AUTOMATIC METHODS OF RECORD LINKAGE AND THEIR USE IN EUSTAT

EUSKAL ESTATISTIKA ERAKUNDEA
BASQUE INSTITUTE OF STATISTICS

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

This publication introduces the work carried out over the last few years on the application of new technologies to the production of statistics in the field of record linkage.

The framework within which such work has been carried out is the Basque Statistics Plan 2005-2008, in relation to R +D + I in statistic methods, the purpose of which is to research and apply new statistic-mathematic technologies in statistical operations. The spirit that stimulates these studies is based on excellence management, as defined by means of a process of continuous improvement.

The growing demand for statistical information, the limited resources of statistical information offices and the objective of avoiding an excessive response load on the people polled makes it necessary to make efficient use of the information proceeding from censuses and polls, as well as that from administrative data.

The application of automatic record linkage methods was initiated in EUSTAT in the nineteen-nineties. First, deterministic methods were applied, and as from 2002 probabilistic methods have been developed, which are the main object of this publication.

I hope this publication is useful for all those who have an interest for this concrete field of statistics.

Vitoria-Gasteiz, March 2007
Josu Iradi Arrieta
Director General.

AUTOMATIC METHODS OF RECORD LINKAGE AND THEIR USE IN EUSTAT.

SUMMARY

Record linkage refers to the task of identifying records that correspond to the same entity in two or more different files when unique identifiers are not available. In EUSTAT[1] we have continuously been working with linkage procedures, but some time ago we started to study new techniques, like the probabilistic record linkage methods, which is what is described in this technical notebook. In this document, first there is a brief introduction to the concept of record linkage. We then establish the situation of EUSTAT when it initiated the study on this new technique. After that we describe the methodology used, which is fundamentally based on the article titled "*A theory for Record Linkage*" by Ivan P. Fellegi and Alan B. Sunter. We then describe the computer program developed at EUSTAT to carry out linkage operations based on an automatic probabilistic method, following the directives established by means of the afore-mentioned methodology. After dealing with the program, some of the linkage applications that have been carried out in our Institute are examined. Lastly, we offer the conclusions we have come to after the study.

[1] EUSTAT wishes to thank Leire Legarreta and Laura Otero for the excellent research work carried out initially by the former and later by the latter, within the framework of the statistical-mathematical methodology research grant promoted by the Basque Statistics Office

Index

Chapter 1

Introduction

Introduction and objectives

Record linkage is defined as the use of algorithmic techniques with the purpose of identifying pairs of records proceeding from two different files when there are no single identifiers available.

Historically, such a task was given to administrative personnel, who had to manually review the lists of records, obtain additional information when there was no such information or when it was contradictory, and take linkage-related decisions following previously established rules. Generally speaking, the lists were sorted by name or by some of the address characteristics in order to facilitate the process. If, for example, a name contained an unusual typographic variation, its correspondence may not have been found. If the files were very big, the lists would then be divided in various sheets, and thus, sometimes, some pairs were lost. In spite of thorough training, the decisions that were taken were not always consistent. All of this brought about the study of computational techniques in order to carry out this linkage task in an automatic way.

There are various techniques to carry out this work. They can be classified in two main groups: deterministic methods and probabilistic methods.

The **deterministic record linkage** seeks exact matches and mismatches on one or more linkage variables for two or more files. For example, it is possible to simply use the Identity Card Number field that is common to two files. However, codification mistakes incurred when registering this variable in any of the files bring about the loss of real *matches* (*match*: a comparison pair of two records in different files of the same person).

The **probabilistic record linkage** method uses the information from a higher number of variables and takes into account all the information provided by any of the matches or mismatches on the linkage variables. Even though it takes into account all the variables, each one of them will have a different decisive power. For example, a coincidence in the Social Security number is more likely to be a match than a coincidence on gender. Also, coincidences on unusual values of a given linkage variable (for example, the surname Galzarsoro) are more indicative than a match on more common values (for example, the surname García).

The main objective of this publication is to present the work carried out in Eustat on record linkage techniques based on probabilistic methods. It was decided within the Basque Statistics Office to study new record linkage methods in order to try to improve some of the results obtained by means of deterministic methods that were being used up to that moment. Starting from the work carried out by Fellegi and Sunter, a SAS language linkage program has been designed, which allows for the linkage of two files using probabilistic techniques.

This notebook describes first the methodology followed for the further development of the linkage program. It also details this program, reviewing each of the macros it is

composed of. There will be an explanation on the functions of each of the macros and on the position they occupy within the linkage program. Then there is a report on the applications of this program carried out by the Basque Statistics Office and mention is made of the results obtained. Lastly, there is a report on the conclusions EUSTAT has reached on this type of probabilistic procedures.

Description of the Project

In order to achieve these objectives, work was initiated on a thorough study of the available documentation on the different record linkage methods, and mainly that on probabilistic methods. Once sufficient information thereon had been gathered, it was decided to implement it and for this purpose a series of macros were developed in SAS language which were to automatically execute the linkage between two files with corresponding records on individuals.

Once the program had been developed, it was executed with various files in order to study its feasibility and to be able to compare the results obtained by means of this method with those obtained using procedures based on deterministic techniques that assigned certain concrete degrees of weight to the variables.

Background

Before starting the study on probabilistic methods, linkage procedures based on deterministic techniques were already being implemented within EUSTAT. Such techniques distinguished the entity from the variables assigning to each a different degree of weight.

One of the applications of this linkage procedure was carried out with the purpose of associating codes corresponding to persons (Basic Population Unit codes), which are stored in the Statistic Population Register and the territory codes (Territorial Emplacement Unit codes) from the Territory database to the records of the Municipal Census.

Only the person codes are associated to the Municipal Census Person Movement files. This process is carried out in various steps and modules, depending on the file.

The various modules are described below:

1) Linkage in the strictest sense.

For the case of the Municipal Census, the intention is to look for person and territory codes in the Statistic Registers of Population and Territory, respectively. The Emplacement Territorial Unit code is sought firstly through the Basic Population Unit code.

The records from the Municipal Census will be associated with the registers from the tables of the Statistic Population Register that contain data such as place of birth, gender, postal address, name, surnames and National Identity Document number, using the following criteria:

- Coincidence in National Identity Document Number. The National Identity Document number is used without the letter. Those with the highest frequency are excluded so as to avoid many duplicates cropping up, as these would be considered to be incorrect National Identity Document numbers.

- Coincidence in name and surnames. The name and surnames are previously normalised and a function of alpha code generation is applied to them. These alpha codes are equivalent literals that facilitate the treatment of those variables. The three alpha codes are used simultaneously for the association.

- Coincidence in postal address, once the coincidences in date of birth and gender have been checked.

Each record in the table of Movement of Persons in the Municipal Census has to be related with the Statistic Population Register by means of the Basic Population Unit code. The process to be carried out is the same as that used for the Municipal Census, without taking into account the comparisons with the postal address, as this does not appear in the Movement of Persons in the Municipal Census file.

Weights are then calculated, and by means of comparisons on the diverse variables, the coincidences are scored and only the ones with the highest value are taken into account.

2) Linkage in a less strict sense.

Since certain valid relations may not be detected because the first linkage criteria may have been too strict, it was decided to make some of these conditions more flexible.

The first adjustment was that, for registers that were not linked with the first process, a search was to be made for matches through combinations of name and the two surnames: name and first surname, name and second surname, first and second surname. The weights used in this process are the same as those in the previous module.

Another criterion is applied to those left without linking after this adjustment; according to this new criterion a search is made for matches of two of the three elements in their date of birth: day and month, day and year, month and year.

With these less-strict linkages, what happened was that sometimes relations that had a total score above the limit were considered valid, albeit due to the coincidence of isolated elements. There could therefore be two erroneous situations: on one hand, that it is the same person but with a different code, and on the other hand, that there are different persons with the same code. Work is now being carried out in the improvement of these cases by means of the normalisation of names and surnames.

Chapter
2

Methodology

The theoretical model in which the program designed in Eustat for the linkage of records is based is the model presented by Fellegi and Sunter in the article "A *theory for Record Linkage*" [1] in 1969. The basics of this model are as follows.

Theoretical Model

Probabilistic record linkage methods carry out the comparison of records from two different files. The files to be compared are noted as A and B and their elements are noted as a and b, respectively.

Both files are supposed to have common elements and therefore, the objective of the linkage is to recognise, from amongst all the AxB pairs that could be formed, those that refer to the same person, object or entity. That is to say, the objective is to divide the set

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

in the union of two disjoint sets

$$M = \{(a, b) \mid a = b, a \in A, b \in B\} \quad (1)$$

and

$$U = \{(a, b) \mid a \neq b, a \in A, b \in B\} \quad (2)$$

which are named sets of *matches* and *non-matches,* respectively.

Each unit of population has some associated characteristics, such as, name, surnames, age or address. It is necessary to identify those records that refer to one single person, object or entity. However, the process of file creation could introduce errors or imprecision (in the form of codification, transcription and typing errors, typographical or phonetic variations, lost data, etc.) in the records generated. As a result of these errors, two members of A and B that do not refer to the same individual may generate identical records and, more frequently, two identical members of A and B could produce different records. The records corresponding to the members of A and B are noted as $\alpha(a)$ and $\beta(b)$ respectively.

It is also supposed that the individuals that have been registered come from random samples selected from A and from B that are noted as As and Bs. The possibility that As = A and Bs = B is not excluded. From now on, to simplify notes, subscript s is eliminated.

The first step when trying to pair records is to compare them. The result of the comparison is a set of codes, which are codified in statements of the following types: “the name is the same in both records”, “the name is the same and it is Pedro”, “the name is different”, “the name field is missing in one of the records” or “there is an agreement in the quarter of the town where the addresses are, but not in the street”. Formally, the comparison vector is defined as a function vector of records α(a) and β(b) in the following manner:

$$\gamma[\alpha(a), \beta(b)] = \{\gamma^1[\alpha(a), \beta(b)], \ldots, \gamma^k[\alpha(a), \beta(b)]\} \quad (3)$$

It can be seen that γ is a function defined on AxB. It can be transcribed as γ(a, b), γ(α, β) or simply γ. The set of all the possible realizations of γ is called *comparison space* and is noted as Γ.

During the linkage operation process, γ(a, b) is observed and it becomes necessary to decide if (a, b) is a match, (a, b) Є M (this decision is called *link* and is noted as A1) or if it is a non-match, (a, b) Є U (this decision is called *non-link* and is noted as A3). However, there may exist situations for which it is impossible to take one of these two decisions for specific levels of error, and thus a third decision is allowed, and is noted as A2, and which is called *possible link*.

In these conditions, *the linkage rule* L is defined as an application of the comparison space Γ on the set of random decision functions D = {d(γ)} where

$$d(\gamma) = \{P(A1|\gamma), P(A2|\gamma), P(A3|\gamma)\};\ \gamma \in \Gamma \quad (4)$$

and

$$\sum_{i=1}^{3} P(Ai \mid \gamma) = 1 \quad (5)$$

In other words, for each observed value of γ, the linkage rule supplies the probabilities of taking each of the three possible decisions.

It is necessary to consider the levels of error associated with each linkage rule. Let us assume that a pair of records [α(a), β(b)] is selected randomly to be compared according to a certain probabilistic process. The resulting vector of comparison γ[α(a), β(b)] is therefore a random variable. The conditional probability of γ , given that (a, b) Є M as m(γ), will be:

$$m(\gamma) = P\{\gamma[\alpha(a), \beta(b)] \mid (a,b) \in M\} = \sum_{(a,b)\in M} P\{\gamma[\alpha(a), \beta(b)]\}.P[(a,b) \mid M] \quad (6)$$

similarly, the conditional probability of γ, given that (a, b) Є U is noted by u(γ). Therefore,

$$u(\gamma) = P\{\gamma[\alpha(a), \beta(b)] \mid (a,b) \in U\} = \sum_{(a,b)\in U} P\{\gamma[\alpha(a), \beta(b)]\}.P[(a,b) \mid U] \quad (7)$$

There are two types of errors associated to this linkage rule. The first happens when comparing pairs of records that do not correspond with *matches*, it is then decided to assign them as *links.* This has the following probability:

$$P(A1|U)=\sum_{\gamma\in\Gamma}u(\gamma).P(A1|\gamma) \quad (8)$$

The second type of error takes place when a *match* is compared and is considered as a *non-link*, and this has the following probability:

$$P(A3|M)=\sum_{\gamma\in\Gamma}m(\gamma).P(A3|\gamma) \quad (9)$$

It is said that a linkage rule in space Γ is the linkage rule in error levels μ, λ (0<μ<1, 0<λ<1) and it is noted as L(μ, λ, Γ) if

$$P(A1 \mid U) = \mu \quad (10)$$

and

$$P(A3 \mid M) = \lambda \quad (11)$$

It is said that the linkage rule L(μ, λ, Γ) is optimal if the relation

$$P(A2 \mid L) \leq P(A2 \mid L') \quad (12)$$

remains true for any L'(μ, λ, Γ ') among all the linkage rules that verify the aforementioned relations.

It is observed that, according to this definition, any optimal decision rule is that which maximises probabilities of adopting positive comparison arrangements (i.e, decisions A1 or A3) subject to certain fixed levels of error. This seems a reasonable decision, since the adopting of decision A2 will require quite costly manual linkage operations. Furthermore, it seems that if the probability of A2 is not small, the linkage process will doubtlessly not be much use.

It is not difficult to realise that with certain combinations of μ and λ, the set of linkage rules that satisfy (10) and (11) is empty. Therefore, only those combinations of μ and λ which can satisfy equations (10) and (11) simultaneously for some set D of decision functions as that defined by (4) and (5). When this happens, it is said that (μ, λ) is a *pair that is admissible in terms of levels of error*.

The authors are proposing an optimal linkage rule. For this purpose, they start by defining a single order within the finite set of all the possible realisations of γ in the following way: if a certain value of γ is such that both m(γ) and u(γ) are equal to zero, then the probability of this γ happening is equal to zero, and therefore it is not necessary to include it in Γ. Following this, an arbitrary order is assigned to each γ for which m(γ) > 0, but u(γ) = 0.

The rest of the γ are sorted in such a way that the corresponding sequence of m(γ) / u(γ) is monotonously decreasing. When the value of m(γ) / u(γ) is the same for over one γ, then the order assigned will be arbitrary.

The ordered set {γ} is indicated by subscript i; (i = 1, 2,…, N_Γ), where N_Γ is the number of points of Γ and is noted as ui = u(γi); mi = m(γi).

Let us suppose that (μ, λ) are an admissible pair of levels of error in that neither are too big. Let us suppose n, n' are whole numbers and that

$$\mu = \sum_{i=1}^{n} u_i \quad \lambda = \sum_{i=n'}^{N_\Gamma} m_i \quad 0 < n \le n' < N_\Gamma$$

The following is defined

$$T_\mu = \frac{m(\gamma_n)}{u(\gamma_n)}$$
$$T_\lambda = \frac{m(\gamma_n')}{u(\gamma_n')} \qquad (13)$$

Then Fellegi and Sunter demonstrated that the best linkage rule in levels of error (μ, λ) was given in the following form:

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

In many applications, it could be possible to tolerate sufficiently high levels of error to eliminate the possibility of action A2. In this case, n and n' or T_μ y T_λ are considered in such a way that the average set of γ in its former form is empty. In other words, each pair (a, b) is localised either in M or in U. In fact, this is the decision we have adopted in our linkage program, and consequently we have established the sole limit $T_\mu = T_\lambda$.

Some instances of simplification

In practice, the set of different values that γ can adopt is so large that the estimation of the corresponding probabilities m(γ) and u(γ) can become quite impracticable. Therefore, it is necessary to carry out certain suppositions that will simplify the distribution of γ.

It is assumed that the components of γ can be re-sorted and grouped in such a manner that $\gamma = \left(\gamma^1, \gamma^2, \cdots, \gamma^k\right)$ and the vectorial components are all statistically independent from each of the conditional distributions. Then,

$$m(\gamma) = m_1(\gamma^1) \cdot m_2(\gamma^2) \cdot \cdots \cdot m_k(\gamma^k) \qquad (14)$$

$$u(\gamma) = u_1(\gamma^1) \cdot u_2(\gamma^2) \cdot \cdots \cdot u_k(\gamma^k) \qquad (15)$$

where m(γ) and u(γ) are defined by (6) and (7) respectively and

$$m_i(\gamma^i) = P(\gamma^i \mid (a,b) \in M)$$
$$u_i(\gamma^i) = P(\gamma^i \mid (a,b) \in U)$$

As a matter of notation simplicity the tendency is to write $m(\gamma^i)$ and $u(\gamma^i)$ instead of the technically more precise expressions $m_i(\gamma^i)$ and $u_i(\gamma^i)$. For example, in a comparison of records corresponding to people, γ^1 could include all the components of comparison referring to surnames, γ^2 all those referring to addresses. The very components γ^1 and γ^2 are vectors in themselves; and the subcomponents of γ^2 can represent, for example, the codified results of comparing the different components of the addresses (town, street, number, etc.). If two records match (that is to say, in reality, they represent the same person), there could be a disagreement in configuration due to errors. The errors in the name, for example, are assumed to be independent from the errors in the address. If two records do not match (that is to say, when in reality they represent two different individuals) then the assumption is that an accidental coincidence in the name, for example, is independent from an accidental coincidence in the address.

In other words, it is assumed that γ^1, γ^2,..., γ^k are distributed in a conditionally independent manner. It should be pointed out that nothing is being assumed about the non-conditional distribution of γ.

It is clear that any increasing monotonous function of m(γ)/u(γ) could equally well be used as a statistic test for the purposes of the linkage rules.

In particular, it will be advantageous to use the logarithm of this ratio and define

$$w^k(\gamma^k) = \log m(\gamma^k) - \log u(\gamma^k) \tag{16}$$

We can then write

$$w(\gamma) = w^1 + w^2 + \cdots + w^k \tag{17}$$

and use w(γ) as our statistic test in the understanding that if u(γ)=0 or m(γ)=0 then w(γ) = +∞ (o w(γ) = -∞) in the sense that w(γ) is higher (or lower) than any given finite number.

γ^k is supposed to be able to assume n_k different configurations $\gamma_1^k, \gamma_2^k, \cdots, \gamma_{n_k}^k$. We can then define

$$w_j^k = \log m(\gamma_j^k) - \log u(\gamma_j^k) \tag{18}$$

It is effective for the intuitive interpretation of the linkage process that the weights are defined as positive for those configurations for which $m(\gamma_j^k) > u(\gamma_j^k)$, and as negative for those configurations for which $m(\gamma_j^k) < u(\gamma_j^k)$, and this property is preserved for weights associated to the total γ configuration.

The total amount of configurations (that is to say, the amount of points $\gamma \in \Gamma$) is obviously $n_1 \cdot n_2 \cdot \cdots \cdot n_k$. However, due to the additive nature of the weights defined for the components, it would be sufficient to determine $n_1 + n_2 + \cdots + n_k$ weights. It will then be possible to determine the weight associated with each and using this additivity, notably reducing in this way the cardinal of Γ.

Calculation of weights

One of the key points in the development of the linkage procedure is the calculation of weights m(γ) and u(γ) for each one of the possible γ that could appear. There are various methods proposed; in fact, the very work by Fellegi and Sunter presents two different ways to tackle this matter. For this work a method has been selected that is based in the frequencies with which each of the recorded data configurations appears in the set of files to be linked.

Let us suppose that one of the components of the records of both files is the *surname* field. The comparison of the surnames from both records will produce as a result a component of the comparison vector. This component may be a simple comparison of the type "*the surname coincides*" or "*the surname doesn't coincide*" or "*the surname doesn't not appear in one or both files*".

In any of the two files the surname may be recorded with some type of a mistake. It is assumed that it is possible to draw up a list with all the realisations of the surnames from both files that is free of error and that also contains the number of individuals from the respective files that have each of these surnames.

Let us suppose that the respective frequencies in A and B are as follows:

$$f_{A1}, f_{A2}, \cdots, f_{Am}; \quad \sum_{j=1}^{m} f_{Aj} = N_A$$

and

$$f_{B1}, f_{B2}, \cdots, f_{Bm}; \quad \sum_{j=1}^{m} f_{Bj} = N_B$$

Let us suppose the corresponding frequencies in A∩B are as follows:

$$f_1, f_2, \cdots, f_m; \quad \sum_{j=1}^{m} f_j = N_{AB}$$

And, to be more specific,

- f_{A1} is the frequency with which a concrete value from one of the variables appears in file A
- f_{B1} is the frequency with which this value appears in file B
- f_1 is the frequency with which this value appears both in A and in B

The following note is necessary:

- e_A o e_B the respective probabilities that the value of a variable has been registered erroneously in each of the two files. (It is assumed that the probability of being erroneously recorded is independent from each of the particular values).

e_{A0} o e_{B0} the respective probabilities that the value of a variable has not been recorded in each of the two files. (It is assumed that the probability of not been recorded is independent from each of the particular values).

e_T The probability that the value of a variable appears in a different manner in both files in spite of being correctly recorded in both. (This can happen, for example, if all files were created at different times and the individual changed his or her name).

Finally, it is assumed that e_A and e_B are sufficiently small for the probability of a coincidence between two identical entries, even if erroneous, is insignificant and that the probabilities of being badly recorded, and not recorded or having changed our independent from each other.

The following rules are given for the calculation of m and u corresponding to the following configurations of γ: the variable coincides and is the j-th listed, the variable does not coincide, the variable is absent in some of the files.

$$m(\text{the variable coincides and is the j-th listed}) =$$
$$\frac{f_j}{N_{AB}}(1-e_A)(1-e_B)(1-e_T)(1-e_{A0})(1-e_{B0}) = \tag{19}$$
$$\frac{f_j}{N_{AB}}(1-e_A-e_B-e_T-e_{A0}-e_{B0})$$

$$m(\text{the variable does not coincide}) =$$
$$\left[1-(1-e_A)(1-e_B)(1-e_T)\right](1-e_{A0})(1-e_{B0}) = \tag{20}$$
$$e_A+e_B+e_T$$

$$m(\text{the variable is missing in some of the files}) =$$
$$1-(1-e_{A0})(1-e_{B0}) = \tag{21}$$
$$e_{A0}+e_{B0}$$

$$u(\text{the variable coincides and is the j-th listed}) =$$
$$\frac{f_{Aj}}{N_A}\frac{f_{Bj}}{N_B}(1-e_A)(1-e_B)(1-e_T)(1-e_{A0})(1-e_{B0}) = \tag{22}$$
$$\frac{f_{Aj}}{N_A}\frac{f_{Bj}}{N_B}(1-e_A-e_B-e_T-e_{A0}-e_{B0})$$

$$u(\text{the variable does not coincide}) =$$

$$\left[1-(1-e_A)(1-e_B)(1-e_T)\sum_j \frac{f_{Aj}}{N_A}\frac{f_{Bj}}{N_B}\right](1-e_{A0})(1-e_{B0}) = \quad (23)$$

$$\left[1-\left(1-e_A-e_B-e_T\right)\sum_j \frac{f_{Aj}}{N_A}\frac{f_{Bj}}{N_B}\right]\left(1-e_{A0}-e_{B0}\right)$$

$$u(\text{the variable is missing in some of the files}) =$$

$$1-(1-e_{A0})(1-e_{B0}) = \quad (24)$$

$$e_{A0}+e_{B0}$$

Proportions f_{Aj} / N_A, f_{Bj} / N_B, f_j / N may be taken, in many applications, as if they were the same one. This would be the case, for example, if the two files were supposed to have been extracted from the same population. These frequencies may be estimated directly from the files themselves.

It should be remembered that according to (18) the contribution of the component of a variable to the total weight is log(m/u) and that the comparisons with a higher weight than a concrete number will be considered as *links*, whereas those with lower weights than the concrete number will be considered *non-links*. It is clear from (19-24) that an agreement on a variable will produce a positive weight and the rarer the value of the variable, the higher the weight. In an analogous manner, a disagreement on a variable will produce a negative weight that will decrease with errors e_A, e_B, e_T. If the value of the variable is missing in some of the records, the weight will be zero.

Creation of standardised lists

In order to be able to use the method of calculation of weights as previously described, it is necessary to have a list available with all the error-free configurations of each of the variables that intervene in the process. This does not imply any problems for variables of the “gender" type, for example, as it is known there are only two correct configurations and that any value that does not correspond with these two values is erroneous.

In the case of variables such as name or surname, this task is not so simple. The number of typing errors that such variables have is usually quite high, and thus, before extracting a list of configuration, it is necessary to correct those that have such mistakes.

We are now going to explain the method that has been developed in EUSTAT and which attempts to relate those configurations that have the same original configuration, from all the configurations of one of these name or surname-type variables, but that are different because each configuration may have more than one possible correct form or because there has been an error in some of them.

The idea is as follows: work is started from an original list C with all the configurations of a variable that appear in both files to be linked. The objective is to generate a list of codes D and an application which is denominated *est*, from C in

D, in such a way that two elements from C that are equivalent to the same value have the same image through *est*. As has been said before, two elements from C may have the same image, either because the same value may be represented by means of different configurations or because when registering some of them, there has been an error.

This process therefore consists of, on one hand, relating those configurations that, although both represent the same value of a variable, are different, albeit correct. Examples of the situation are names such as ENEKOITZ-ENEKOIZ, LEZURI-LEXURI or JAVIER-XABIER. Therefore, it must be guaranteed that the image through *est* of these configurations is the same.

For this purpose, certain standardisation criteria have been followed, some of which are as follows:

- Errors produced by the occurrence of the symbol ¥ instead of letter Ñ.
- Dashes, full stops and commas are eliminated.
- Values such as he HASN'T GOT ANY, THERE ARE IN ANY, WE DON'T KNOW OF ANY, WE DON'T KNOW are substituted by “missing”.
- Accents or diereses are eliminated from those vowels that appear with any type of accent or diereses.
- Numbers (1, 2, 3, 4, 5, 6, 7, 8, 9) are eliminated.
- Certain characters (‘, ´, `, ¨, ^, *, “, º, ª, #, Ç, ¥) are eliminated.
- Zeros are substituted by the letter ‘O’.
- Certain particles (D, DE, DEL, DA, DI, DO, L, LA, LAS, EL, LOS, Y) are eliminated
- Certain diminutives are translated.
- Characters or groups of characters with phonetical or graphical similarity are identified.
- Pairs of names and surnames that coincide in all the letters except one are studied, and an analysis is carried out to determine if they could proceed from the same configuration. These are studied separately because there are graphic characters that, although they do not have an evident phonetic or graphic similarity in some cases, may eventually represent the same character.

On the other hand, those configurations for which some type of error has taken place must be detected. Once they have been identified, their image through the *est* application will be the same as that in the correct configuration from which they proceed. Thus, configurations such as FERANNDO, FERNNDO and FERANANDO correspond through the *est* application with the same FERNANDO code.

The first obstacle that can be found when considering this task is how to detect those configurations in which an error has taken place. In order to come to logical

deductions, we started from the frequencies with which the configurations appear in the files to be linked. For each cЄC, $freq_0(c)$ is to represent the frequency with which c appears in both files.

Let us suppose c is a string of characters with a length of n that appears in some of the files to be linked and has been, supposedly, correctly recorded.

This is denoted as

e = Prob(correctly recorded character)

1-e = Prob(erroneously recorded character)

Given the fact that the string of characters c has n characters, the probability of that string of characters having been correctly recorded (as has been supposed) will be of:

P(0 errors | length (c) = n) = e^n.

If the number of individuals of the original population of which the variable takes value c is freq(c), given that $freq_0(c)$ represents the number of cases in which configuration c takes place in the files that are being considered, then:

freq(c) * P(0 errors | length (c) = n) = $freq_0(c)$

Therefore, freq(c) can be estimated as:

freq(c) = $freq_0(c)/e^n$

Analogously, for each correct configuration of C can be determined the total amount of those configurations for which one error, two errors, etc have taken place during its transcription.

$$P(1\,\text{error}\,|\,\text{length}\,(c)=n)=\binom{n}{1}e^{n-1}(1-e)=n\cdot e^{n-1}\cdot(1-e)$$

It is thus possible to estimate the total amount of configurations that can have been generated starting from c with an error like:

$$freq_1(c)=P(1\,\text{error}\,|\,\text{length}\,(c)=n)\cdot\ freq(c)=\frac{freq(c)\cdot n\cdot(1-e)\cdot e^n}{e}$$

In the same manner it is possible to estimate the total amount of configurations that could have been produced as from c with two errors, as follows:

$$P(2\,\text{errors}\,|\,\text{length}\,(c)=n)=\binom{n}{2}e^{n-2}(1-e)^2=\frac{n\cdot(n-1)\cdot e^{n-2}\cdot(1-e)^2}{2}$$

$$freq_2(c)=P(2\,\text{errors}\,|\,\text{length}(c)=n)\cdot\ freq(c)=\frac{freq(c)\cdot n\cdot(n-1)\cdot(1-e)^2\cdot e^n}{2\cdot e^2}$$

In this situation the objective is to relate those erroneous configurations from a variable to the correct ones, as we know the estimate that has been carried out on the total number of erroneous configurations that can appear. Those cases with an amount of errors above 2 are not considered, because it is supposed that the probability with which they happen is very small. In any case, if desired, they could be taken into account following the reasoning carried out for the cases with 1 or 2 errors.

The starting point is to compare the configurations that may be erroneous with those that can be supposed to be correct, and generate a list in which both appear related in those cases in which the number of possible errors that separate some configurations from others is below or equal to 2.

An intuitive idea is to suppose that those configurations that take place with high frequencies in the files are correct, taking into account that it is not very probable that the same error is committed too often. Therefore, the limit is established, depending on which set C is divided into the following subsets:

C_1 = {cЄC | freq(c) > lim}

C_2 = {cЄC | freq(c) ≤ lim}

The values in C_2 are compared with those in C_1 with the objective of relating them in those cases in which the number of errors that separate them is low. In this way, a list is drawn up in which each individual in C_2 may appear as related with none, one, or several individuals from C_1 and vice versa.

Then, from all these relations a selection is made of those that are finally going to be linked with each other. For this purpose, starting from the previously detailed cases, the total amount of configurations related with a given configuration can never be above the previously estimated value.

In this manner, each element from C_2 has been related with one or several elements from C_1 and vice versa. From all this set of relations, a final subset is to be selected in which each element from C_2 is related to, at most, one from C_1, and the total amount of configurations related with a given c from C_1 is not above $freq_1$ + $freq_2$.

$$freq_1 + freq_2 = \frac{freq(c) \cdot n \cdot (1-e) \cdot e^n}{e} + \frac{freq(c) \cdot n \cdot (n-1) \cdot (1-e)^2 \cdot e^n}{2 \cdot e^2}$$

In order to carry out this selection the following criterion is established: from the starting list, a subset is selected from which each element from C_2 can only be related with one from C_1, given that a possibly erroneous configuration can only proceed from a correct original configuration. In case the configuration from subset C_2 appears related with more than one from C_1, the one that appears most frequently in the files is selected, or, alternatively, the one that differs in the least amount of errors.

Once a list of these characteristics has been generated, it is probable that each element from C_1 is related with more than one from C_2. Therefore, from amongst

them all, those that are more likely to proceed from C_1 are selected, after having made some type of error, bearing in mind that the total number of configurations related with one of these cannot be above the estimate that has been carried out.

The process is repeated until all the elements from C_2 have been related with some from C_1 or all the elements from C_2 have been related with a number of configurations that is equal to the estimate carried out.

Apart from this method that is used to try to correct errors in the variables of the name or surname type, later in the program another type of considerations on these variables is carried out for those pairs that are not linked in the first tries and for which a somewhat more thorough study is carried out.

First, a check is made to see if any of the names is a diminutive of the other taking into account certain concrete and previously known cases.

A revision is also carried out, both for names and surnames, to see if a compound value appears in one of the files whereas in the other file only one of the components appears.

Another way of establishing the relation between two names or two surnames is by means of the Jaro string comparator. This calculates the number of common characters in the two strings, their lengths and the number of transpositions to calculate the measure of similitude that goes from 0.0 and 1.0.

The transpositions are determined by pairs of common characters that are out of place. Then, the value of the Jaro character string comparer for strings α and β is given by the following formula:

$$S_J = \frac{1}{3}\left(\frac{common\ ch.}{length(\alpha)} + \frac{common\ ch.}{length(\beta)} + \frac{common\ ch. - transpositions}{common\ characters}\right)$$

It can be seen that if the strings are identical (common characters = length (α) = length (β) and transpositions = 0), then $S_J = 1$.

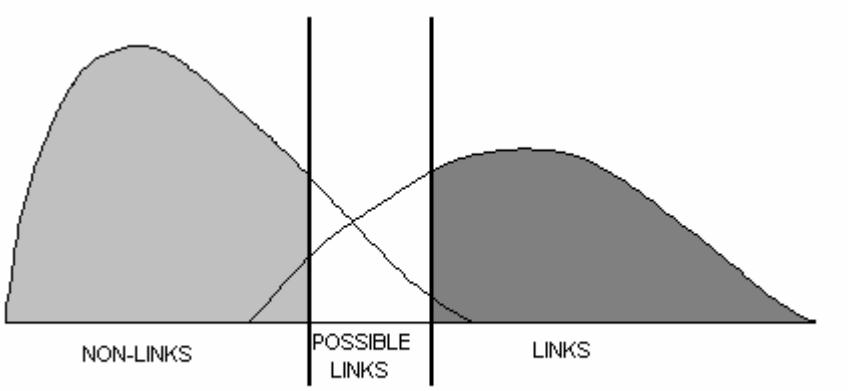
Another consideration that is made when the two surnames of the individual are used as linkage variables is to check if there has been an interchange. As explained below, when observing the results that were obtained after the execution of the linkage program, there were hundreds of pairs of records that corresponded to the same individual and that did not link because the order of the surnames had been changed.

The last contrast has to do with the peculiarity of our situation. In the Basque Autonomous Community there are two official languages, Spanish and Basque. Therefore, there are certain name equivalences that can produce different configurations in two files for a single occurrence. What then as to be done is to create a database with various equivalences of Spanish - Basque names and each pair of names is checked to see if they are included in this list. If so, it is then established that there is a relation between them.

After all these appreciations, the weights are calculated once again and those that now have a weight that is superior to the limit previously established by the user are selected.

Method to establish the limit

Having specified all the relevant γ_j^k specifications and having determined their associated weights $w_j^k; k = 1,2,\ldots,K; j = 1,2,\ldots,n_k$ it is necessary to establish the limit values T_μ and T_λ corresponding to the given μ and λ. These two T_μ and T_λ limits divide the pairs of records into three zones.



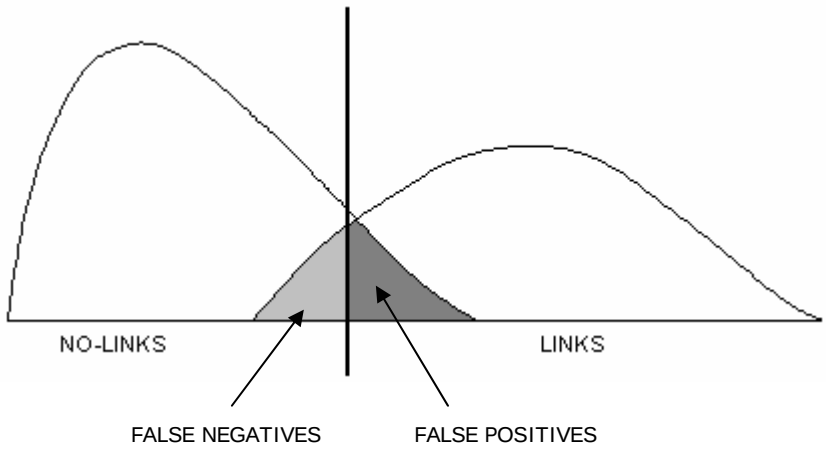
The area to the left of the limit T_μ corresponds to those pairs of records that are going to be classified as *non-links*. In the area to the right of T_λ are the pairs of records that are to be filed as *links*. And in the intermediate zone is where the pairs of records that are called *possible links* are filed and that need manual revision in order to decide what group to classify them in.

When taking these decisions two types of errors are being committed. First, an error is being committed when establishing as a link two records that do not belong to the same individual. And another error is being committed when establishing as a *non-link* of pair of records that in reality correspond to the same individual. These errors may be quantified as follows:

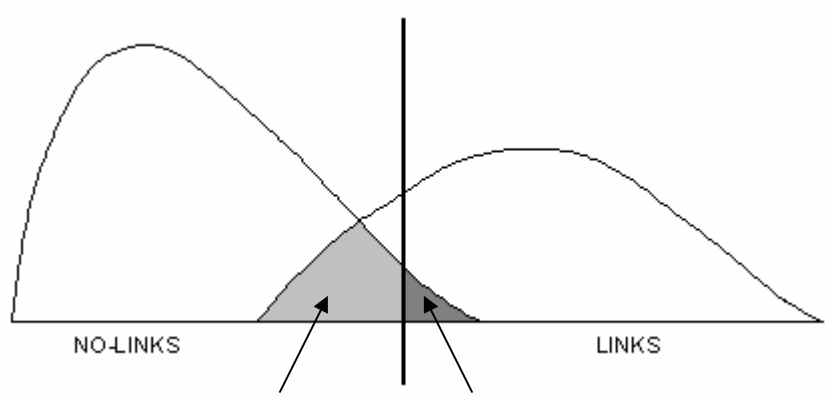
$$\mu = P(A1 \mid U) = \sum_{\gamma \in A1} u(\gamma)$$ Proportion of *link* pairs in U.

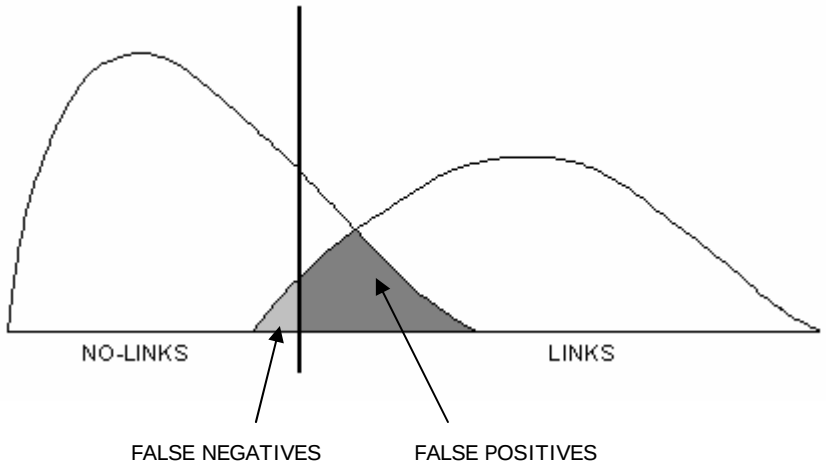
$$\lambda = P(A3 \mid M) = \sum_{\gamma \in A3} m(\gamma)$$ Proportion of *non-link* pairs in M.

In order to avoid manual processing of the pairs of records that are between both limits, in EUSTAT it was decided to adopt a single limit $T = T_\mu = T_\lambda$.



Equalling both limits, as the value of the single limit increases, the number of false negatives will also increase whereas the number of false positives will also decrease. In an analogous manner, as the value of the single limit decreases, the number of false negatives is to decrease whereas the number of false positives will increase.

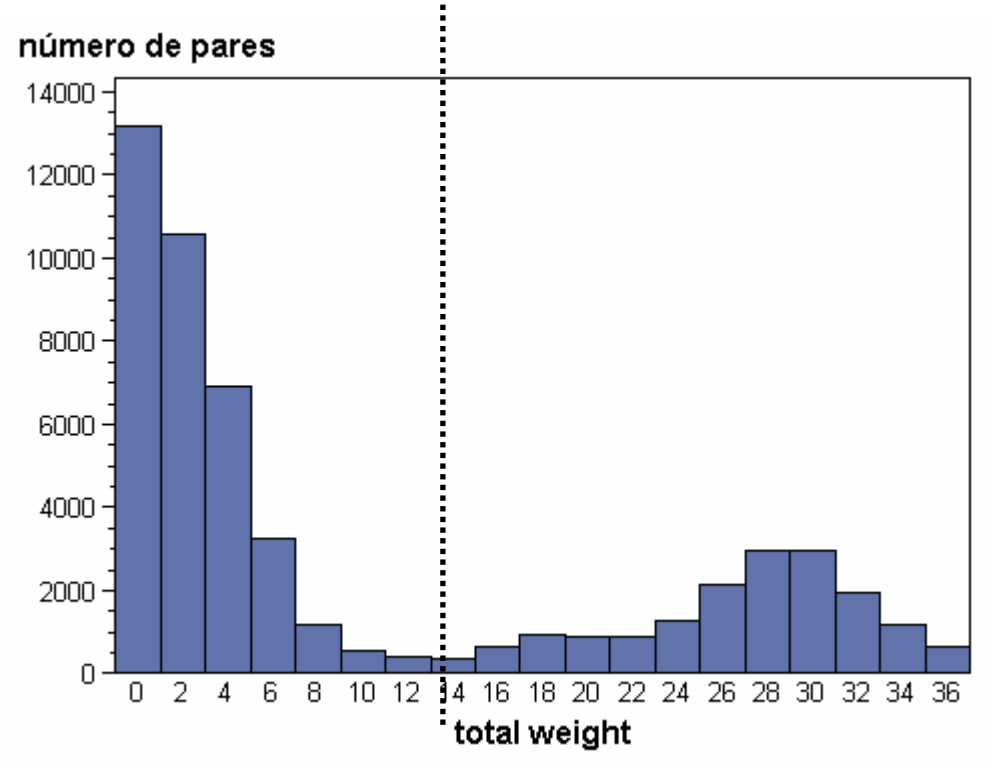




A simple method with which to adequately establish this limit is based on the observation of the histogram of frequencies of all the calculated total weights. The histogram provides valuable information to aid in the decision-making process, based

on the form the histogram adopts for the various weight values. That is, the pairs corresponding to the same individual should have a high weight whereas pairs of records not corresponding to the same individual will have a low weight (what is more, a negative weight).

In practice, this is not exactly so due to errors in the variables and casual coincidences in the variables. In any case, the frequency histogram will have a similar form to the following graph:



This is how the value of the limit is established and all those pairs of records with a weight over that value are declared *links* whereas those pairs of weights below that limit are declared *non-links*.

Blocking

An ideal state of affairs when linking two files A and B would be to compare each one of the records from A with all the records from B, but this, for two medium-sized files, would mean a total of |AxB| comparisons, which is an excessively high number.

Therefore, what is then to be done is to establish a scheme that only extracts pairs of records that are reasonably susceptible of corresponding with a *match*. This process is called *blocking*.

In this manner, the intention is to reduce the amount of comparisons the program is to carry out by sorting out the files into mutually exclusive and thorough blocks

designed to increase the proportion of observed *matched pairs* while reducing at the same time the number of pairs of records to be compared.

Blocking is generally implemented by means of classifications of the two files over one or several variables. For example, if both files are classified by postal code, the pairs that would be considered would be those in which the postal code coincides. The pairs of registers that do not have the same postal code recorded will not be analysed and would therefore be automatically considered *non-matches*.

Obviously, if for each record its correspondence were searched in the whole of the other file, the probability of finding a true *match* would be higher as no record is excluded from consideration.

However, the cost of such a process would be excessive. Thus, *blocking* is a compromise between the computational cost (examining an excessive number of pairs of records) and the proportions of false *non-matches* (classifying pairs of records as *non-matches* if the records are not members of the same block). In order to make this inconvenience more tolerable, it is advisable to use more than one *blocking* variable and that such variables are independent from each other.

It is also very important to choose correctly the variable selected as *blocking* criterion. A component of an ideal *blocking* would contain a considerable amount of values that are quite uniformly distributed. That is to say, none of the blocks into which the file is divided should be excessively big nor of the amount of blocks should be too small for the savings in execution time to be real. They should also be variables that have a low probability of registering errors (that is to say, a high weight) because if an error has been committed in that variable the affected records will be placed in different blocks and they would therefore not be compared and consequently could not be linked.

We are now going to describe here which are the types of *blocking* implemented in our linkage program. It should be mentioned that such criteria have been programmed in a modular manner so as to facilitate future inclusion of other different criteria.

Date of birth criterion.

This criterion establishes blocks according to the recorded date of birth variable. That is to say, it will only study those pairs of records where the date of birth coincides.

Records for which the value of the variable is missing will not be considered, as there is no reason that indicates that believing that this variable does not appear in one file implies it will not appear in the second file either.

This is one of the main reasons why it is advisable not to use only this *blocking* criterion, especially if there is a high probability that the date of birth variable is empty.

Key criterion.

This criterion uses a key-type variable for the classification of files such as, for example, the variable *housing identifier*. This variable does not identify each

record, as there may be several individuals with the same *housing identifier* since they may all reside in the same house.

Text code 1 criterion.

This criterion is based on a certain character code that is formed as from the individual's surnames. For each file two new variables are created which store the combination of the first letter and the two following consonants of each surname. The final *blocking* variable is constructed by concatenating the two former variables in such a way that an alphabetical code is created as from the two surnames. This series of characters is used instead of full surnames to avoid the errors that may have happened when recording the values of this variable.

Text code 2 criterion.

This criterion is based on a certain character code that is formed as from the individual's surnames. For each file two new variables are created which store the first three letters from each surname. The final *blocking* variable is constructed by concatenating the two former variables in such a way that an alphabetical code is created as from the two surnames. As in the previous criterion, this series is used instead of full surnames to avoid the errors that may have happened when recording the values of this variable.

Chapter 3

SAS Programming

Below is the general description of the linkage program developed in EUSTAT detailing the SAS macros it is composed of.

Input Variables

Before executing the program, the user has to introduce certain necessary data for its correct functioning. The data that have to be defined are as follows:

Folder localisation.

The user must define three variables with the localisation of the following folders.

1 – A folder in which are the two files that are to be linked. This folder is also where the program result files are going to be kept. There will be three such files: one with the pairs of linked records and another two with the registers that it has been impossible to link from each of the two files.

2 – A folder in which the data files created during the execution of the program are stored.

3 – A folder in which the auxiliary data files that are created and eliminated during the execution of the program are stored.

File names.

The user also has to indicate the names of the two files to be linked.

Names and types of the linkage variables.

For each of the linkage variables that are going to be used it will be necessary to introduce in the program the name that the mentioned variable has in each of the files and the type of variable it is; the following types are possible:

0 – Record identification code.

1 – Name-type variable.

2 – Surname-type variable.

3 – Gender-type variable.

4 – Date-of-birth-type variable.

5 – Year-of-birth-type variable.

6 – Number-only code-type variable.

7 – Number-and-letter-code-type variable.

It is compulsory to define at least one identifying code-type variable to be able to distinguish the records, together with one name-type, one surname-type and one gender-type variable, respectively.

Gender indicators.

It is also necessary to tell the program which are the values that indicate masculine and feminine genders.

Types of *blocking* criteria.

It is also necessary to decide which *blocking* criteria have to be used to execute the program.

Possible types are:

1 – Date of birth.

2 – Code.

3 – Text code 1.

4 – Text code 2.

Each one of these criteria has been detailed in the corresponding section of the previous chapter.

Parameters.

For each linkage variable it is necessary to establish the value of certain parameters that are necessary for the calculation of weights. These are as follows:

- Probability of a value from that variable having been badly recorded in the first file.
- Probability of a value from that variable having been badly recorded in the second file.
- Probability of a value from that variable being different in both files in spite of having been correctly written into both.

Besides, values should also be assigned to the following general variables:

- Probability of a character having been correctly recorded.
- Value used to establish when a name or surname is frequent or not.

Limit.

It will also be necessary to establish a limit value that discriminates which pairs of records are to be linked. At a given moment, the execution of the program stops, a histogram is displayed and based on that, the limit value is established and introduced.

SAS Macros

Name of the macro	Description
ejecución	This is the main macro of the program. This macro is called upon twice. The first time, files and linkage variables are studied, together with the standardisation and homogenisation of the variables, the procedure to assign values according to the previously-described standardised lists is carried out, weights are calculated, blocks are established and the distribution of weights of the pairs of records that belong to a block are studied. This distribution is displayed in the form of a histogram. The program returns the histogram to the user, from whom it establishes the value of the limit it will use to distinguish between *links* and *non-links*. The macro will be called upon for the second time to carry out the last step consisting of studying the pairs of records, the weight of which is below the limit but very near it.
paso1_analisis	The libraries that are going to be worked on are created. File1 is distinguished from file2 (file1 will be the smallest, the one with the least number of registers, as it is supposed that this file is intended to be linked completely). An analysis is made of the available variables and checks are carried out to make sure there are no incongruence in the data introduced by the user.
recuento	The number of records in a data file is calculated.
paso2_confeccion_listas	This macro calls up on other macros to carry out a process of homogenisation and standardization of the name and surname-type variables, given the fact that these variables are susceptible of containing errors and besides, one single configuration may appear represented in different ways.

enies	This corrects those configurations in which the symbol ¥ appears as substituting the letter Ñ.
estandarizar	This standardises the name or surname-type variables according to certain criteria such as the elimination of spelling or punctuation signs, or the identification of characters with a phonetic or graphic similarity.
emparejar	Some characters may be used with the same meaning in spite of not having a relevant phonetic or graphic similarity. Therefore, the macro studies the configurations that only differ in one character and evaluates if they proceed from the same origin.
comunes_y_no_comunes	This macro studies the frequency with which each of the configurations appears in the set of two files to be linked. After this study, it is possible to distinguish the values as common and non-common according to the frequency. It is convenient to note that those values that have had the same code as a result after the previous standardisation measures will be considered as of the same configuration.
buscar_errores	Values with low frequencies (non-common values) are supposed to be erroneous configurations proceeding from some of the frequent values (common values). A measure of distances is then established and pairs of values are selected, formed by a common value and a non-common value, and which have a sufficiently small distance to be able to consider they could proceed from the same configuration.
confeccion_listas	From all the pairs extracted from the previous macro, those that are in accordance with certain frequency and error-possibility criteria are selected and there are sufficient indicators to consider that they correspond to pairs of names or surnames proceeding from the same original configuration.
indicador_sexo	Once the standardisation of the name-type variable has been carried out, this macro identifies if each value belongs to a man or to a woman.

paso3_asignar_estandarizados	This macro studies those configurations of names that appear both in the list of men and in the list of women and decides on whether to include them in one table or the other according to the frequencies of such configuration in both listings. Once this work has been carried out it restarts the codes assigned to each configuration of the men's and women's name variable and the surname variable. There will be two series of codes, one for surnames on the other both for men's and women's names, but in this order; that is to say, there will be a concrete code number that will be stored in a variable and for which any code inferior to it will correspond to a masculine name and any code above it will correspond to a feminine name.
est_final	This is an auxiliary macro that is called upon from macro paso3_asignar_estandarizados and which re-initiates the codes so that they are a consecutive series of natural numbers.
paso4_calculo_pesos	This macro calculates the weights m and u described in the methodology.
paso5_blocking	All the *blocking* criteria established by the user are executed one by one. All the weights recovered from these *blocking* procedures are put together and the histogram is drawn up (with weights above 5, so that the histogram is visibly clearer) from which to take up a position on the value of the limit variable.
blocking_fecha_nacimiento	This macro carries out a *blocking* of all the records for coincidence in the date-of-birth variable.
blocking_clave	This macro carries out a *blocking* of all the records for coincidence in a key variable.
blocking_codigo_texto1	This carries out the *blocking* of the records discriminating coincidences of certain strings of characters composed by the first letter on the two following consonants of each of the surnames.
blocking_codigo_texto2	This macro carries out a *blocking* of the records discriminating coincidences of strings of characters composed by the first three letters of each of the surnames.

fusion	This is an auxiliary macro called upon from each of the macros corresponding to a *blocking* criterion and that carries out the calculation of the m and u weights for all the configurations of each variable.
paso6_posibles_links	With this macro a selection is made of the pairs of records that are considered *matches*. First it takes those with a total weight above the established limit and after that it carries out a somewhat more thorough study in which some pairs, which are near the limit but do not go above it, are selected.
asignación_lineal	Given a group of pairs of records that are susceptible of proceeding from the same origin, this macro takes biunivocal assignments, and when there are several possibilities, it chooses the one with the highest weight.
iml	This macro carries out a procedure of lineal assignment (lsap) for 1-1 assignments.
comparador_strings	This macro calculates the Jaro string comparator to measure the distance between two configurations.

.

Chapter
4

Applications

Below is the description of some of the linkage applications[2] that have been created in EUSTAT and that use probabilistic linkage methods. The first few of them are particular cases in which some details of the probabilistic method were introduced. The linkage programme described in the previous chapter was developed as a consequence of their results and of the existing relevant literature on the matter. The rest of the applications are the consequence of the results obtained after executing that program.

Marriage Statistics and Statistical Population Register

The first tests for the implementation of record linkage methodologies using probabilistic techniques within the Basque Institute of Statistics were carried out with the purpose of linking the file that contains information on personal data of title-bearers of marriages and the file that contains the data from each Basic Population Unit, that is to say, the Statistical Population Register.

Linkages that had been carried out up to that moment obtained a low linkage percentage, given that in those files no single identifier like the National Identity Card existed nor did any other type of identifying code.

Firstly, two data files are taken from the Oracle database, one with marriages and dated after January 1996 and the other with data from all the records corresponding to each Basic Population Unit.

It was then necessary to determine which linkage variables were to be used. For this purpose, the formats in which the variables appear were studied. It was then possible to see quite easily that some variables were certainly not useful for the pairing process, given that some of them appeared as not reliable, or were reappointed in a different manner in each file. In other cases, as in the gender variable, previous re-codification was needed in order to use that variable in the process, as in one file such information was recorded as ‘H’ for women and ‘V’ for men and in the other as ‘6’ and ‘1’ respectively.

The variables that were going to be used in this case as linkage variables were:

Name
First surname
Second surname
Gender
Date of Birth

[2] All linkage applications mentioned in this text were created within EUSTAT and under the protection of statistic secrecy (as regulated in Law 4/1986 on Statistics of the Autonomous Community of the Basque Country), which affects the whole of the personnel involved in its development.

In total there was a set of 64,384 records of marriages of which it was possible to link, using the aforementioned procedures, 30,802, that is to say, 47.84% of the records.

Firstly the records that directly coincided in all the linkage variables were linked by means of a SAS procedure in order to apply the probabilistic linkage procedure to those records in which the coincidence was more complicated to justify. The records that were not linked during this approach were fed into the probabilistic link program using two different blocking criteria. The first criterion was through the date of birth and the second through the initials of name and first and second surnames. In total, it was finally possible to link 60,055 records, which represented 93.28% of the original file.

Register of Companies and Directory of Economic Activities

Another of the first cases in which probabilistic linkage procedures were applied was the adaptation of the general linkage programme of two files to link companies in the Register of Companies with the Eustat Directory of Economic Activities. The objective of this linkage was to be able to use economic information provided by the Register of Companies as a source of updating for the Directory of Economic Activities, and, in this way, increase the coverage of the operation, since the Directory of Economic Activities was the framework up to which such information was to be fed.

The first step to carry out this task was to obtain a single file from the Register of Companies. This was the union of 6 equally structured files, 2 per Historical Territory. In each territory one of the files was digitally originated whereas the other was obtained after a scanning process and then an Optical Character Recognition (OCR) process. From the resulting file, duplicate records are eliminated (conserving the most recent one).

The following process was to cross it with the Directory of Economic Activities in the CIF/DNI (Fiscal Identification Code/National Identity Document) field. As a result of that process, a file was obtained with the records that had not directly coincided with the directory. It is to this file with the records that had not coincided directly that the linkage process was applied, once the CIF/DNI and the names had been standardised. Of all the records analysed, the new records coinciding with the directory and that fulfil certain criteria were recovered.

The variables that are considered linkable are:

CIF/DNI
Name
Postal code
Telephone.

Different linkages are carried out in accordance to the following variables:

- CIF/DNI, Name, Postal code and Telephone.
- CIF/DNI and Name.
- Name.
- CIF
- Telephone
- Postal Code

In the cases of linkages through CIF and telephone number, there was also the requisite that there should be a certain similitude in the names, apart from the values of the standardised CIF and telephone numbers coinciding. This similitude was calculated by means of the execution of a SAS macro designed for this purpose and that considered four instances of possible relations:

1. Compressed names coincide.

2. The Jaro string comparator is over 0.85.

3. One of the company names is an abbreviation or an acronym of the other name.

4. The number of letters in common words in both names is over that of a certain value.

The macro did not only select pairs of records whose names fulfilled some of the aforementioned criteria but also marked which of the four criteria was the one that was fulfilled for each pair of records of establishments.

Up to now, nothing of what has been described for this exploitation refers to probabilistic methods. These are used in the last linkage, that of the postal code. Evidently, linking two companies because the postal code coincides does not have any logic since many companies will have been registered under the same postal code. In this case, what was done was that, apart from checking for a certain similitude between the names of the companies as done in previous linkages, a certain weight was assigned to each pair of records depending on the frequency of the name.

That is to say, those records for which there was a postal code coincidence and the names of which were related by means of some of the criteria previously listed. Each of these pairs is assigned a weight. In this way it is directly related to the frequency with which words common to both names appear in both files.

Lastly, the pairs of records that fulfil the following are linked:

(1) Coincidence in CIF (standardized).

(2) Coincidence in name (standardized).

(3) Coincidence in telephone number and there is a relation between the names.

(4) The variable named weight is above a certain given value.

(5) If all the words that appear in one of the names are in the other name.

The results were that, of the 30,000 initial records of companies, it was possible to link 22,500 by means of the exact procedure. The 7,500 remaining unlinked records, were studied by means of the procedure described and it was possible to relate approximately 900 of them. That is to say, in total, there was an increase in the linkage percentage from 75% to 78% thanks to this procedure.

Be it fit to mention that many of the records that could not be linked in any of the two phases correspond to the agricultural sector and that their particular conditions make it impossible for them to be linked by any method.

Personal and Family Income Statistic

Another of the experiences of EUSTAT in which certain processes related with probabilistic linkage were used was that carried out on the Personal and Family Income Statistic for the 2001 fiscal year.

Linkage of records in the Personal and Family Income Statistic was carried out between the files received from the Statutory Regional Finance Departments (*Haciendas Forales*) and the Statistical Population Register.

This linkage was conceived as a process of comparison of fields from different sources, detection of equal values and assignment of a linkage weight for each coincidence in order to, later on, calculate a total weight, depending on which corresponding Basic Population Unit code was assigned (from the Statistical Population Register). That is to say, an exact linkage process was applied using certain determined weights.

The novelty in this application is that before the execution of the linkage process, a homogenisation treatment was carried out with the variable *name and surnames* based on the probabilistic methods developed.

Below is a description of the homogenisation treatments used for the variable *name and surnames*, which are indispensable in the linkage process.

Three files are received from the respective Local Statutory Finance Departments with different record designs and a homogenisation treatment is necessary for the variables. The *name and surnames* field is to adjust to a three-subfield scheme: name, first surname and second surname, separated by asterisks.

This is an explanation on which exactly were the problems found when dealing with this variable:

- In some records in which one of the surnames is compound, the *name and surnames* field is erroneous. If the first surname is a compound surname, the last particle of this surname is converted into the second surname, and this goes as the name without being separated with an asterisk from the real name. If the second surname is the compound surname, the last particle thereof overflows into the name field without being separated with an asterisk from the real name.

- Other records adjust to the name, first surname and second surname scheme, separated by blank spaces.

As a basis for homogenisation, databases generated from the names and surnames of the individuals registered in the 2001 Population and Housing Census and in the Survey on the Population in Relation to Activity are used.

The data from names and surnames have been taken from these files, the frequency with which each of them appears in the two files has been calculated and the following databases were created in relation to such frequencies:

Name of the data base	Description
listado_final variables: *variable* *resultado*	This is generated as from the data on names and surnames of individuals from the 2001 Population and Housing Census and the Survey on the Population in Relation to Activity. Each one of the components from the name and surnames fields has been taken into account and they have been assigned a value in the *resultado* (result) field according to the frequency with which they appear. The *resultado* field will always have one of these possible values. 1 = Name 2 = Less frequent name. 3 = Surname 4 = Less frequent surname. 5 = Name and surname
datos_2 variables: *variable* *var1* *var2* *resultado* *n_datos*	This is also generated as from the names and surnames of the individuals from the 2001 Population and Housing Census and the Survey on the Population in Relation to Activity. It was generated in order to try to detect data corresponding to names and/or the surnames composed by two components. Fields var1 and var2 will be the corresponding components of the composed name and/or surnames, n_datos will have a value of 2 to indicate that they are two components and the *resultado* field takes one of the following values: 1 = Compound name 3 = Compound surname
datos_3 variables: *variable* *var1* *var2* *var3* *resultado* *n_datos*	This is analogous to the previous database but with the data corresponding to compound names and surnames composed of 3 components. Fields var1, var2 and var3 will be those components, n_datos has a value of 3 and the *resultado* field takes one of the following values: 1 = Compound name 3 = Compound surname
datos_4 variables: *variable* *var1* *var2* *var3* *var4* *resultado* *n_datos*	This is analogous to the previous database but with the data corresponding to compound names and surnames composed of 4 components. Fields var1, var2, var3 and var4 will be those components, n_datos has a value of 4 and the *resultado* field takes one of the following values: 1 = Compound name 3 = Compound surname

The file that is entered is the file from the Local Statutory Finance Department with the variables *antes* (before), *variable* and *después* (after). The treatment is only applied to the *variable* field which corresponds to the names and surnames.

Firstly, certain small modifications are made for the standardisation of names and surnames (accents, diereses, symbols, numbers, repeated consecutive letters except RR and LL are eliminated, and certain contractions of known names and surnames are translated such as M for MARIA or PZ for PEREZ).

These modifications are carried out in order to facilitate the identification of the names and surnames of the files from the Local Statutory Finance Department with those obtained from the 2001 Population and Housing Census and the Survey on the Population in Relation to Activity, since these same modifications were carried out to them beforehand.

Now the *name and surnames* field from the Finance Department files is separated by words. Those words that are used as linking words for certain compound names and surnames are eliminated (such as DE, LA, LOS, SAN,…). Loops of words are taken and compared with those compound names and surnames that are stored in the listado_final, datos_2, datos_3 and datos_4 databases previously described. If some coincidences are found, the corresponding number of the variable *resultado* and that of the variable *n_datos* are duly stored.

It is necessary to take into account that there may be more than one possible combination when considering compound names or surnames (for example, PEDRO JOSE RUIZ OTALORA may be disassembled as PEDRO*JOSE*RUIZ-OTALORA or as PEDRO-JOSE*RUIZ*OTALORA). Therefore, what is done is to highlight these records by means of a variable that acts as a marker and that indicates that this record is a special case to be studied more carefully.

The corresponding codes are assigned according to the result obtained when searching for the names and surnames in the Finance Department records with the reference databases. The linking words occasionally used in compound names and surnames and which previously had been eliminated are also codified but with symbols instead of numbers.

These changes are as follows:

Particle	Code
DE, DEL, Y	&
Any one-letter word except Y	+
DA, DI, DO, SAN, DOS, SA, SANTA, SANTO, SAO, BON	*
LA, LAS, EL, LOS	$

Lastly, those records that have been considered special by means of a marker variable are taken aside in a database to study their specificities more thoroughly. Depending on the values they have, a series of decisions are taken on which of them to keep.

Final objective is to obtain the result files *_asignados* (assigned) and *_no_asignados* (non-assigned) with the final values of the *name and surnames* field

Once the homogenisation treatment of names and surnames has been carried out, the linkage of files in itself is carried out and in this case this is a determinist linkage.

The following variables intervene in the assignment of weights (with repercussion in the RPF treatment):

- NAME (name alphacode, surname 1 alphacode and surname 2 alphacode).
- RESIDENCE (Territory, Municipality, District, Section, Entity, Street, Number, Floor and Left or Right).
- DNI (8, 7, 6, 5 and 4-digit National Identity Document number).
- DATE OF BIRTH (Day, Month and Year of Birth).

The assignment of the Basic Population Unit code would be carried out once all the weights have been computed, in such a way that the function would eliminate the records of lesser weight, and then the duplicates, in accordance to the weights assigned according to the set of dominions. This is how the Basic Population Unit code with the highest weight would be selected and, at the same time, that code would not have been previously stricken out. With the process of homogenisation described and the determinist linkage procedure it is possible to link over 98% of the records.

Population and Housing Census and Survey on the Population in Relation to Activity

These linkage techniques have also been used as an application for the validation on the 2001 Population and Housing Census of the Autonomous Community of the Basque Country, as from data obtained from the Survey on the Population in Relation to Activity. This operation is carried out every three months with the objective of providing continuous statistical information on the characteristics of the main groups in which the population of the Basque Country can be classified according to their participation in the various economic activities. Then, in order to apply the linkage process, works starts as from two files. One corresponds to the 2001 Population and Housing Census, which contains 2,082,587 records. And the other corresponds to the Survey on the Population in Relation to Activity from the nearest quarter to the date of reference of the Census, and which has 10,831 records.

The fields that are common to both files and that are used as linkage variables are:

Name
First surname
Second surname
Gender
Housing code
Date of birth

The following *blocking* criteria were used:

Date of birth
Housing code
Text code formed by the first letter and the two following consonants from both surnames
Another code formed by the first three letters of the two surnames.

After applying the linkage process the results were that 10,535 records were found in the 2001 Census of Population and Housing out of the 10,831 in the Survey on the Population in Relation to Activity. This means a linkage percentage of 97.27%.

Population Register of Vitoria-Gasteiz and Statistical Population Register

Linkage has also been carried out between files from the Statistical Population Register and the Population Register of Vitoria-Gasteiz. The number of records contained in each of these files is as follows:

Municipal Census of Inhabitants of Vitoria-Gasteiz	233,670
Statistic Population Register	2,456,831

The objective is to localise in the Statistical Population Register all the individuals registered in the Population Register of Vitoria-Gasteiz.

The linkage variables used were:

First surname
Second surname
Name
Date of birth
Gender

And the following *blocking* criteria were used:

Date of birth
Text code1
Text code2

The results obtained in the first approach were:

Number of linked pairs of records	231,646
Number of non-linked records from the Population Register of Vitoria-Gasteiz	2,357
Number of non-linked records from the Statistical Population Register	2,225,188

As a matter of common sense, if a sum was made of the number of linked records with the number of non-linked records in each of the files, the result should be the total number of records from the original files. But this was not the case here.

This is because there are records in the Population Register of Vitoria-Gasteiz that have been linked with more than one record from the Statistical Population Register, as this latter Register has some duplicates.

The same thing happens, but to a lesser extent, the other way round. Several records from the Population Register of Vitoria-Gasteiz have been linked with the same record in the Statistical Population Register.

In total, there are 231,313 different records from the Population Register of Vitoria-Gasteiz that have been linked with some records in the Statistical Population Register. This means a linkage percentage of 98.99%.

As was indicated in the "Background" section, the linkage procedure had already been carried out with these files by means of non-probabilistic techniques. As the results of this other linkage operation were available, it was decided to compare the results obtained through both methods.

There are 635 records from the Population Register of Vitoria-Gasteiz that have been linked with a different record from the Statistical Population Register in each linkage method. These records were revised manually and the two records with which they got linked were subjected to comparison. In general it was observed that approximately in 4 out of every 5 records, the record linked by means of the probabilistic method is more accurate. And on the other hand, it seems that the deterministic method works better only in 1 out of every 25 records. In the rest of the records, either no criterion was found to determine which of the two was better, or both seemed to have linked with the same individual but with different codes.

On the other hand, there are 535 records that have been linked by means of a probabilistic linkage program that had not been linked by means of the previous procedure , and analogously, there are 1,126 records that were linked by means of the previous procedure that that were not linked during the execution of the programs described in this document.

Some of the observations obtained while revising the latter records are as follows:

- The deterministic method poses problems when linking names of foreign individuals since in some cases various registers with the same name are linked with each other in spite of having different surnames. This happens in general because the second surname of many foreign individuals is not registered and the deterministic method establishes that the comparison between two empty fields is positive and, together with the coincidence of the name, this produces a *match* result in spite of there not being a coincidence in the first surname. This problem is solved with the probabilistic method as this evaluates coincidences or similitude of all the fields on the whole.

- In both methods it has been detected that there are errors when distinguishing brothers, especially in the cases of twins.

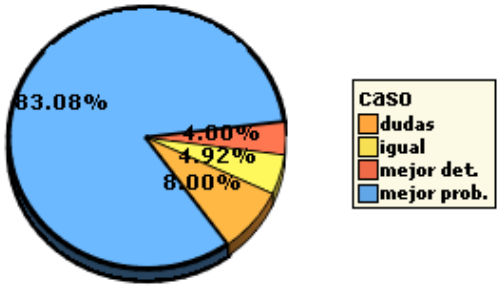
- In some cases the surnames of the individual are the wrong way round. That is to say, what in one file appears as the first surname corresponds with what in another file is the second surname and vice versa. This interchange of surnames does not allow the probabilistic method to identify them as a *match* even though it is one (the deterministic method does identify them correctly).

Bearing in mind this last observation, it was proposed to include code in the linkage program that will deal with this situation. This was done and the results were calculated again. In this second approach, the results were as follows:

Number of linked pairs of records	231,818
Number of non-linked records from the Population Register of Vitoria-Gasteiz	2,185
Number of non-linked records from the Statistical Population Register	2,225,016

As had happened before, there are records from the Population Register of Vitoria-Gasteiz that got linked with more than one record from the Statistical Population Register and therefore produced more than one linked pair. In total, there are 231,485 different linked records from the Population Register of Vitoria-Gasteiz, which means 99.06% of the total. With this modification it was possible to link another 172 records.

After this second execution of the program, we once again searched for the records that had been linked in a different manner depending on the method used. In this occasion, the result was that in 83% of the cases the probabilistic method worked better, in 4% the deterministic method was better, and in 5% both methods worked equally well and there was finally an 8% residue of cases in which it was impossible to reach a decision.



Survey on the Population in Relation to Activity and Statistical Population Register

Another file that was linked with the Statistical Population Register was the Survey on the Population in Relation to Activity. But in this case no attempt was made to localise all the records in the Survey on the Population in Relation to Activity but only those that did not have a Basic Population Unit code.

The Survey on the Population in Relation to Activity is based on a continuous probabilistic sample, that is to say, a panel of housing units that has been continuously renewing itself. The sample of housing units is randomly extracted from the Housing Directory in a stratified manner at the Historical Territory (province) level.

Once the sample of housing units has been obtained, the data from other people living in the selected housing units are associated as from the Statistical Population Register.

For this reason, when extracting the sample of all the individuals of each housing unit, they all possess a Basic Population Unit code.

However, it sometimes happens that when a house is visited for polling purposes, the people who reside in that house are not those who appear in the sample (for example, if those who used to live there have moved).

In this case, the people who currently live in the house are polled. As a consequence of that, these new people will not have the Basic Population Unit code field filled in, and it is going to be these individuals that we are going to try to localise later on in the Statistical Population Register by means of the linkage program.

The file from the Survey on the Population in Relation to Activity had 41,568 different individuals, of which, a little over 10% did not have a Basic Population Unit code. Then, the following files were linked with the following record numbers:

Survey on the Population in Relation to Activity without Basic Population Unit code	4,117
Statistical Population Register	2,456,831

The linkage variables that were used were:

First surname
Second surname
Name
Date of birth
Gender

And as *blocking criteria*, the following were used:

Date of birth
Text code1
Text code2

The results obtained were as follows:

Number of linked pairs of records	3,277
Number of non-linked records from the Survey on the Populataion in Relation to Activity without Basic Population Unit code	855
Number of non-linked records from the Statistical Population Register	2,453,556

There were in total 3,262 different records from the Survey on the Population in Relation to Activity that did not have a Basic Population Unit code and that could be found by means of probabilistic linkage. This means 79.23% of the total of records.

Looking only at the linkage percentage it may seem that the result is not as favourable as that obtained in the exploitation previously described on the Population Register. In this case it is necessary to clarify that it is impossible to find in the Statistical Population

Register some of these individuals interviewed in the Survey on the Population in Relation to Activity, as they may simply not be there. This is the case of individuals that are localised in their housing units and do not appear in the original sample (and therefore lack a Basic Population Unit code) and that correspond to individuals who come from outside the Basque Autonomous Community and who, therefore, are not registered in the Statistical Population Register.

To verify how many of the records from the Survey on the Population in Relation to Activity do not have a Basic Population Unit code and have not been linked because they do not appear in the Statistical Population Register is a very difficult task. What has been possible to check is how many of these records have now been linked because they are individuals who were born after the latest update of the Statistical Population Register. In this case, there are 187 records in the Survey on the Population in Relation to Activity, the dates of birth of which are after the latest update registered in the Statistical Population Register.

Chapter 5

Conclusions

After all the results obtained up to the moment, the general conclusion is that the probabilistic linkage method is considerably useful when linking files that do not have an identifying record code as a common variable.

The main advantage of the probabilistic method with respect to the deterministic method is a greater efficiency in the most difficult cases, as against the simplicity and quicker results of the deterministic method. This implies the elimination of manual treatment in some cases and therefore a reduction of costs.

One requisite of the probabilistic method is a high computational capacity, which is something that has become less and less important nowadays.

EUSTAT is considering not only continuing work with this method for the linkage of individuals but also expanding it so as to consider the linkage of company files.

For company files, it has been observed that the main difference between linking individuals and linking companies is the type of variables that are available in each case. In company records, the *name of the company* variable cannot be treated in the same way as is done when it is the name of an individual as the casuistics are completely different. There is more diversity in company names and besides, the probability of error is higher, since the complexity is also bigger. Another difference is that a variable that appears very often in company files and that has not been treated in the case of individuals, is the address variable. Such a variable may appear recorded in many different ways and its standardisation will mean additional work.

On the other hand, the fact that probabilistic methods have been observed to be considerably useful does not imply that we shall stop using deterministic methods within the Basque Office of Statistics. Since the linkage results depend to a considerable extent on the quality of the files, there will be cases in which it will be better to use other types of methods instead of the probabilistic ones. What is more, it may well be that for some concrete files the ideal option would be a combination of various types of methods.

Currently, EUSTAT is working on the construction of a linkage module that would optimise and facilitate the computational use of all these methods so as to be able to apply the most adequate method in each case.

Chapter 6

Bibliography

[1] FELLEGI, IVAN P. AND SUNTER, ALAN B.

A theory of Record Linkage. Journal of the American Statistical Association, December vol. 64, nº 328, pp. 1183-1210 (1969).

[2] JARO, M.A.

Record Linkage research and the calibration of record linkage algorithms. U.S. Bureau of the Census (1984).

[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. International Journal of Epidemiology (2002).

[5] AYESTARÁN, MARINA AND LEGARRETA, LEIRE.

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