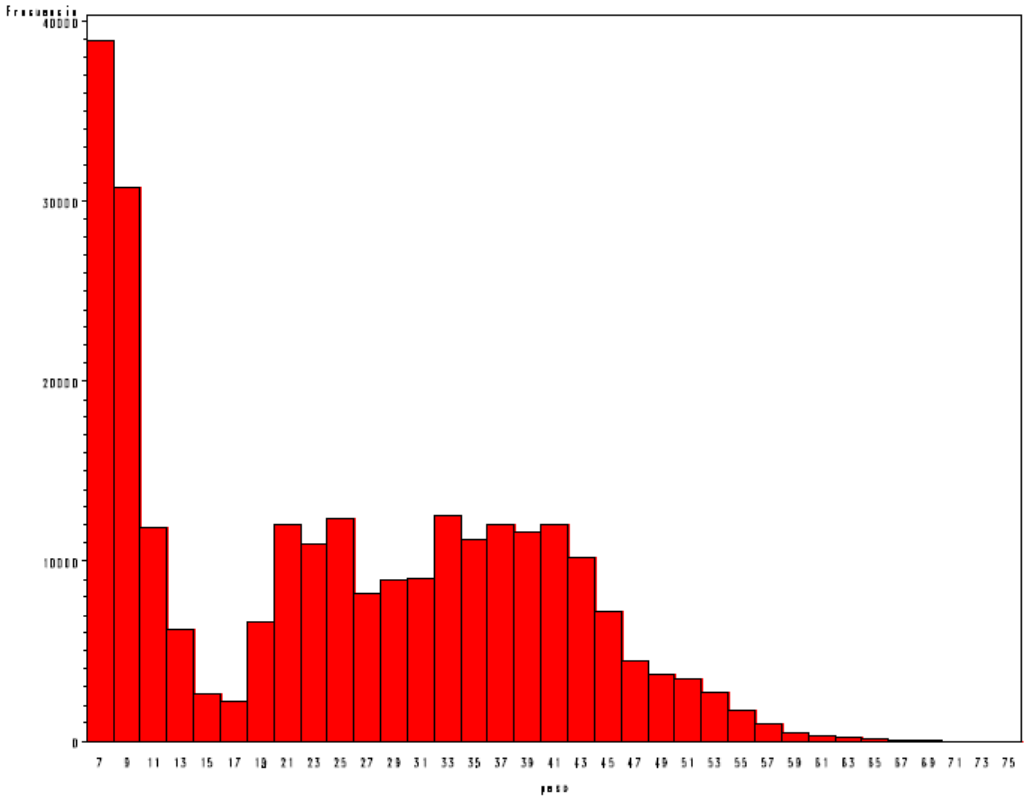
On the other hand, if the two elements represent different economic units, the linkage variables should not coincide and the pair of elements should have a very low weight (even negative). However, there can be accidental coincidences in some linkage variables that have increased the weighting.

Therefore, the histogram would be made up of two clear sections in an ideal theoretical situation. One section to the right around a high weighting would correspond to the pairs of elements that represent the same economic unit, and another section to the left around a low weighting that would correspond to the pairs of elements that represent different economic units.

In practice, these sections are not so clearly differentiated and there values in the intermediary weightings that form a curve.

The user must therefore observe the histogram and establish a weighting based on which all the pairs of elements with weighting higher or equal to the weight established as a limited will be considered to be linked. It is recommended to take the valley of the histogram as the limit weighting.

An example of a weighting histogram is included below:




Once the user has decided the value of the limit weighing, a linear assignation procedure is performed with the pairs of units that have been selected in the blocking phase to establish the elements considered to be linked.

Chapter 5

# ANALYSIS OF THE RESULTS

This section describes the application of the record linkage program aimed at administrative files with the aforementioned economic unit information to the following files: the EUSTAT Economic Activities Directory (DIRAE) and the Social Security file.

## Description of the files

### Economic Activities Directory (DIRAE)

The Economic Activities Directory information is provided by the following three data sets:

- A file that contains the DIRAE data for the local economic activity units.

  The local economic activity unit is a sub-division of the local unit based of activity criteria, exclusive from DIRAE.

- A file that contains the DIRAE data for the corporate units.

  The corporate units are both bodies corporate whose existence is recognised by the law independently from the individuals or institutions that own them or are members of them and the individuals who, as the self-employed, exercise an economic activity.

- A file that contains the DIRAE data for the local units.

  The local unit corresponds to a company or a part of a company (workshop, factory, warehouse, offices, mine, deposit) located in a topographically delimited place. The economic activities are carried out at that place or part of it by one or several people (including the case, of part time work) - except where there are exceptions – on behalf of a single company.

### Social Security

The information of the Social Security file is provided by two data sets:

- A file that contains Social Security data on entrepreneurs

- A file that contains Social Security data on contribution units

The Social Security file is updated every quarter and therefore 4 different files are received in a calendar year. The file used in record linkage is a file that calculates all the variants of a record during a year. In other words, when an establishment has varied any of its characteristics, that record is added to the file and there may be establishments that have even been quadrupled.

## Analysis of the files

The step prior to the linkage of the administrative files is the construction of the data sets that are going to be linked. Therefore, a single set of data is constructed for each source with all the information required for the linkage, that is, the keys that identify each unit in the record and all the common variables to both records used for the linkage. In our case, the **ula.sas7bdat** data sets are constructed using each DIRAE local unit and **uco.sas7bdat** with the information of the Social Security contribution units.

The following table shows the variables to be considered in the linkage:

| DIRAE | | | SOCIAL SECURITY | | |
|---|---|---|---|---|---|
| BODY CORPORATE: | UJA_CIF | TAX NUMBER (COMPANY) | ENTREPRENEUR | EMP_CIFDNI | TAX No. /ID No. |
| | UJA_NOMBRE | Name | | EMP_NOMBRE | Name |
| | UJA_PROV | Province | | EMP_TH | Province |
| | UJA_MUN | Municpality | | EMP_MUN | Municpality |
| | UJA_CP | Post code | | EMP_CP | Post code |
| | UJA_CALLE | Street text | | EMP_T_CALLE | Street text |
| LOCAL UNIT | ULA_TH | Province | CONTRIBUTION UNIT | UCO_TH | Province |
| | ULA_MUN | Municpality | | UCO_MUN | Municpality |
| | ULA_CP | Post code | | UCO_CP | Post code |
| | ULA_CALLE | Street text | | UCO_T_CALLE | Street text |

As can be seen, there are few linkage variables available. The *CIF/DNI* variable is a linkage variable that univocally identifies the company. The *Name* variable is also a very powerful discriminator, even though it is likely to contain many errors. On the other hand, the other location variables (*Province, Municipality, Post Code and Street text*) are not very representative.

Apart from the discriminating power of each linkage variable, it is also important to bear in mind that not all the elements of the administrative record contain information for all variables. The following table shows the number of elements that are absent in each of the variables:

| DIRAE (ula.sas7bdat) | | | SOCIAL SOCIAL (uco.sas7bdat) | | |
|---|---|---|---|---|---|
| 200,675 records | | | 554,177 records | | |
| **Variable** | **NMISS** | **PCTMISS** | **Variable** | **NMISS** | **PCTMISS** |

| | | | | | |
|---|---|---|---|---|---|
| UJA_CIF | 0 | 0% | EMP_CIFDNI | 265 | 0.05% |
| UJA_NOMBRE | 0 | 0% | EMP_NOMBRE | 1 | 0.00% |
| UJA_PROV | 0 | 0% | EMP_TH | 143673 | 25.93% |
| UJA_MUN | 0 | 0% | EMP_MUN | 143673 | 25.93% |
| UJA_CP | 18 | 0.01% | EMP_CP | 140996 | 25.44% |
| UJA_CALLE | 179 | 0.09% | EMP_T_CALLE | 143673 | 25.93% |
| ULA_TH | 0 | 0% | UCO_TH | 112027 | 20.22% |
| ULA_MUN | 1437 | 0.72% | UCO_MUN | 112027 | 20.22% |
| ULA_CP | 5958 | 2.97% | UCO_CP | 101856 | 18.38% |
| ULA_CALLE | 1653 | 0.82% | UCO_T_CALLE | 104213 | 18.81% |

As can be seen in the previous table, even though the DIRAE is very comprehensive, the location variables of the Social Security file have many absent values, which greatly hinders the linkage.

Tests have been performed using different blocking criteria. The NIF blocking has been used in all cases as it is a variable that provides a great deal of information and it is also quite well recorded.

This is known thanks to the programmed macro to study the validity of the control character. The following table contains the results obtained in that analysis:

| **Result of ValidarNIF.sas** | **DIRAE** | **Social Security** |
|---|---|---|
| -3 | 8 | 1 |
| -2 | 38 | 0 |
| -1 | 55 | 6 |
| 0 | 140 | 265 |
| 1 | 102109 | 424931 |
| 2 | 92616 | 112602 |
| 3 | 5706 | 16372 |
| 4 | 3 | 0 |
| **Total** | 200675 | 554177 |

It should be remembered that the negative values indicate that the control character is incorrect and the 0 indicates that the NIF format is not the standard one (that is, it does not have 9 characters). Therefore, the table shows that:

- With regard to the DIRAE local units, 0.05% of the establishments have an incorrect NIF and 0.07% have an incomplete NIF format. There are also 3 establishments with a temporary NIF (NIF validation code=4) that cannot be validated.

- With regard to the Social Security contribution units, 0.001% have an incorrect NIF and 0.05% an incomplete NIF format.

## Results of the linkage

The following tables contains a summary of the results obtained when running the linkage program with the different blocking criteria.

| **Blocking criteria** | **Limit weighting** | **Run Time** | **Linkages** | | |
|---|---|---|---|---|---|
| | | | **Phase 1** | **Phase 2** | **Total** |
| NIF | 20 | 10 hours | 14748 | 111948 | 126696 |
| NIF + Calle | 28 | Aborted after 39 hours | -------- | --------- | --------- |
| NIF + NTHMUN | 32 | 6 hours | 14748 | 80040 | 94788 |

During phase 1, the pairs of elements were linked where all the linkage variables coincided, that is, it is a direct linkage. In phase 2, the linkage was carried out of the pairs of elements selected by linear assignation algorithm among those that have a greater weighting than the limit weighting, that is, they are the pairs of elements linked by the probabilistic procedure.

Chapter 6

# CONCLUSIONS

Using administrative information in the statistics environment is not always straightforward. Administrative records are usually not conceived or designed for statistical purposes and using the information that they provides usually requires a preliminary processing.

This project has turned the social security record into a useful instrument for statistical use when constructing and maintaining the DIRAE: The DIRAE is considered to be a fundamental reference directory that is essential for providing reliable and first-rate statistics. The Social Security information is used to update such important variables as the number of employees.

The results of applying the probabilistic linkage method together with the preliminary processing show more than acceptable percentage of record linkage, taking into account the quality of the linkage variables. As we have described in the technical notebook, many of the auxiliary variables used in the linkage have missing values that hinder any type of processing. The linkage of a “direct” mode would not reach even 10% of the establishments.

The general idea of the project sought to link any two records of establishments, as had been done with individuals. After studying the records of available establishments, it was seen that the specific features of each prevented a joint processing. It was therefore decide to program a sequential macro and modulate in such a way that the modules can be reused and adapted for any future establishment linkage.

This study can be improved as administrative records can be improved band by studying new channels to perform an effective blocking, that enables the file to be divided into sufficiently small blocks for the running time (number of comparisons) to be acceptable and sufficiently broad not as to loose quality and possible data pairs.

A quality control was performed to assess the results obtained by selecting different random samples of pairs of linked units in the second phase of running the linkage programme. During phase 1, the pairs of economic units were linked where all the linkage variables coincided. During the second phase, the pairs of records that determine the probabilistic methodology used were linked. It was checked that the linkage was correct or at least logical for the majority of the records linked. For example, single establishments of the same company could be linked even though important linkage variables were absent.

The result of this project has therefore been more than satisfactory and will enable us to have a tool that will improve the quality the EUSTAT economic activities directory.

Chapter 7

# BIBLIOGRAPHY

[1] EUSTAT

AUTOMATIC METHODS OF RECORD LINKAGE AND THEIR USE IN EUSTAT. http://en.eustat.es/documentos/datos/ct_15_i.pdf

[2] I.P. FELLEGI AND A.B. SUNTER

A theory for record linkage. *Journal of the American Statistical Association, 64(328):1183-1210, 1969*

[3] JARO, M.A.

Advances in record linkage methodology as applied to matching the 1985 Census of Tampa, Florida. Journal of the American Statistical Association.

[4] BLAKELY, T. AND SALMOND, C.

Probabilistic record linkage and a method to calculate the positive predictive value. Internation Journal of Epidemiology (2002).

[5] AYESTARAN, MARINA AND LEGARRETA, LEIRE.

Applying methods of record linkage for census validation in the Basque Statistics Office. Instituto Vasco de EstadÌstica (2004).

[6] WINKLER, WILLIAM E.

Matching and Record Linkage. Bureau of the Census (1993).

[7] CHRISTEN, PETER AND CHURCHES, TIM.

Febrl – Freely extensible biomedical record linkage. Australian National University (2003).

[8] YANCEY, WILLIAM E.

An Adaptive String Comparator for Record Linkage. U.S. Bureau of the Census, Statistical Research Division (2004).