APPLYING METHODS OF RECORD LINKAGE FOR CENSUS VALIDATION IN THE BASQUE STATISTICS OFFICE

Leire Legarreta and Marina Ayestarán

EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADISTICA

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

APPLYING METHODS OF RECORD LINKAGE FOR CENSUS VALIDATION IN THE BASQUE STATISTICS OFFICE

Leire Legarreta and Marina Ayestaran

SUMMARY

The aim of this paper is to describe the experience of the Basque Statistics Office in applying novel methods of record linkage for census validation.

The probabilistic methods of record linkage, pioneered by Fellegi and Sunter, make a feasible and efficient comparison of large databases in a statistically justifiable way, when unique identifiers are not available. Despite their relative complexity, they provide numerous advantages compared to other ad hoc procedures used for this purpose. Initial experiences in Eustat in applying this new methodology and the results thus obtained, with linkage percentages rising for some files from 48% to 93%, clearly showed the need to carry on with this approach.

Based on the theory introduced by Fellegi and Sunter, we have implemented an automatic linkage methodology. One of the improvements we have performed refers to new methods for creating standardised lists for the configurations that variables can adopt. As pairs of strings often show typographical variations, new techniques to compare strings have been developed. These techniques take into account the frequency of the multiple configurations that can be found in the files to be linked, as well as the specific spelling and phonetics of our area. Linkage, integrated as an automatic system, allows the calculation of all the parameters implied in the process from the data files. It uses classic methods to calculate weights, allows the development of each stage of linkage with control over tolerable errors and makes use of different blocking criteria to minimise running time.

These linkage techniques are being used to validate the 2001 Population and Housing Census in the Basque Country using data obtained from two different sources: the Census Validation Survey itself, and the quarterly labour force survey of the Basque Country. Applying our improved linkage techniques to these two different sources allows us to compare two different validation techniques.

Firstly, an introduction to record linkage is made and the objective of the current task is defined. Next, the data and methodology used to carry it out is noted, detailing the most innovative parts of the programmes. Finally, the results that have been obtained from the linkage and validation with the labour force survey and some conclusions that can be drawn are summarised as well as any outstanding tasks.

Keywords: Quality and Quality Assurance in the Use of Administrative Data, Data Processing, Improving Process Quality.

Index

Chapter
1

Introduction

As we move into the 21st century, the quality of information to be worked with is becoming a key issue. It is often necessary to integrate a large amount of different sources, including censuses, surveys or administrative files, to obtain more complete quality information about the same individuals. In this context, record linkage methodology, as pioneered by Fellegi and Sunter, is becoming a vital tool, given that it allows the linkage of large data files in a statistically justifiable way, even when we do not have unique identifiers for the records that we wish to compare.

In the Basque Statistics Office the need was seen to implement computerised record linkage. The lack of unique identifiers in some files, the poor quality of information that had occasionally been gathered and the large number of typographical and phonetic variations on the names and surnames of individuals that had been registered are just some of the problems that the ad hoc procedures developed for this purpose had tried to solve without much previous success. The first trials by Eustat in the application of the new methodology and the results obtained, with linkage percentages for some files that rose from 48% to 93%, made clear the need to carry on with this work.

Chapter
2

Objective

In the light of the results obtained by the first linkage tests, the possibility was presented of applying these techniques to validation work.

At present, Eustat carries out a Census Validation Survey, which sets out to provide quality indicators for the information that has been gathered. On one hand, it measures the degree of coverage of the census operation and on the other, the quality of the responses.

Given the availability of sampling surveys, providing reliable high quality information, the possibility was considered of using some of them as a Validation Survey, with the possible outcome of eliminating this in the future, thus saving time and effort.

To be precise, it was considered that the ideal survey to carry out this work would be our Labour Force Survey or Survey on the Population in Relation to Activity (P.R.A.). This is a Survey made with quarterly periodicity, whose information is reliable, since it is a panel in which each household is asked over six consecutive months, meaning that it contains high quality information. Additionally, it collects exactly the variables that are required to be validated in the Census, such as the socio-demographical variables and the variables related to activity. It also has the advantage that the sample is a fairly wide one.

Chapter
3

Methodology

As previously stated, the underlying theoretical model for the whole linkage project is that introduced by Fellegi and Sunter. There follows a brief comment on this model and some of the most innovative or relevant apects of the project are looked at in detail.

Theoretical model

The probabilistic methods of record linkage set out to compare individuals or records from two different files. The files to be compared are denoted A and B. The objective is to recognise, from among all the pairs of A times B that could be formed, those that refer to the same person, object or entity. Therefore, the objective is to break down the set

$$A \times B = \{(a, b) : a \in A, b \in B\}$$

in the union on the disjoint sets

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

and

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

denoted matches and nonmatches respectively.

Each population unit was registered in a file under certain associated characteristics, such as name, surnames, age or address. This process could introduce errors or imprecisions (coding, transcription or typing errors, typographical and phonetic variations of the area, lost data, etc.) in the records created. As a result of these errors, two members of A and B which do not refer to the same individual could generate identical records and, more frequently, two identical members of A and B could produce different records.

We denote the files generated by A and B LA and LB, and the records corresponding to a and b are denoted $\alpha(a)$ and $\beta(b)$ respectively.

The first step in trying to link records from two files is to compare them. The result of the comparison is a set of codes, coded in statements such as “the name coincides and it is Pedro”, “the name coincides and it is Juan”, “the name does not coincide” or “the name is lost in one of the two files”. Formally, we define the comparison vector as a function vector of the records $\alpha(a)$ and $\beta(b)$ in the following manner: if $a \in A$, $b \in B$ and $\alpha(a) = (\alpha 1, \alpha 2, \ldots, \alpha k)$ and $\beta(b) = (\beta 1, \beta 2, \ldots, \beta k)$ they collect information that refers to the same variables, we generate a comparison vector γ:

$$\gamma = \gamma[\alpha(a), \beta(b)] = \{\gamma^1[\alpha(a), \beta(b)], \ldots, \gamma^k[\alpha(a), \beta(b)]\} = \\ = (\gamma_1, \gamma_2, \ldots, \gamma_k)$$

where each γj with j = 1, 2,…, k represents one of the possible results of the comparison between α j and βj. Each γj will have nj possible associated configurations, which is to say that there will be nj possible results of comparing α j and βj:

α_j^1 : α_j and β_j coincide and are equal to a certain configuration.

α_j^2 : α_j and β_j coincide and are equal to another certain configuration.

⋮

α_j^s : α_j and β_j do not coincide.

⋮

$\alpha_j^{n_k}$: α_k is lost.

The set of all the possible realisations of γ is called comparison space, and is denoted Γ.

In the process of the linkage operation, depending on the value of vector γ(a, b) we decide if (a, b) is a match, that is to say if (a, b)∈M (we will call this decision, known as A1, link) or if it is a nonmatch, (a, b)∈U (we will call this decision, known as A3, nonlink). However, situations could arise in which we are unable to make one of these decisions for specified levels of error, so that we can make a third division, known as A2, which we will call possible link.

A rule of linkage L could be defined as an application of comparison space Γ upon the set of random decision functions D={d(γ)} 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 γ, the linkage rule assigns a status of link or nonlink to each comparison pair.

Let us assume that the comparison vector γ[α(a), β(b)] is a random variable. We can denote the conditional probability that is produced γ given that (a, b)∈M as m(γ), which will be:

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

Similarly, we can denote the conditional probability of γ given that (a, b)∈U for u(γ). Therefore,

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

We must consider the two types of error levels associated with each linkage rule. The first occurs when we compare a nonmatch and we decide that it is a link, the probability of which is:

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

The second type of error occurs when a match is compared and we decide, however, that it is a nonlink, the probability of which is:

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

Fellegi and Sunter defined a linkage rule according to the levels of error μ and λ, which we denote $L(\mu, \lambda, \Gamma)$ optimal, if the relation

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

is maintained for any $L'(\mu, \lambda, \Gamma)$ among all the linkage rules that verify the previous relations.

It can be seen that, according to this definition, an optimal decision rule is that which maximises the probabilities of making positive comparison provisions (which is to say decisions A1 or A3), subject to fixed levels of error. This seems a reasonable decision, given that taking a decision A2 would require expensive operations of manual linkage; on the other hand, if the probability of A2 is not a small one, the linkage process will be of little use.

The authors proposed an optimal linkage rule. For this, they began by defining a unique order in the finite set of all the possible realisations of γ as follows: if any value of γ is such that both $m(\gamma)$ and $u(\gamma)$ are equal to zero, then the probability that it produces this γ is equal to zero and it is therefore not necessary to include it in Γ. Following this, an arbitrary order is assigned to any γ for which $m(\gamma) > 0$, but $u(\gamma) = 0$.

The other γ are ordered so that the corresponding sequence of

$$m(\gamma) / u(\gamma)$$

is monotonely decreasing. When the value of $m(\gamma) / u(\gamma)$ is the same for more than one γ, then the assigned order is arbitrary.

We indicate the ordered set $\{\gamma\}$ by the sub-index i; ($i = 1, 2, \ldots, N\Gamma$) where $N\Gamma$ is the set of points of Γ; and we write $ui = u(\gamma i)$; $mi = m(\gamma i)$.

Let (μ, λ) be an acceptable pair of levels of error in that neither are too big. Let n, n´ be whole so that

$$\mu = \sum_{i=1}^{n} u_i , \qquad \lambda = \sum_{i=n'}^{N_\Gamma} m_i , \qquad 0 < n \leq n' < N_\Gamma,$$

If we define

$$T_\mu = \frac{m(\gamma_n)}{u(\gamma_n)}$$

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

then Fellegi and Sunter demonstrated that the best linkage rule at levels (μ, λ) was given by the form:

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

In many applications, it could be possible to tolerate sufficiently high levels of error so as to eliminate the possibility of action A2. In this case, we would consider n and n´ so that the average set of γ in (1) was empty. In other words, each pair (a, b) is located either in M or in U. In fact, this is the decision that we adopted in our linkage programme, thereby establishing a single limit $T\mu = T\lambda$.

Some simplification assumptions

This theoretical model that we have implemented presents numerous advantages compared to the ad-hoc procedure used for this purpose until now. However, in trying to implement the theory, we come across a large amount of practical problems.

Some have been tackled previously by various authors and various solutions have been presented. For example, the one that refers to the elevated number of values that the vector γ can adopt, so that estimation of the corresponding probabilities of $m(\gamma)$ and $u(\gamma)$ could become totally impracticable. To reduce the large amount of calculations and storage requirements, the following simplification assumptions are made concerning the distribution of γ. We assume that the components of the vector are mutually statistically independent with respect to each one of the conditional distributions. So we can state:

$$m(\gamma) = m_1(\gamma^1) \,.\, m_2(\gamma^2). \ldots . m_k(\gamma^k) \qquad (24)$$

$$u(\gamma) = u_1(\gamma^1) \,.\, u_2(\gamma^2). \ldots . u_k(\gamma^k) \qquad (25)$$

Furthermore, it is clear that any monotonely increasing function of $m(\gamma) / u(\gamma)$ could work just as well as a statistical test for the purpose of our linkage rule. It is especially worthwhile to use the logarithm of this weight and define

$$\omega^k(\gamma^k) = log\ m(\gamma^k) - log\ u(\gamma^k). \qquad (26)$$

We can then write

$$\omega(\gamma) = \omega^1 + \omega^2 + \ldots + \omega^k \qquad (27)$$

and use $\omega(\gamma)$ as a statistical test, accepting that if $u(\gamma) = 0$ or $m(\gamma) = 0$ then $\omega(\gamma) = +\infty$ (or $\omega(\gamma) = -\infty$) in the sense that $\omega(\gamma)$ is bigger (or smaller) than any given finite number.

Supposing that γk can take nk different configurations, $\gamma_1^k, \gamma_2^k, \ldots, \gamma_{n_k}^k$. We define

$$\omega_j^k = \log m(\gamma_j^k) - \log u(\gamma_j^k). \qquad (28)$$

A priori, the total number of configurations (which is to say, the number of points of $\gamma \in \Gamma$) is obviously n1.n2….nk. However, after the simplification assumption that we have introduced and also due to the additive quality of the weights defined in components, it is sufficient to determine n1 + n2 + … + nk weights. We can always determine weights associated with any γ using this additivity and thus significantly reduce the cardinal number of Γ.

Calculation of weights

Another problem that has been tackled on several occasions is that referring to how to calculate weights m(γ) and u(γ). A great number of methods have been put forward for one of the major problems presented by record linkage. In our case, we have chosen a weight calculation method based on the frequency that each of the configurations of the registered data appears in the files to be linked, which was also proposed by Fellegi and Sunter. This method allows us to obtain all the weight information for each one of the different configurations, using only the data contained in the files that are to be linked. It also presents other advantages, such as its relative simplicity. The idea is basically that an agreement on the name variable, for example, will produce a positive weight and that the more unusual the name, the greater the weight; a disagreement on the name will produce a negative weight which will diminish according to the coding errors that the file has. This result seems intuitively attractive.

Creation of standardised lists

Some of the problems raised by linkage have already been resolved or tackled. However, there are other aspects which give rise to problems and to which we must offer a possible solution. One the most conflictive arises from the need to posses standardised lists with all configurations error-free for each one of the variables that are part of the process.

Record linkage generally means that the files have been compiled at different times or by different organisations, using different guidelines to register and abbreviate the data. Even when they are compiled by the same organisation, different individuals may answer to differing information needs and therefore may use different formats for the name and address. Furthermore, the same individual may reply in differing ways at different times, omitting, for example, an initial of a name in one case, but including in another, or abbreviating part of an address at different times.

As well as this problem, transcription or typing errors may occur. Then there is the existence in our case of two official languages, which could mean that in many cases the same individual registers their name in two different languages in different files.

All these situations make the creation of a standardised list of names and addresses a difficult task for a computerised linkage programme and the comparison of pairs of files complicated, unless the information is standardised in a much more consistent manner.

In the case of variables such as sex, for example, it is obvious that there are two possible configurations, and that any value that differs from these will not be valid. However, this is not so simple for variables such as first name, surnames or street names.

Certain typographical or phonetic variations are easily corrected, eliminating H´s, unifying V´s and B´s or presenting all the character strings in upper case. However, there are other types of problems that are complex and occasionally insurmountable.

The problem of transcription errors has been tackled with the creation of string comparators such as the Jaro one, which set out to identify transpositions, suppressions or deletions of characters, giving a measure of difference between the strings. These comparators take into account the number characters shared or otherwise by two strings, the number of transpositions and their length. However, it is not feasible to compare a large quantity of configurations of a variable using only this type of technique.

Every day, new names and surnames are added to our files. The set of names used to register newly-born children gets ever wider and the number of foreign individuals of numerous nationalities that have become part of our data bases multiply the number of phonetic and spelling peculiarities very different to our own. It is obvious that it would be extremely expensive to maintain up-to-date look-up tables for all the variables. Additionally, it is evident that it is not possible to compile lists with all the possible combinations of errors that could be made when registering a string.

To solve these problems as much as possible, we have developed a method of configuration of standardised lists, making set modifications on the configurations as they appear in the original files, with the aim of matching those configurations that correspond in the original itself.

The first step is to make a simple standardisation process in the variables, consisting in actions such as eliminating the H´s, identifying V´s y B´s, putting all the configurations in upper case or eliminating accents and punctuation marks. Next we relate those configurations that are identical, except characters with similar sounds, such as ADAM and ADAN or ENEKOITZ and ENEKOIZ. We also make transformations in configurations in order to correct problems with diminutives or apellatives, changing FDEZ for FERNANDEZ or relating CONTXI with CONCEPCIÓN.

Following all this and as a final and most complicated step, we have developed an algorithm that relates similar configurations, except differences that may be due to transpositions, deletions or insertions of characters. For these related pairs, an estimation is made of the probability that one string really comes from the other, bearing in mind the following questions: the number of errors that would have been made under our hypothesis, estimations concerning the probability of making transcription errors that we are prepared to admit, or the length of the strings. Depending on these estimated probabilities we decide if both intend to record the same characteristic or not.

Once all this process has been carried out, we assign a numerical code to each string. Those that, in our opinion, intend to record the same information are given the same code.

This method of compiling standardised lists has the advantage that it is carried out in each linkage project, so that no auxiliary information is necessary and the lists are created only from the information collected in the files. This way, the incorporation of new configurations not previously taken into account does not present a problem. Given that the frequency with which each one of the configurations appears is taken into account, it seems more logical that more errors are made when registering a very frequently occurring string and in the same way, limits the quantity of related configurations to a reasonable number. The main disadvantage would appear to be the running time.

Establishing the limit

Another problem posed by taking on the task of linkage is deciding where to fix the limits $T\mu$ y $T\lambda$ which determine the boundary between links and possible links, and between these and nonlinks. In the case of our project, we have considered $T\mu = T\lambda$ with the aim of always making a decision about each pair, eliminating posterior manual checking which is usually the case with doubtful pairs.

We can make estimations of the errors that we accept according to the limit that we decide to consider. However, a simple method to choose an adequate limit is based on the observation of frequency histograms with the weights resulting from all the comparisons. The histogram supplies us with valuable information to assist us in making a decision, based on the form that this takes for the different values of the weights.

Blocking criteria

With two populations, comparing all the possible pairs of records that could constitute a match would have too great a computational cost. Instead of considering all the pairs *A x B*, a common method in the problem of linkage consists of considering only those pairs that coincide in certain identifiers or *blocking* criteria.

Blocking criteria lessen the number of pairs that we have to take into account, as well as increasing the number of false nonmatches, since they eliminate the true matches that, due to error, do not coincide in the blocking criteria.

The ideal blocking variables are those for which we are sure do not contain any unusually large category and about which we are fairly certain as to the accuracy with which the data has been registered.

Occasionally, it is recommendable to combine different criteria. It is possible to use one blocking criteria to extract the greatest possible number of matches, and then to use others in order to capture those pairs that could not be previously identified, due to errors in the blocking variables. On the other hand, in our linkage programme, we have chosen all the criteria one by one, storing the resulting pairs with the corresponding weights. Of these, we will eventually select the pairs that make the final weight as great as possible. In this way, we guarantee that the final assignment between pairs of records is the optimal one, despite the handicap of the somewhat more extended running time.

SAS programming

With all these and a few other modifications, we have gone on to implement a series of macros, all programmed in SAS, which automatically allow the development of record linkage of two data files after the user has introduced only initial data. This initial data refers to the localisation and name of the files to be linked, the variables that we wish to use as linkage variables, what type each one of the variables is and the blocking criteria we wish to use.

Using this initial information, the programme gets underway, in the following manner: firstly, the information supplied by the user is analysed, the linkage variables are identified, renamed or modified correctly and a check is carried out to see whether all the necessary variables have been introduced, etc. If an absence of certain information is detected, the running of the programme is halted and the user is requested to introduce what is missing.

Once all the initial information has been gathered and analysed, the standardisation process begins upon each one of the variables, according to the type that each one belongs to, as explained previously. The corresponding standardised configurations are assigned, the linkage weights are calculated and the various blockings are executed. Using the weights that have been obtained, the programme creates a frequency histogram, which is automatically displayed to the user. The information contained in the histogram should therefore allow the user to take the correct decision on which limit to consider. Then, the programme requests the user to introduce the value that they wish to use to assign each comparison pair the status of link or nonlink. Once the limit is introduced, the programme continues to run.

On comparing some pairs with others, it is possible that more than one record from a file is related with a record from another. As we have mentioned, the programme has been developed in order to execute all the *blocking* criteria one by one and store the results. Given that the objective of linkage is to make the relation between the records one-to-one, we apply a lineal assignation algorithm, which selects those combinations of pairs or records so that the final weight is as great as possible, in accordance with the previous restriction.

In the final stage of the programme, we analyse those pairs whose weight, while not being above the established limit, is high enough to be taken into consideration. Checks are carried out upon these pairs of records in order to take a final decision as to their status. For example, an algorithm will help us to detect if the names or surnames from one of two related records are diminutives of the other, or if one is part of the other. We can also check if an error regarding sex could have been made, given that the name coincides, or if a transcription error has taken place that has not been correctly detected by our identification algorithm, for which we use the Jaro string comparator. Once these and other simple checks have been carried out, the weights for the comparison pairs are recalculated and to this result a lineal assignation algorithm is applied once more, thus obtaining the final result.

Following this process, the programme generates three data files in the localisation requested by the user. One contains the records that have been linked from both files, whereas the other two store those records that it has not been possible to link. Additionally, the programme gives the linkage percentage achieved and the running time.

Chapter
4

Results and conclusions

- Once the whole process had been carried out, the linkage results obtained were 97%, which is to say that 97% of the records that had been registered in the selected Labour Force Survey were also present in the Population Census. Initially, the results appear to be fairly positive.

- The analysis of the coverage has been carried out based on the households to which the linked individuals belong in both sources, the Census and the Survey on the Population in Relation to Activity.

- There are 10,414 individuals in the Survey on the Population in Relation to Activity for those households, while we have managed to link 10,227 in the Census. Therefore, the coverage is 98%.

- On the other hand, all the individuals of the Census in the matched households have been linked with individuals of this Survey, so in this sense the coverage is 100%.

- Once the individual linkage of the Survey on the Population in Relation to Activity with the Census has been made, the comparison of the common characteristics has been carried out from the two sources.

- The file with the common individuals in both sources (as a result of the linkage) has been generated and the answer consistency has been analyzed.

- The overall consistency index, which measures the proportion of records equally classified in all the categories of one variable has been calculated. Therefore, it shows the total stability of the variable. Over 85% it is considered an acceptable value.

- The percentage of identically classified has also obtained for each variable value. Consequently, it measures the answer stability in each category. An acceptable value is the one which is over 80%.

- The characteristics, for which this index has been generated, are demographic variables such as year of birth (every five years), sex, marital status, head of household and level of instruction; variables relationated to the activity like the relation to the activity itself and characteristics of the working population such as job, activity sector and professional situation.

- The validation results are good in general for demographic variables. The overall consistency index is about 95%, except the level of instruction, which is only 61%, but with a great variety by categories.

- The relation to the activity variable has an overall consistency index of 86%. The most consistent categories are the working population category and the unemployed population category. The categories relating to unemployed persons show a low consistency.

- As regards the characteristic variables of unemployed population, the activity sector shows the highest consistency, 83%, being the professional activity only 55%.

- These results are similar to the obtained results from the 1996 Validation Survey which was used to validate the results from the 1996 Population and Housing Census. Therefore, it seems feasible that in future, these techniques could be employed as validation tools.

- Furthermore, in time, record linkage could tackle a great number of existing applications, such as the application in companies' records which have different problems to the person's records due to the variable types that have to be taken into account.

Chapter

5

References

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

A theory or 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 (1989)