

The Control of Non-Sampling Errors on Linked Data: an Application on Population Census

*Il controllo dell'errore non campionario su dati abbinati: un'applicazione
al Censimento della popolazione*

Giorgio Alleva¹, Marco Fortini², Andrea Tancredi¹

¹Dipartimento di studi geoeconomici, linguistici, statistici e storici per l'analisi regionale, Sapienza Università di Roma, Via del Castro Laurenziano, 9 - 00161 Rome

²Istituto Nazionale di Statistica, Via Balbo 16, - 00184 Rome
giorgio.alleva@uniroma1.it, andrea.tancredi@uniroma1.it, fortini@istat.it

Riassunto. Una qualsiasi indagine statistica è soggetta a mancate risposte, errori di copertura e di misurazione. Nel caso di indagini di elevato interesse pubblico viene spesso condotta una seconda rilevazione con il duplice obiettivo di valutare l'entità dell'informazione mancante e la precisione delle misurazioni. In questi casi, per la validazione dell'informazione statistica complessiva, tutte le fasi del processo di *linkage* necessitano di un'accurata analisi metodologica. A partire dal censimento della popolazione residente in Italia e dalla successiva indagine di copertura, nel lavoro è valutata la discrepanza che intercorre nell'informazione statistica rilasciata dalle due indagini, e discusso l'eventuale miglioramento che si ottiene integrando i dati ai fini della stima delle caratteristiche della popolazione residente straniera.

Keywords: post enumeration survey, record linkage, response errors and under coverage errors.

1. Data quality management: some guidelines

Error management represents one of the main aspects of quality control in the statistical data production process. For more than ten years quality improvement has been a strategic objective for the Italian Statistical System (SISTAN) as can be seen from planning documents and the investments made in the improvement of production processes, but above all from the results that have so far been achieved and consolidated. In particular, the relevant innovations made in the different phases of the data production process, in new surveys and in the revision of traditional surveys, have taken place by way of a complex optimization process in response to the growth in demand for statistical information within a framework of increasing bonds fixed by national and European regulations and by a progressive reduction of financial resources (Alleva, 2004).

During the process of quality improvement, and considering the experiences matured by the National Statistical Offices, it is now possible to identify elements that represent new paradigms for all those involved in improving the productive system of official statistics.

This paper, though the result of a close collaboration, was specifically elaborated as follows: sections 1 and 4 by G. Alleva, sections 2 and 3.1 by M. Fortini, section 3.2 by A. Tancredi

First of all, evaluation and quality control still cannot be considered the final elements of the production process but strictly belong to the survey design and hence the planning phase. Data quality still cannot be conceived as a derived element of processes or products, but instead as one of the main factors that operates in the decisional procedures and contributes to defining the present statistical outline. (Buratta and Zannella, 2004). Data quality should be planned before it is measured, and the quality evaluation has to be organized according to predefined and diffused standards. Subsequent to this, and in coherence with the data quality definition, there is greater awareness the quality is planned according to user requirements.

On the other hand, this trend of attention given to the users presents different implications, and each of these should promote specific reflections on SISTAN's organization and resource attribution. It is strategic to plan and control statistical quality not only by considering the single production process but also the entire system and the whole set of processes and producers involved in the building up of statistical information on a specific subject (environmental and sustainable development represents one of the most emblematic cases, with greater scope for improvement if the production and diffusion of SISTAN is planned according to the quality standards established directly by the users). Informative systems and integrated modalities of data diffusion in general, also belonging to different organs of SISTAN, represent a theme worth investing in with greater conviction, as an effective way of responding to user demand. It is important to pay greater attention to the diffusion of documentation on data collection and elaboration, the unique possible answer to users who want to evaluate whether particular statistical data can be considered useful for their own specific purposes; micro data diffusion, compatible with the statistical disclosure limitation, should be made more active and promoted in order to reduce the gap with other European countries in this mode of diffusion. Quality evaluation should not end with data release, and so should not neglect to take into account a customer satisfaction approach (the results of these analyses should play a certain role in quality planning). Finally, it is necessary to ask ourselves if producers, inside statistical quality evaluation, should also evaluate and measure the real use of statistical information by the users to whom the information is directed, moving from an evaluation of products (outputs) to an evaluation of the global results that their action produced for society (outcome) (Giovannini, 2007).

The introduction of new technologies for data collection and management, as well the experimentation and application of methodologies for the improvement of singular aspects of data quality, represent the implemented actions taken in the last few years, and those that are expected in the future. In this framework, besides early estimates and small area estimation methods, a third element of frontier is represented by the development and application of record linkage techniques.

Indeed, record-linkage represents one of the elements that could best assure advantages from the informative patrimony accumulated by data producers, both in the way of a greater supply of joint statistical information on the same population, in the building of statistical informative systems or integrated (or federated) archives, and in the way of the data quality improvement of a survey through information obtained from other surveys or administrative archives. However, counter to the potential advantages offered, it is necessary to take into account the complexity and the further uncertainty that record linkage can also introduce.

With this aim in mind, the present paper will attempt to verify how it is possible to discern between response errors and under-coverage errors in the Italian Population Census of 2001, through a post enumeration survey and in so doing, apply a record linkage procedure that could reduce these errors but give rise to other sources of error. These different sources of error are defined in paragraph 2, as well as their interaction with linkage errors. Paragraph 3 will present an application regarding the estimation of some characteristics of the foreign resident population. Some concluding remarks are drawn in paragraph 4.

2. Sources of survey errors and their relationship with linkage errors

A statistical survey involves different phases and procedures and is prone to various sources of errors. A typical classification divides the survey errors into sampling and non-sampling ones, arranging the non-sampling errors into further classes that identify coverage, non response, measurement and processing errors.

To control the errors arising during the various phases of a survey and their complex interactions, a control system is usually planned. The most important goal of a survey control system is to estimate the mean square error (MSE) regarding the statistics produced by the survey and divide it into the error components that are kept under control. Recently, however, instead of producing a direct evaluation of the accuracy estimates, the survey quality researchers are moving towards the production of quality indicators that serve only as proxy variables for the MSE but are easier and cheaper to calculate and moreover, give direct insights on the process phases eventually going out of control (Biemer and Lyberg, 2003). As an example of mixed strategy “Survey on Structure and Production of Agricultural Firms” can be mentioned, where a direct estimate of the measurement errors is produced together with many quality indicators as timelines of the field operations, data capture error rates and item non response rates, etc..

In fact, the census usually provides the opportunity to implement a more sophisticated survey control system, thanks to the higher financial resources available and to the relevance of the information to produce. In this context a post enumeration survey (PES) is often carried out to evaluate the census population totals for the under-coverage errors. Several decisions can be taken on the basis of the PES results, from the simple evaluation of the information produced, as done in Italy, to a limited or even complete adjustment of the census count, as done in the USA (Brown et al., 1999) and the UK (Brown et al., 1999b) respectively. Based on a capture-recapture method, the basic counts needed to estimate the coverage errors are obtained by identifying the individuals counted in both surveys through a record linkage between the records belonging to the census data and those from PES.

Unfortunately, the same kinds of errors afflicting the master survey (Census) will impinge also on the control survey, and the errors, all together, will interact with the record-linkage procedures. For this reason, an accurate analysis of the difficulties arising from these errors has to be conducted in order to evaluate the data correctly.

In the present context we will refer to the census counts, focusing on the relationship between the linkage procedures and two of the non-sampling error types that affect both the census and the PES estimates, namely the under-coverage and the response errors. The existing literature considers the effect of linked data on regression estimates (Lahiri and Larsen, 2005), measurement error estimation (Neter et al., 1965; Brancato et al.,

2004), and under-coverage estimation (Biemer, 1988; Ding and Fienberg, 1994). In the present paper we consider how frequency distributions can be modified by the combined effect of under-coverage, response and linkage errors and how the effect of these factors can be evaluated when dealing with a control survey. We considered a probabilistic record-linkage approach in opposition to a deterministic one, since the explicit definition of a probabilistic model of the former allows a more natural estimation of the linkage errors than the latter. In the following section we will consider the errors already introduced in more detail and the effect of their associations on the estimates.

Coverage errors

These regard the omission, duplication or wrongful inclusion of the units in the population or sample. This is called under-coverage in the case of omissions, whereas it is called over-coverage when duplication or wrongful inclusion occurs. In a sample survey the coverage errors in the sampling frame cause both the inclusion of out-of-scope units (over-coverage) and the impossibility of selecting some in-scope units (under-coverage).

An estimate of over-coverage can be obtained by doing an eligibility check on the surveyed units, whereas under-coverage can be studied if an independent list of units is available through control surveys, administrative data or other sources.

Under-coverage is generally more difficult to investigate than over-coverage, since statistical models and record-linkage procedures need to be applied, resulting in hypothesis constraints and the occurrence of processing errors. A classical approach to this problem is based on the capture-recapture method, so called since it was formerly used for animal abundance estimation (Seber, 1982) and only afterward applied to human populations (Sekar and Deming, 1949). Under this method a simple estimate of the unknown population total $\hat{N}$ is obtained as

$$\hat{N} = \frac{N_{CEN} N_{PES}}{N_{11}} \quad (1)$$

by means of two independent lists of units. The frequencies x_{1+} , x_{+1} and x_{11} come from the observed table

Figure 1. *Contingency table for a capture recapture experiment*

	Second list		
First list	<i>Present</i>	<i>Absent</i>	<i>Total</i>
<i>Present</i>	N_{11}	N_{12}	N_{CEN}
<i>Absent</i>	N_{21}		
<i>Total</i>	N_{PES}		$\hat{N}$

which represents the cross-tabulation of the two lists and corresponds to the units captured by the first (x_{1+}), by the second (x_{+1}) and by both (x_{11}) the lists. From the table above, we can see that the fourth cell is of course unknown, as it refers to the total number of units captured by neither lists. The estimate of $\hat{N}$ follows from the multinomial distribution on the observed frequencies and it can be seen as a particular case of a multinomial model with structural zeros (Bishop, Fienberg and Holland, 1975). Fundamental hypotheses for the correct estimation of $\hat{N}$ under the model

sketched above are the statistical independence between the two captures, the homogeneity of the capture probabilities for each single list and the ability to recognise without error the units captured by both lists. A review of the possible statistical models based on the capture-recapture method when applied to human populations, either for complete or area sampling of the second capture, is given in Wolter (1986). It is worth noting that, whereas the effect of coverage errors regards mainly the population counts, if the coverage rates differ through the categories of one or more classification variables, their joint and marginal frequency distributions will also be affected.

Measurement errors

These consist in the difference between the observed and the true value with respect to each surveyed statistical unit. These errors are called response errors when resulting from the interview process, i.e. problems affecting the questionnaire such as: the questions posed by the interviewers, the answers given by the respondents or even the interaction between respondents and interviewers. When they depend instead on survey process failures they are called processing errors. The simplest model for measurement error is given by

$$y_i = m_i + b_i + e_i \quad (2)$$

where: y_i is the observed value for the i -th unit, m_i is the true value for the i -th unit, b_i is a systematic error component for the i -th unit and e_i is a random error component for the i -th unit so that $E(e_i)=0$ and $V(e_i)=\sigma_e^2$

Following this model, the classical theory of non-sampling errors provides an estimate of σ_e^2 by means of a second measure carried out without reconciliation, with the requirement that both measures take the same error distribution. When a re-interview with reconciliation is available, the bias component $B=\sum_{(i)} b_i$ can also be estimated. Extensions of the simple model can account for other sources of measurement errors such as/like the interviewer effect, the mode of data collection and some field conditions (Biemer and Stokes, 1991).

Possible sources for the second measure are given by re-interview surveys or administrative data collections. It is worth mentioning that when the second measure is collected through administrative data or by means of an area sample (as done for the PES), a record-linkage is needed in order to match the first and the second measurement for each of the in-scope units. An illustration is given by the European Statistics on Incomes and Living Conditions (EU-SILC), which is carried out for Italy by Istat, where tax files are linked to the survey records to validate the answers given by the survey respondents on the subject of household incomes. Measurement error results in the inaccuracy of the estimates, both for the mean and for the distribution of the investigated characters.

Probabilistic linkage and linkage errors

Record-linkage consists in matching the records belonging to different data sets if they correspond to the same statistical unit (Belin and Rubin, 1995).

Let us consider two partially overlapping files **A** and **B** consisting in the same type of units (e.g. individuals) and let Ω be the set of all possible pairs of records from **A** and **B**, i.e. $\Omega=\{(a,b): a \in A, b \in B\}$. Suppose also that the two files consist of vectors of variables $(\mathbf{X}_A, \mathbf{Z}_A)$ and $(\mathbf{X}_B, \mathbf{U}_B)$, either quantitative or qualitative, and that $\mathbf{X}_A$ and $\mathbf{X}_B$ are sub-vectors of common identifiers, called key variables in what follows, so that any single

unit is univocally identified by an observation $\mathbf{x}$. Given the definition above we can formally define record-linkage as the problem of assigning each pair $(a,b) \in \Omega$ to either one of the two subsets M or U , and this identifies the matches and the unmatches respectively. Fellegi and Sunter (1969), defined the likelihood ratio

$$w_{(a,b)} = \frac{\Pr(\gamma^{ab} | M)}{\Pr(\gamma^{ab} | U)} = \frac{m_{(a,b)}}{u_{(a,b)}} \quad (3)$$

where γ^{ab} represents the vector of indicator variables regarding the pair (a,b) that assumes 1 in the j -th position if $x_{a,j}^A = x_{b,j}^B$ and 0 otherwise. By means of the ratio $w_{(a,b)}$ the following test can be conducted:

$$\begin{cases} \log(w_{(a,b)}) < K_1, & (a,b) \in U \\ K_1 \leq \log(w_{(a,b)}) \leq K_2, & (a,b) \in Z \\ \log(w_{(a,b)}) > K_2, & (a,b) \in M \end{cases} \quad (4)$$

where K_1 and K_2 are two thresholds chosen so as to assign the pair (a,b) to either one of the sets U , M or Z , where the pairs falling in set Z are held out for clerical review in order to be classified either as matches or nonmatches.

The probabilities $\mathbf{m}$, $\mathbf{u}$ and p , where p is the probability to randomly sample a match from the whole set of the pairs Ω , can be estimated through the EM procedure for a mixture model, as shown in Jaro (1989), provided that the comparisons are conditionally independent given the belonging of the pairs to M or U . Conditional independence can also be partially released in the framework of a latent class estimation model provided that model parameters remain identified (Armstrong and Mayda, 1993; Hagenaars J.A. 1993).

Often, further constraints require that each unit of file $\mathbf{A}$ is linked to at the most one unit of file $\mathbf{B}$ and vice versa; in this case the linkage is referred to as a 1-1 problem and can be efficiently solved by means of a linear sum assignment procedure (Jaro, 1989).

Throughout the linkage operation, two types of errors can be made, consisting in the matching records that fail to be linked and in the nonmatching records which are incorrectly linked. As a consequence, it is important to carefully control the number of false match and false non-match errors that arise from the linkage procedures in order to avoid biases and inefficiencies in the analysis. In our case, the linkage errors can affect the capture-recapture estimation since they introduce a bias into the count of the units enumerated by the two surveys. In a similar way the linkage errors induce measurement errors since wrong values are assigned when pairs are erroneously matched.

3. An application to the 2001 Census Post Enumeration Survey

3.1 Data sources and experimental context

The 2001 Census Post Enumeration Survey (PES) was planned in order to estimate the rate of coverage errors in the 2001 Italian Census with respect to people and families which are resident in Italy. In this illustration we describe only the main features of PES which were needed for the development of our data analysis. Details on the 2001 PES can be found in Istat (2007)

Since the inclusion of double enumerations was prevented to a large extent by the accurate checks implemented during the field operations carried out by the municipalities, the PES focused mainly on the estimation of under-coverage errors. For this purpose, a sample of 98 municipalities (out of 8,103) and 1,098 census enumeration areas (out of the 352,219 enumeration areas including at least one person) was selected, so as to represent the five broad Italian geographical areas and a population classified according to four categories of municipalities. The selected sample of enumeration areas was completely re-enumerated immediately after the end of the census field operations finding 173,109 people in-scope belonging to 66,384 families.

The 182,519 Census records resulting in the enumeration areas sampled for the PES were then matched to the PES data findings of 169,980 people counted in both surveys. To achieve these results a complex record-linkage strategy was implemented including both deterministic and probabilistic record-linkage steps. The results of the linkage were used as an input to apply a capture-recapture model to the data and estimate the rate of census under-coverage.

In the present context, the 4,375 foreigners included in the sample by either the PES, the Census or both the surveys, were selected for our analyses, due to their peculiar status as an elusive population and also because of its interest as an emerging social issue. The gender and the occupational status of the foreigners were taken into account to identify the degree of difference between demographic and socio-economic variables with respect to errors effect, the former being considered easier to measure than the latter. In so doing, it is possible to show how survey results are affected by the many operational choices made during the data production process. Particular consideration will be given to the effect that record-linkage procedures can have on the accuracy of results. The sampling weights were not considered in our application in order to keep it manageable and more focused on the non-sampling errors. Such a decision, however, prevents the following results from being directly extended to the whole Italian population.

3.2 *The effect of error sources on the foreign population*

In this section we describe the results obtained by linking a subset of the records from the census population survey (CEN, in the following) with a subset of the records from the PES sample. The file extracted from the census population survey comprises 3,675 records with 18 variables some of which are: standardized name, standardized surname, sex, day of birth, month of birth, year of birth, place of birth. The PES file has 3,437 records and the same variables as the CEN file.

Table 1: *Gender distributions for the foreigners obtained from the census survey records (columns 2-3), from the post enumeration survey records (columns 4-5), from the census survey linked records (columns 6-7) and from the post enumeration survey linked records (columns 8-9).*

Gender	Census Survey		Post Enum. Survey		Linked records			
					Census Survey		Post Enum. Survey	
	N_i	$f_i \%$	N_i	$f_i \%$	N_i	$f_i \%$	N_i	$f_i \%$
Male	1,815	51.45	1,685	49.96	1,388	50.79	1,381	50.92
Female	1,779	48.65	1,688	50.04	1,345	49.21	1,331	49.08
Missing value	11		64		4		25	
Total	3,675	100.00	3,437	100.00	2,737	100.00	2,737	100.00

The main differences between the two files should be ascribed principally to coverage problems. In fact, some of the resident foreigners, selected also for the post-enumeration survey, were not reached in the first survey but listed in the second; vice-versa some resident foreigners were reached in the first but not in the second survey. Measurement errors may also affect the results for some variables and thus lead to differences between the observed data distributions in the two occurrences. The first columns of Table 1 show the absolute and relative frequencies for the gender distributions in the two files.

Notice that the male frequency for the foreigners is 51.45 percent in the CEN file and 49.46 percent in the PES file. In order to measure the distance between the two gender distributions we calculated the dissimilarity index

$$z = \frac{1}{2} \sum_{i=1}^k |f_{A,i} - f_{B,i}| \quad (5)$$

where $f_{A,i}$ and $f_{B,i}$ are the relative frequencies for the i th modality in the two distributions, obtaining a value for z equal to 0.015.

In order to eliminate the effect of the different coverage in the two distributions, we performed a standard record-linkage analysis on the two files considering the enumeration area as block variables. Standardized name, standardized surname, day of birth, month of birth, year of birth, level of education and occupational status have been used as key variables. For each block the parameters $\mathbf{m}$, $\mathbf{u}$ and $\mathbf{p}$ have been estimated by the E-M algorithm. Finally, the common records between the same block of the two files are estimated by solving the constrained maximization of the function

$$\sum_{ab} X_{ab} w_{ab} \quad (6)$$

with respect to the Boolean variables X_{ab} , where w_{ab} are the logarithm of Fellegi-Sunter estimated weights. With this procedure we estimated 2,737 couples of common records. 97 percent of these couples have more than 4 key variables equal in both files. The last 4 columns of Table 1 show the gender distribution for the 2,737 estimated common units when this variable is observed in both the census survey records and the post enumeration survey records. The male frequency in the subset of linked records for the CEN file is 50.79 percent, whereas it is 50.92 for the PES file. By focusing on the estimated common records we have strongly reduced the distance between the two distributions. In fact, the dissimilarity index between the two gender distributions is much smaller than before and equal to $z=0.0013$. However, there is still a slight difference due to both the measurement error (which, however, should be small for this variable) and the introduction of the linkage error (a false match could be effectively a male in one file and a female in the other).

One of the goals of the record-linkage application is to produce a new enlarged data set with all the records from one survey and the records of the other survey which have not been linked. In Table 2 we report the gender distributions obtained by adding the 700 records of the PES file to the CEN file which have not been linked (file DB1) and by adding the 938 records of the CEN file to the PES file also not linked (file DB2). The greater sample size has reduced the coverage error, and now the differences between the

gender distributions in the two enlarged data set are smaller than before. In fact, the male frequencies are 50.61 percent and 50.70 per cent, respectively and the dissimilarity index z is equal to 0.00085

Table 2: Gender distributions for the foreigners in different enlarged databases. (Database DB1 is obtained by enlarging the census file with the records of the post enumeration survey which have not been linked for the analysis of the gender distribution. Database DB2 is obtained by enlarging the post enumeration survey file with the census records which have not been linked)

Gender	DB1		DB2	
	N_i	$f_i \%$	N_i	$f_i \%$
Male	2,189	50.61	2,182	50.70
Female	2,136	49.39	2,122	49.31
Missing value	50		71	
Total	4,375	100.00	4,375	100.00

In order to completely eliminate the coverage error we estimate the total number of males by capture-recapture methods. We can obtain two estimates, depending on the source considered in drawing the gender of the linked records. Let N_M be the unknown number of male foreigners, $n_{M,CEN}$, and $n_{M,PES}$ be the total number of males in the CEN and PES files. Moreover let $t_{M,CEN}$ and $t_{M,PES}$ be the number of males in the linked records when drawing the information on the gender from the CEN or PES files respectively. The total number of male foreigners can be estimated by

$$\hat{N}_{M,CEN} = \frac{n_{M,CEN} \times (n_{M,PES} - t_{M,PES} + t_{M,CEN})}{t_{M,CEN}} \quad (7)$$

or by

$$\hat{N}_{M,PES} = \frac{(n_{M,CEN} - t_{M,CEN} + t_{M,PES}) \times n_{M,PES}}{t_{M,PES}} \quad (8)$$

and analogous estimators can be found for the unknown number of females.

Table 3 shows the capture-recapture estimates for the absolute and relative frequencies of the gender distribution. Note that the two estimators give similar results with a slight greater frequency of females. The dissimilarity index in this case is equal to 0.0015

Table 3: Gender distribution for the foreigners obtained by capture-recapture methods drawing the gender of the linked records from the census survey (columns 2-3) and from post enumeration survey (columns 4-5).

Gender	Census survey based		Post enumeration survey based	
	N_i	$f_i \%$	N_i	$f_i \%$
Male	2,212	49.78	2,206	49.63
Female	2,232	50.22	2,238	50.37
Total	4,444	100.00	4,386	100.00

We now consider the occupational status. Table 4 shows the distribution for this variable both for the CEN and PES file. The two distributions show quite different

patterns. For example, the student frequency is 8.48 percent in the CEN file and 4.87 percent in the PES file. The dissimilarity index z is equal to 0.056. Such a difference between the two distributions is caused by the strong measurement error which affect the occupational status.

In performing a record-linkage analysis, the occupational status was taken away from the key variables and the sex variable was introduced. Table 4 shows the occupational status distribution with the linked records for both surveys. The difference between the two distributions seems wider than before. For example the employed frequency is 57.22 percent for the linked records of the CEN file, while it is 61.24 percent for the linked records of the PES file. The dissimilarity index is equal to 0.068.

Table 4: Occupational status distribution for the foreigners obtained from the census survey records (columns 2-3), from the post enumeration survey records (columns 4-5), from the census survey linked records (columns 6-7) and from the post enumeration survey linked records (columns 8-9).

Occupational Status	Census Survey		Post Enum. Survey		Linked records			
					Census Survey		Post Enum. Survey	
	N_i	$f_i \%$	N_i	$f_i \%$	N_i	$f_i \%$	N_i	$f_i \%$
Employed	1,801	58.27	1,635	59.91	1,323	57.22	1,362	61.24
Waiting for first occupation	108	3.49	129	4.73	80	3.46	95	4.27
Unemployed	248	8.02	188	6.89	197	8.52	154	6.92
Waiting to begin job just found	31	1.00	13	0.48	25	1.08	11	0.49
Student	262	8.48	133	4.87	200	8.65	96	4.32
Housewife	494	15.98	497	18.21	388	16.78	411	18.48
Retired	75	2.43	78	2.86	50	2.16	48	2.16
Other conditions	73	2.43	56	2.06	49	1.12	37	1.66
Missing values	584		708		441		539	
Total	3,675	100.00	3,437	100.00	2,753	100.00	2,753	100.00

Table 5: Occupational status distributions for the foreigners in different enlarged databases. (Database DB3 is obtained by enlarging the census file with the records of the post enumeration survey which have not been linked for the analysis of the occupational status. Database DB4 is obtained by enlarging the post enumeration survey file with the census records which have not been linked)

Occupational Status	DB3		DB4	
	N_i	$f_i \%$	N_i	$f_i \%$
Employed	2,074	57.52	2,113	60.23
Waiting for first occupation	142	3.94	157	4.48
Unemployed	282	7.82	239	6.81
Waiting to begin job just found	33	0.92	19	0.54
Student	299	8.29	195	5.56
Housewife	580	16.08	603	17.19
Retired	105	2.91	103	2.94
Other conditions	91	2.52	79	2.25
Missing values	753		851	
Total	4,359	100.00	4,359	100.00

Table 5 gives the occupational status distribution with the two enlarged data set. The dissimilarity index is equal to 0.044. Finally, table 6 shows the capture-recapture estimates for the absolute and relative frequency of the occupational status categorized according to the two classifications employed/non employed

Table 6: *Categorized occupational status distributions for the foreigners obtained by capture-recapture methods drawing the occupational status of the linked records from the census survey (columns 2-3) and from the post enumeration survey (columns 4-5)..*

Occupational Status	Census survey based		Post enumeration survey based	
	N_i	$f_i \%$	N_i	$f_i \%$
Employed	2,172	57.49	2,209	59.73
Non Employed	1,606	42.51	1,489	40.27
Total	3,778	100.00	3,698	100.00

4. Concluding remarks

The application of record linkage techniques with the aim of enlarging the database collected by the Population Census, through the information made available by a planned post enumeration survey, seems to allow a more accurate evaluation of the dimension of response and under-coverage errors. This makes it possible to implement a more pondered data release procedure which considers different levels of detail. A further development of the present work could be represented by the proposal of specific tests to decide whether or not to release statistical data at a certain level of detail, based on specific threshold values of the error. In particular, the application shows that it is advisable to take error sources into account as much as possible, rather than apply only partial adjustments to the estimates. In fact, when both linkage and capture-recapture methods are applied, the distributions resulting from the classification variables registered in the two files show greater coherence. As expected, gender seems to be less affected by measurement and coverage errors than occupational status. Moreover, with respect to the employed, the disagreement resulting from the various figures should suggest caution to the users in interpreting differences that are smaller than 3 or 4%.

Our application did not explicitly consider to what extent linkage errors can affect the disagreement between the various estimates. Further analyses would be to consider the variability due to the uncertainty in the matching status for the pairs coming from different sources, so as to allow better evaluations. The approaches currently considered rely on resampling schemes and montecarlo methods that make use of the matching error probability estimates coming from probabilistic linkage procedure.

References

- Alleva G. (2004) Introduzione alla Sessione su Programmazione e qualità della produzione statistica, in: *VII Conferenza nazionale di statistica, 9-10 novembre 2004*.
- Armstrong J. A., Mayda J. E. (1993) Model-based Estimation of Record Linkage Error Rates, *Survey Methodology*, 19, 137-147.

- Belin, T. R., and Rubin, D. B. (1997) A method for calibrating false-match rates in record linkage, in: *Record Linkage – 1997: Proceedings of an International Workshop and Exposition*, 81-94.
- Biemer P.P. (1988) Modeling Matching Error and its Effect on Estimates of Census Coverage Error, *Survey Methodology*, 14, 117-134.
- Biemer P.P., and Lyberg L.E., (2003) *Introduction to Survey Quality*, Wiley. New York.
- Biemer P.,P. and Stokes S.L., (1991) Approaches to modelling measurement error, in: *Measurement Errors in Surveys*, P. P. Biemer et al. (Eds.) Wiley 487-516.
- Bishop, Y. M .M., Fienberg S. E. and Holland P. W. (1975). *Discrete multivariate analysis: theory and practice*, MA MIT Press, Cambridge.
- Brancato G., D'Orazio M., Fortini M. (2004) Response Error Estimation in Presence of Record Linkage Errors: the case of the Italian Population Census, in: *Proceedings of the European Conference on Quality and Methodology in Official Statistics*. Mainz, 24-26 May 2004. CD ROM.
- Brown D.L., Eaton M.L., Freedman D.A., Klein S.P., Olshen R.A., Watcher K.W., Wells T.M., Ylvisaker D. (1999) Statistical Controversies in Census 2000, *Jurymetrics* 39, 347-375.
- Brown, J., Diamond, I.D., Chambers, R., and Buckner, L. (1999 b) The role of dual system estimation in the 2001 Census Coverage Surveys of the UK, in: *Population Association American Annual Conference*.
- Buratta V., Zannella F. (2004) Programmazione e qualità: una sfida per la statistica ufficiale, in: *VII Conferenza nazionale di statistica, 9-10 novembre 2004*.
- Ding Y., Fienberg S. E. (1994) Dual System Estimation Of Census in Presence of Matching Error, *Survey Methodology*, 20,149-158.
- Fellegi I.P., Sunter A.B. (1969) A Theory for record linkage, *Journal of the American Statistical Association*, 64, 1183-1210.
- Giovannini, E. (2007) Qualità della statistica nella società della conoscenza, *Giornale del Sistan*, 34/2007.
- Hagenaars J.A. (1993) *Loglinear Models With Latent Variables*, Sage Publications Inc
- Istat (2007); *Conoscere il 14° Censimento generale della popolazione e delle abitazioni - La qualità dei dati*, Collana Censimenti, Istat, Roma, (in press).
- Jaro M.A. (1989) Advances in record linkage methodology as applied to matching the 1985 Census of Tampa, Florida, *Journal of the American Statistical Association*, 89, 414-420.
- Lahiri P., Larsen M.,D. (2005), Regression Analysis with Linked Data, *Journal of the American Statistical Association*, 100, 222-230.
- Neter, J., Maynes, E.S., and Ramanathan, R. (1965) The effect of mismatching on the measurement of response errors, *Journal of the American Statistical Association*, 60, 1005-1027.
- Seber, G. A. F. (1982) *The Estimation of Animal Abundance (2nd ed.)*, Edward Arnold London.
- Sekar C. C., Deming W. E. (1949) On the method of estimating birth and death rates and the extent of registration, *Journal of the American Statistical Association*, 101-115.
- Wells T.M., Ylvisaker D. (1999) Statistical Controversies in Census 2000, *Technical Report 537*, Department of Statistics, U.C. Berkeley
- Wolter K.M. (1986) Some coverage error models for census data, *Journal of the American Statistical Association*, 81, 338-346