ISSN 1977-0375

eurostat
Methodologies and
Working papers

Insights on Data Integration Methodologies

ESSnet-ISAD workshop, Vienna, 29-30 May 2008

2009 edition

eurostat
EUROPEAN COMMISSION

eurostat
Methodologies and Working papers

Insights on Data Integration Methodologies

ESSnet-ISAD workshop, Vienna, 29-30 May 2008

2009 edition

eurostat
EUROPEAN COMMISSION

Luxembourg: Office for Official Publications of the European Communities, 2009

ISBN 978-92-79-12306-1
ISSN 1977-0375
DOI 10.2785/20079
Cat. No. KS-RA-09-005-EN-N

Theme: General and regional statistics
Collection: Methodologies and working papers

Table of contents

Foreword

Data integration is a methodological area which represents a valid response to a major challenge for NSIs in the ESS. Current informative needs for official statistics require an increasingly sophisticated use of multiple sources for the production of statistics while budgetary constraints and increasing public concern about data privacy and burden on respondents are rising.

The ESSnet project in the area of Integration of Survey and Administrative data (ESSnet ISAD) finalised in June 2008 represents a first attempt to create a common methodological basis for the application of statistical methodologies for the integration of different sources. It aimed at reviewing and promoting knowledge and application of sound methodologies for the joint use of available information in existing data sources for the production of official statistics.

The main findings of the project were presented in the final workshop which took place in Vienna on 29-30 June 2008. The workshop gave the opportunity to bring together experts from the ESS and universities on the topic of data integration. Different experiences from NSIs were presented and new insights and methodological solutions proposed by key academic and NSI's researchers were proposed and discussed.

Eurostat is proud of the results achieved by the ISAD project which put together expertises on a relatively new domain which were isolated thorough the ESS. The objectives to share know-how, to identify common issues and to develop knowledge which can be beneficial for the whole system, were successfully achieved.

Drawing on the workshop output, this document reviews different experiences and presents methodological solutions and directions for future work in the domain of data integration. It aims to raise awareness of the topic and to the transfer of knowledge in the ESS.

Pedro Díaz Muñoz

Director

Acknowledgements

The European Commission expresses its gratitude and appreciation for the work carried out by the members of the ESSnet ISAD: Mauro Scanu (coordinator), Nicoletta Cibella, Paolo Consolini, Marco Di Zio, Marcello D'Orazio, Marco Fortini, Monica Scannapieco, Tiziana Tuoto (ISTAT); Eric Schulte Nordholt (CBS); Jaroslav Kraus, Robert Šanda, Ondřej Vozár (CZSO); Francisco Hernández, Esteban Barbado, Gervasio Fernández, Miguel Guigó (INE); Alois Haslinger (STAT). Special thanks also to the other colleagues of ISTAT, INE, CSZO, STAT and CBS who also have contributed to the realisation of the project.

Finally the European Commission thanks all the other NSIs which participating to the initiatives proposed the ESSnet ISAD members in the different phases of the project have indirectly contributed to its success.

Introduction

The idea of establishing ESSnets in the field of Statistics was launched in 2005 as a way to reinforce cooperation between National Statistical Institutes. In this way the various institutes in Europe could benefit from each other experiences and together raise the level of their statistical production process.

The ESSnet – ISAD (Integration of Surveys and Administrative Data) project is a one-and-a-half year project partially funded by Eurostat whose activities began in December 2006 and ended in June 2008. An overview of the activities in this ESSnet can be found at the ESSnet website (http://cenex-isad.istat.it). The institutions involved in the project are:

- ISTAT (Italy, coordinator),
- CBS (The Netherlands),
- CZSO (Czech Republic),
- INE (Spain),
- STAT (Austria).

Among the project activities there was a workshop on Integration of Surveys and Administrative data. This workshop was held at the STAT premises in Vienna, 29-30 May 2008.

The aim of the workshop was:

1. to disseminate the ESSnet-ISAD project results in the ESS;
2. to allow presenting and discussing the different experiences in the ESS MS;
3. to gather researchers from the ESS countries interested in the topic, and create a community of experts;
4. to present and compare new insights and methodological solutions from ESS NSIs and key academic researchers.

The workshop was attended by 59 invited attendees from 24 countries (plus Eurostat personnel).
The workshop consisted of 7 sessions:

Session 1	ESSnet ISAD results
Session 2	Record linkage
Session 3	Statistical matching and forecasting
Session 4	Conceptual aspects for integration
Session 5	Integration of registers and samples 1
Session 6	Integration of registers and samples 2
Session 7	Register based statistics

The papers presented during the workshop were both methodological papers and practical applications. The topic of integration of surveys and administrative data confirmed to be a very active area of research. Innovative solutions were discussed clearly in the papers. Great emphasis was also given to open problems that still need to be appropriately tackled.

We truly hope that these papers can be a stimulus for further research in the topic, and a way to reinforce cooperation between the NSIs in the ESS as well as with academic institutions.

Workshop material is freely available on the project website (http://cenex-isad.istat.it).

Mauro Scanu (ISTAT, project coordinator)
Alois Haslinger (STAT)
Francisco Hernández Jiménez (INE)
Jaroslav Kraus (CZSO)
Eric Schulte Nordholt (CBS)

1

Record linkage

Record linkage, correct match probabilities and disclosure risk assessment

Chris Skinner
University of Southampton, Southampton SO17 1BJ, United Kingdom
e-mail: C.J.Skinner@soton.ac.uk

Abstract: The use of probabilistic record linkage methods in disclosure risk assessment for microdata is outlined. The disclosure risk is defined as the probability of identification, which is the probability that a match obtained from the record linkage method is correct. The nature of this probability and its estimation is investigated in the context of disclosure risk assessment. There is a particular focus on the impact of misclassification of identifying variables.

Keywords: confidentiality; identification; microdata; misclassification; probabilistic record linkage.

1. Introduction

Record linkage has many valuable uses in official statistics, but it also represents a threat. It is conceivable that the confidentiality of data, made available by a statistical agency to researchers for valid analytical purposes, might be compromised if an 'intruder' with access to these data succeeded in linking the data to some external data source on known individuals (or other units). Fienberg (2006) suggests that this threat is growing.

In this paper we consider how the agency might assess the risk of disclosure arising from such a threat. We suppose the agency wishes to release an 'anonymised' microdata file, consisting of records for a sample of individuals, for which each record contains the values of various variables of research interest but no direct identifiers, such as name or address. The agency is concerned that an intruder might 'identify' one of these records by linking it to an external data source on known individuals, using a subset of the variables, and that this could enable the intruder to disclose the values of the remaining variables on one or more known individuals. The aim of the paper is to consider how the agency may assess the *risk of identification* (c.f. Reiter, 2005; Skinner, 2007), which we define as the probability of a correct match. We note that false matches may also be of concern to an agency (Lambert, 1993), but it is not possible to control statistically the opportunity for intruders to make erroneous guesses and we restrict attention in this paper to correct matches. One context where risk assessment is often used in practice is to support decisions about the choice of statistical disclosure control methods to apply to the microdata to obtain a masked file and the choice of 'parameters' of these methods, e.g. the degree of masking (Willenborg and de Waal, 2001).

A basic difficulty faced by the agency in its assessment of identification risk is that the external data source is unknown. This may lead the agency to constructing a 'surrogate' external file. Options include using:

- genuine external files, e.g. FCSM (2005) reports that the US National Center for Education Statistics uses certain commercially available school files;

- other datasets the agency collects;
- the original unmasked microdata file as a highly conservative option, when a masked file is to be released (Winkler, 2004);
- one or more sets of *key variables,* i.e. variables which might be matchable to an external file, and then constructing a synthetic surrogate file (just for these variables) from the microdata, typically employing modelling assumptions.

Since the surrogate file may be constructed from the original file, the term 're-identification' is often used in risk assessment (e.g. Lambert, 1993). Given a surrogate file and the microdata file to be released, there are at least three ways in which an agency could assess the probability of a correct match for a given record linkage technique applied to the files:

(1) empirical proportion of claimed matches which are correct (termed *the risk of true identification* by Lambert, 1993);
(2) Bayesian posterior probabilities of identification employing the intruder's prior probabilities (termed *the risk of perceived identification* by Lambert, 1993);
(3) model-based probabilities of a correct match which may be determined by the intruder and are publicly defensible.

We reject approach (1) as a primary approach, since it can fail to control adequately for the information available to the intruder. In particular, the intruder could not determine this proportion since it requires knowledge of the true identities of the records in the microdata, information unavailable to the intruder. Suppose, for example, that the overall proportion of correct matches is 5% and that the agency considers this sufficiently low. Suppose, however, that the intruder could determine which 5% of his claimed matches are correct and which 95% are incorrect. Then the intruder could claim some matches with 100% confidence and this might be deemed an unacceptable disclosure risk. On the other hand, suppose the agency chooses to calculate its proportions separately according to different areas and observes that the proportions vary across areas from 0% to 70%. It might deem the release of data for those areas with proportions as high as 70% as unacceptable. However, if the intruder could only determine that the overall rate of a correct match was 5% and was unable to identify areas where it was higher, the agency's judgment would seem over-conservative.

We also reject approach (2) on the same grounds that we decided not to consider false matches, i.e. since 'the agency cannot control the intruder's perceptions' (Lambert, 1993, p.322).

Our preferred approach is (3). We define the probability of a correct match as the probability conditional on data which is assumed available to the intruder. And we require that this probability can be estimated reliably (in a sense to be discussed) from these data. If an intruder is not able to present sufficient evidence that a claimed match is correct, we take the view that this is not of a matter of concern to the agency. We suppose that the agency might use empirical proportions of correct matches as a means of validating these probabilities but not as the primary source, as in (1).

These considerations differ somewhat from discussions of false matches in the record linkage literature. In conventional applications of record linkage, incorrect matches (false positives or false negatives) are only of interest because of their statistical consequences for samples as a whole. Fellegi and Sunter (1969, p. 1196) state that 'we are not concerned with the *probability* of [these two kinds of erroneous matches]…but rather with the *proportion* of occurrences of these two events in the long run'. Belin and Rubin (1995) similarly focus on false match rates for a linkage procedure. In contrast, requirements to protect the confidentiality of every individual imply that an agency may be interested in the probability of a correct match for a single individual.

So far, our discussion has related to any record linkage technique. We shall focus in the remainder of the paper on probabilistic record linkage techniques as developed by Fellegi and Sunter (1969, hereafter referred to as FS) and others e.g. Jaro (1989). Other kinds of record linkage, especially of a distance function minimisation form, have also been considered in relation to disclosure risk assessment (Domingo-Ferrer and Torra, 2003).

2. The problem: the threat of disclosure from record linkage

Consider a survey microdata file containing records for a sample of responding units s_1 drawn from a finite population P. Each record will typically include a large number of variables, needed for substantive research, but will not include directly identifying variables like name and address. Suppose an intruder has access to this file and wishes to identify one or more units in s_1, with the ultimate aim of disclosing the values of the variables about these units. Suppose the intruder seeks to achieve this by matching the file to an external file of records for another sample of units $s_2 \subset P$, for which the identities are known and for which it is feasible that the intersection $s_{12} = s_1 \cap s_2$ is non-empty. (We assume here that the definition of the population P is public and that the intruder can thus remove from consideration any records in the external file which do not belong to P – hence we do not require the full generality of the approach of FS, which allows s_1 and s_2 to be drawn from different populations.)

We suppose the matching is undertaken based upon the values of variables, which appear in both files. These variables are often called *key variables* in the disclosure control literature (Bethlehem et al., 1990) and the match key (Herzog et al., 2007, p.82) or the match fields (Jaro, 1995, p.492) in the record linkage literature. Let $\tilde{X}_a$ denote the value of the vector of key variables for unit a in the microdata ($a \in s_1$) and X_b the corresponding value for unit b recorded in the external database ($b \in s_2$). The difference in notation between $\tilde{X}$ and X allows for the possibility that the variables are recorded in a different way in the two data sources. These variables might differ for various reasons, including measurement error (in either source) or the application of a perturbative statistical disclosure control method to the microdata file. Following FS, we suppose the intruder undertakes linkage by calculating a comparison vector $\gamma(\tilde{X}_a, X_b)$ for pairs of units $(a,b) \in s_1 \times s_2$, where the function $\gamma(.,.)$ may take values in some finite comparison space Γ.

Example 1: Exact matching on categorical key variables

Suppose that $\tilde{X}$ and X may only take K possible values, denoted $\{1,...,K\}$ without loss of generality. Let $\Gamma = \{1,2,...,K+1\}$ and define the comparison vector by

$$\gamma(\tilde{X}, X) = j \text{ if } \tilde{X} = X = j,\ \ j = 1,2,...,K$$

$$\gamma(\tilde{X}, X) = K + 1 \text{ otherwise.}$$

In this case, an intruder might consider any pair $(a,b) \in s_1 \times s_2$ for which $\gamma(\tilde{X}_a, X_b) \leq K$ as a potential match, but rule out of consideration any pair for which $\gamma(\tilde{X}_a, X_b) = K + 1$.

We suppose the intruder seeks to use the comparison vectors to identify one or more pairs $(a,b) \in s_1 \times s_2$ which contain identical units, i.e. are of the form (a,a) where $a \in s_{12}$. Since the number of pairs in $s_1 \times s_2$ may be very large, we suppose the intruder only considers pairs which fall in a set $\tilde{s} \subset s_1 \times s_2$. We discuss the choice of $\tilde{s}$ further in section 4. We partition $\tilde{s}$ into $M = \{(a,b) \in \tilde{s} \mid a = b, a \in s_{12}\}$, the set of pairs of common units, and $U = \{(a,b) \in \tilde{s} \mid a \in s_1, b \in s_2, a \neq b\}$, the set of pairs of different units. The problem faced by the intruder is how best to use comparison vector values to classify pairs from $\tilde{s}$ into M or U. An optimum strategy is shown by FS to be based upon a comparison of the probability distributions of the comparison vector between M or U, i.e. a comparison of

$$m(\gamma) = \Pr[\gamma(\tilde{X}_a, X_b) = \gamma \mid (a,b) \in M] \;, \qquad \text{and} \qquad (1)$$

$$u(\gamma) = \Pr[\gamma(\tilde{X}_a, X_b) = \gamma \mid (a,b) \in U] \;, \qquad \gamma \in \Gamma . \qquad (2)$$

We discuss the nature of these probabilities in the next section. FS show that an optimal approach for the intruder is to order pairs in $\tilde{s}$ according to the likelihood ratios $m(\gamma)/u(\gamma)$, treating pairs with higher values of this ratio as more likely to belong to M. Our aim in this paper is to explore the probability of a correct match for pairs selected in this way.

3. The probability of a correct match

Given a pair (a,b), linked according to a record linkage approach as defined above in terms of in (1) and (2), it is usual to define the probability that the pair represents a correct match, that is $a = b$, as $p_{M|\gamma} = \Pr[(a,b) \in M \mid \gamma(\tilde{X}_a, X_b)]$, i.e. the conditional probability that the pair is in M given that it is in $\tilde{s}$ and that the comparison vector takes the value γ. To express $p_{M|\gamma}$ in terms of $m(\gamma)$ and $u(\gamma)$, let:

$$p = \Pr[(a,b) \in M], \qquad (3)$$

be the probability that the pair is in M given that it is in $\tilde{s}$ and, using Bayes theorem, we obtain

$$p_{M|\gamma} = m(\gamma)p / [m(\gamma)p + u(\gamma)(1-p)] \quad . \qquad (4)$$

Sorting pairs according to this 'posterior' probability is equivalent to sorting pairs according to the likelihood ratio $m(\gamma)/u(\gamma)$. From the statistical disclosure control perspective, expression (4) may be interpreted as the identification risk for a pair (a,b), i.e. the probability that a and b are identical, given the value of the comparison vector. From the record linkage perspective, expression (4) is the probability of a correct match or alternatively one minus the probability of a false match (Belin and Rubin, 1995).

The expressions in (1), (2) and (3) are, of course, dependent on the way the probabilities are defined. Our primary approach in this paper is to suppose that the probabilities are defined with respect to the following three processes:
(i) a random selection (with equal probability) of the pair (a,b) from $\tilde{s} = M \cup U$;
(ii) a random process of generating $\tilde{X}_a$;
(iii) a specified random probability design for the selection of s_1 from P ;
where the population P and the values X_a for units in the population are treated as fixed. Evaluating the probabilities over (i), holding s_1 and the $\tilde{X}_a$ fixed, we may write

$$m(\gamma) = E[n_{M\gamma} / n_M] \ , \ u(\gamma) = E[n_{U\gamma} / n_U] \ , \tag{5}$$

where n_M and n_U are the numbers of pairs in M and U respectively, $n_{M\gamma}$ and $n_{U\gamma}$ are the corresponding numbers of these pairs for which the comparison vector takes the value γ and the expectation is taken with respect to (ii) and (iii). We may thus interpret $m(\gamma)$ and $u(\gamma)$ as the expected relative frequencies of the different comparison vectors within M and U respectively. Similarly, we may write

$$p = E(n_M / \tilde{n}) \, , \tag{6}$$

where $\tilde{n}$ is the number of pairs in $\tilde{s}$ and the expectation is with respect to (iii).

To explore the form of $p_{M|\gamma}$ further under (i), (ii) and (iii), we consider two special cases.

Example 1 with no misclassification

Suppose that exact matching is used as in Example 1. Suppose that: $\tilde{X}_a = X_a$ for all units $a \in P$ (i.e. no misclassification); $s_2 = P$ and $\tilde{s} = s_1 \times s_2$. Let $n_1 = | s_1 |$ and $N = | P |$. Noting that $n_M = n_1$ and $\tilde{n} = n_1 N$, we obtain from (5) and (6):

$$m(j) = E[f_j / n_1] \ , \qquad u(j) = E\left(\frac{f_j (F_j \text{-} 1)}{n_1 (N - 1)} \right) , \qquad j = 1, \ldots, K$$

$$p = E[n_1 / (n_1 N)] = 1 / N \, , \tag{7}$$

where f_j and F_j are the numbers of units with $X_a = j$ in s_1 and P respectively. Using Bayes theorem we have:

$$\Pr[(a,b) \in M \mid \gamma(\tilde{X}_a, X_b) = j] = E\left(\frac{f_j / (n_1 N)}{f_j / (n_1 N) + f_j (F_j - 1) / (n_1 N)} \right) = 1 / F_j \, . \tag{8}$$

Note that this result if free of any assumptions about the sampling scheme. Expression (8) is familiar in the disclosure risk literature (e.g. Duncan and Lambert, 1989). It is common to argue, however, that agencies should design release strategies so that an intruder could not know the value of F_j from external information (Skinner, 2007). Note that, in particular, this requires assuming that $s_2 \neq P$. Otherwise, the intruder could determine F_j from knowledge of X_a for $a \in P$. If F_j is unknown to the intruder, the uncertainty about F_j needs to be integrated out of the expression for the identification risk, subject to conditioning on the information available to the intruder. This integration is most naturally done by revising the probability mechanisms (i)-(iii) above to include a process which generates the values X_a for units in the population. Under this extended probability mechanism, the identification risk becomes $E(1/F_j \mid data)$, where *data* represents the data available to the intruder. We shall return to this issue in Section 5. First, we extend the result in (8) to the case when $\tilde{X}_a$ may be derived from X_a by a process of misclassification and s_2 may be any proper subset of P.

Example 1 with misclassification

Suppose again that exact matching is used and that $\tilde{s} = s_1 \times s_2$. We consider two extensions to the previous result. First, we allow s_2 to be any proper subset of P. Second, we suppose that each $\tilde{X}_a$ is determined from X_a as follows

$$\Pr(\tilde{X}_a = j \mid X_a = k) = \theta_{jk} \text{ , for all } a \in P \text{ ,} \tag{9}$$

where θ_{jk} is an element of a misclassification matrix with columns which sum to 1. We now obtain

$$m(j) = E[f_j^{12} / n_{12}] \text{ ,} \qquad u(j) = E\left(\frac{\tilde{f}_j f_{\mathrm{j}} - f_j^{12}}{n_1 n_2 - n_{12}} \right) \text{ ,} \qquad j = 1, \ldots, K$$

$$p = E[n_{12} / (n_1 n_2)] \text{ ,}$$

where f_j^{12} is the number of units in s_{12} with $X_a = j$ and $\tilde{X}_a = j$, $\tilde{f}_j$ is the number of units in s_1 with $\tilde{X}_a = j$ and f_j is the number of units in s_2 with $X_a = j$. If we suppose that Bernoulli sampling is employed with inclusion probability π we have $n_{12} \doteq n_2 n_1 / N$ so that $p \doteq 1/N$ and $n_1 n_2 - n_{12} \doteq (N-1) n_{12}$.

It follows that

$$\Pr\left[(a,b) \in M \middle| \gamma\left(\tilde{X}_a, X_b\right) = j\right] \doteq E\left(\frac{f_j^{12} / \left(n_{12} N\right)}{f_j^{12} / \left(n_{12} N\right) + \left\{\left(\tilde{f}_j f_j - f_j^{12}\right) / \left[n_{12}\left(N-1\right)\right]\right\}\left\{\left[\left(N-1\right)/N\right]\right\}} \right)$$

$$= E\left(\frac{f_j^{12}}{\tilde{f}_j f_j} \right) \text{,}$$

where the expectation is with respect to both the sampling and the misclassification mechanisms. $E(f_j^{12}) = \pi\theta_{jj} f_j$ and $E(\tilde{f}_j) = \pi\tilde{F}_j$, where $\tilde{F}_j$ is the number of units in P with $\tilde{X}_a = j$ (imagining that the misclassification takes place before the sampling). Hence we may write

$$\Pr[(a,b) \in M \mid \gamma(\tilde{X}_a X_b) = j] \doteq \frac{\theta_{jj}}{\tilde{F}_j} \ . \tag{10}$$

Note that this expression applies for any choice s_2, which may be selected arbitrarily.

The expression in (4) for the probability of a correct match and the special cases in (8) and (10) apply to a pair of records (a,b) with a specific agreement pattern γ. This notion may be extended to apply to a class of pairs, $\hat{M}$, for which the likelihood ratio is above some threshold, say $\hat{M} = \{(a,b) \mid \gamma(\tilde{X}_a, X_b) \in \Gamma_M\}$, where Γ_M is the set of agreement patterns γ for which $m(\gamma)/u(\gamma)$ is above a threshold specified by the intruder as determining which pairs to declare as links. The empirical *correct match rate* for this approach is the number of correctly matched pairs divided by the number of declared match pairs (c.f. Larsen and Rubin, 2001).

A key issue for identification risk assessment is how to estimate the probability $p_{M\mid\gamma}$ and, more specifically, how to estimate $p, m(\gamma)$ and $u(\gamma)$. We shall discuss this in section 5. Before then, we expand on the nature of the record linkage approach.

4. Taking account of key variable structure

In section 2 the comparison vector $\gamma(\tilde{X}_a, X_b)$ was defined generally. In practice, it is usual that $\gamma(\tilde{X}_a, X_b)$ is based upon the separate comparisons of C key variables. We write $\tilde{X} = (\tilde{X}^1, ..., \tilde{X}^C)$ and $X = (X^1, ..., X^C)$ and suppose that

$$\gamma(\tilde{X}_a, X_b) = [\gamma^1(\tilde{X}_a^1, X_b^1), ..., \gamma^C(\tilde{X}_a^C, X_b^C)], \tag{12}$$

where $\gamma^c(\tilde{X}^c, X^c)$ denotes the comparison vector (or scalar) for the c$^{\text{th}}$ key variable. A simple and commonly used approach (c.f. Larsen and Rubin, 2001) to defining $\gamma^c(.,.)$ is as follows.

Example 2 Comparison vectors for simple agreements between key variables
Suppose that each $\gamma^c(\tilde{X}^c, X^c)$ is a binary indicator of agreement between $\tilde{X}^c$ and X^c, i.e. $\gamma^c(\tilde{X}^c, X^c) = 1$ if $\tilde{X}^c \sim X^c$ and $\gamma^c(\tilde{X}^c, X^c) = 0$, otherwise, $c = 1, 2, ..., C$, where $\sim$ is a specified agreement relation. With categorical key variables this relation may simply

be one of equality. With continuous variables $\tilde{X}^c \sim X^c$ may indicate that$|\tilde{X}^c - X^c| < \varepsilon$ for a specified value ε. Define $\gamma(\tilde{X}_a, X_b)$ as in (11). Then

$$\Gamma = \{(\gamma^1, \gamma^2, ..., \gamma^C) \mid \gamma^c = 0,1;\ c = 1,2,...,C\} = \{0,1\}^C \text{ and } |\Gamma| = 2^C .$$

A more complex choice, building on Example 1 is as follows.

Example 3 Comparison vectors for agreements between categorical key variables
Suppose that each of $\tilde{X}^c$ and X^c is categorical, taking t_c possible values, denoted without loss of generality $j^c = 1,2,...,t^c$, and that $\gamma^c(.,.)$ is defined analogously to Example 1: $\gamma^c(\tilde{X}^c, X^c) = j^c$ if $\tilde{X}^c = X^c = j^c$, $j^c = 1,2,...,t^c$, $\gamma^c(\tilde{X}^c, X^c) = t^c + 1$ otherwise, $c = 1,2,...,C$. Again, define $\gamma(\tilde{X}_a, X_b)$ as in (12). Then

$$\Gamma = \{(\gamma^1, \gamma^2, ..., \gamma^C) \mid \gamma^c = 1,...,t^c + 1, c = 1,2,...,C\} \text{ and } |\Gamma| = \prod_{c=1}^{C} (t^c + 1) .$$

Given the large potential size of Γ when C is at all large, it is common to restrict attention to a subspace Γ^* of Γ. A common approach, following FS, is to block.

Example 4 Blocking
Partition the key variables into two subsets, $X^1,...,X^D$ and $X^{D+1},...,X^C$ and suppose that the set of possible values of $X^{D+1},...,X^C$ (or equivalently of $\tilde{X}^{D+1},...,\tilde{X}^C$) is partitioned into blocks (e.g. Jaro, 1995). We then assume that the set $\tilde{s}$ of pairs examined by the intruder for matching only includes pairs $(a,b) \in s_1 \times s_2$ for which $\tilde{X}_a^{D+1},...,X_a^C$ and $X_b^{D+1},...,X_b^C$ fall in the same block. For example, if $X^{D+1},...,X^C$ are categorical and the intruder only considers pairs which match exactly on each of $X^{D+1},...,X^C$ then the blocks will consist of the different combinations of categories of these variables. Restricting attention to pairs falling into the same block will typically be equivalent to imposing a restriction on the possible values of $(\gamma^{D+1},...,\gamma^C)$, and hence defining Γ^* as a proper subset of Γ.

The probabilities $m(\gamma)$ and $u(\gamma)$ in (1) and (2) play a fundamental role in probabilistic record linkage and their estimation is clearly challenging if $|\Gamma|$ is large, as is likely in Examples 2 and 3 if C is at all large. It is therefore common to make simplifying assumptions, in particular, following FS, that the agreement patterns $\gamma^1(\tilde{X}_a^1, X_b^1),...,\gamma^C(\tilde{X}_a^C, X_b^C)$ in (11) may be treated as independent within M and U, i.e. it is assumed that

$$m(\gamma) = m_1(\gamma^1)m_2(\gamma^2)...m_C(\gamma^C) \text{ and } u(\gamma) = u_1(\gamma^1)u_2(\gamma^2)...u_C(\gamma^C) \qquad (13)$$

where

$$m_c(\gamma^c) = \Pr[\gamma^c(\tilde{X}_a^c, X_b^c) = \gamma^c \mid (a,b) \in M]$$

and

$$u_c(\gamma^c) = \Pr[\gamma^c(\tilde{X}_a^c, X_b^c) = \gamma^c \mid (a,b) \in U],$$

$c = 1, 2, \ldots, C$. We refer to this assumption as *independence of agreement patterns***.**

Example 3 (continued)
In the categorical variable case considered earlier we may write:

$$\Pr[\gamma^c(\tilde{X}_a^c, X_b^c) = j^c] = \Pr[\tilde{X}_a^c = X_b^c = j^c], \qquad j^c = 1, 2, \ldots, t^c.$$

A sufficient condition for the independence of agreement patterns is that misclassification operates independently, variable by variable, and that the key variables are themselves independent. Under these assumptions we have:

$$\Pr(\tilde{X}_a = j \mid X_a = k) = \theta_{jk} = \theta^1_{j^1k^1}\theta^2_{j^2k^2}\ldots\theta^C_{j^Ck^C},$$

where $\theta^c_{j^ck^c} = \Pr[\tilde{X}^c = j^c \mid X^c = k^c]$. And, following FS (sect. 3.3.1) we have:

$$m_c(j^c) = \theta^c_{j^cj^c}\Pr[X_a^c = j^c \mid a \in s_{12}], \tag{14}$$

$$u_c(\gamma^c) = \Pr[\tilde{X}_a^c = j^c \mid a \in s_1]\Pr[X_b^c = j^c \mid b \in s_2],\ j^c = 1, 2, \ldots, t^c. \tag{15}$$

5. Estimation

We now consider the problem of estimating the probability of a correct match, as defined in section 3. One potentially attractive option, in particular because it could be applied to any record linkage technique, would be to use the empirical match rate. However, we reject this as a method for directly estimating identification risk for the reasons given in section 1. Instead we consider approaches which estimate $p_{M|\gamma}$, as defined in (6) and (7), for a given observed value γ of the comparison vector for a pair for which a match might be claimed.

We assume that the estimation of $p_{M|\gamma}$ can only use data which may be available to the intruder and thus, in particular, rule out the possibility of using a training sample (c.f. Belin and Rubin, 1995).

We distinguish two possible broad approaches:
(a) a *direct* approach, where either $p_{M|\gamma}$ or its components $p, m(\gamma)$ and $u(\gamma)$ are expressed in terms of the underlying data generation process, the parameters of this process are clarified and perhaps approximated theoretically and then estimated, possibly using external sources (c.f. FS Method 1);

(b) a *mixture model* approach: where $p, m(\gamma)$ and $u(\gamma)$ are treated as unknown parameters in a model for the observed values of the comparison vectors. The model is a mixture of models for M and U, treated as latent classes, and maximum likelihood estimation is used for parameter estimation (e.g. FS Method 2; Jaro, 1989; Larsen and Rubin, 2001).

We only consider the direct approach further here. The mixture model approach has found some success in record linkage applications where very strong identifying information, such as name and address, is available. On the other hand, it has been less successful when the distributions of the comparison vectors for M and U are not well-separated or are not each unimodal (Larsen and Rubin, 2001; Herzog et al., 2007) and this seems more likely to be the case in practice in the disclosure control context, at least for social survey data. In any case, further exploration of this idea seems worthwhile but is not attempted here.

We consider the direct approach first by returning to expressions for $p_{M|\gamma}$ in two examples in Section 3. We then discuss the estimation of $p, m(\gamma)$ and $u(\gamma)$.

Example 1 with no misclassification

In this case we obtained $p_{M|\gamma} = 1/F_j$ in expression (8) but argued, following this expression, that a more suitable measure will usually be $E(1/F_j \mid data)$. Skinner and Shlomo (2008) have discussed the evaluation of this conditional expectation under the assumption that the F_j are generated from a Poisson log-linear model and that the sample frequencies f_j represent the *data*. Treating the pairs (f_j, F_j) as independent, the conditional probability may then be expressed as $E(1/F_j \mid f_j)$ and a closed form expression may be obtained under the Poisson log-linear model and a Bernoulli sampling assumption. The conditional probability will be highest for cases which are unique in the sample, i.e. $f_j = 1$. The conditional probability may be estimated by estimating the log-linear model parameters and plugging these estimates into the expression for the conditional probability.

Example 1 with misclassification

In this case we obtained the approximate expression $p_{M|\gamma} \doteq \theta_{jj} / \tilde{F}_j$ in expression (10). As above, we may argue that in practice $\tilde{F}_j$ will be unknown and a more suitable measure is $\theta_{jj} E(1/\tilde{F}_j \mid \tilde{f}_j)$. The second component of this expression, $E(1/\tilde{F}_j \mid \tilde{f}_j)$, may be estimated by applying the methodology of Skinner and Shlomo (2008) to the observed microdata. We comment on the estimation of θ_{jj} below.

Let us now consider the separate estimation of $p, m(\gamma)$ and $u(\gamma)$. Consider p first. If $\tilde{n}$ is large we have from (6) that $p \doteq n_M / \tilde{n}$. The intruder knows the value of $\tilde{n}$ and so needs to estimate n_M in order to estimate p. We know $n_M \le n_{12}$, where $n_{12} = |s_{12}|$. And if we take the worst case, where the intruder selects $\tilde{s}$ in such a way that it includes all possible common pairs (i.e. all (a,a) where $a \in s_{12}$) then we have $n_M = n_{12}$. Thus, in order to estimate p, it suffices to estimate n_{12}. We *s*uppose the intruder can determine inclusion probabilities $\pi_i = \Pr(i \in s_1)$ for $i \in s_2$. This is plausible. Often inclusion probabilities are equal or else they will vary by strata which may be known for units in s_2. Since we have $n_{12} = E(\sum_{i \in s_2} \pi_i)$, where the expectation is with respect to the sampling scheme for s_1, the intruder can estimate n_{12} by $\hat{n}_{12} = \sum_{i \in s_2} \pi_i$ and hence estimate p by $\hat{p} = \hat{n}_{12} / \tilde{n}$. Often in social surveys the inclusion probabilities π_i will be small, say 1/10,000, and so $\hat{n}_{12}$ is only likely to be a reasonable estimator (with reasonable relative precision) if the size of the external database is large, representing a substantial proportion of the population. Note also that some adjustment will usually be necessary for nonresponse (most simply by multiplying π_i by a response rate).

Let us now turn to the estimation of $m(\gamma)$ and $u(\gamma)$. Consider Example 1 with misclassification again, where we wish to estimate $m(j)$ and $u(j)$ for $j = 1,...,K$. We may write $m(j) = \theta_{jj} E[n_{12j} / n_{12}]$, where n_{12j} is the number of units in s_{12} with $\gamma = j$. And under Bernoulli (or equal probability) sampling we may write $E[n_{12j} / n_{12}] = f_j / n_2$, so that $m(j) = \theta_{jj} f_j / n_2$. And to first approximation (Jaro, 1989) we have:

$$u(j) = (\tilde{f}_j / n_1)(f_j / n_2) .$$

To estimate $p_{M|\gamma}$ in (4) we only need to estimate the ratio $m(j)/u(j)$, which we may approximate in this case by $m(j)/u(j) = \theta_{jj} /(\tilde{f}_j / n_1)$. The quantities $\tilde{f}_j$ and n_1 are known from the microdata so the intruder just requires an estimate of θ_{jj}. This might be obtained from some approximating assumptions and external evidence on the misclassification process.

One first assumption may be that some of the key variables are subject to no misclassification, as is commonly assumed for blocking variables, and that misclassification on the remaining variables is not dependent upon the values of such correctly classified variables.

A further assumption may be that the remaining key variables are misclassified independently. This may be related to but is not the same as the earlier assumption of independence of agreement patterns. That assumption would follow if different key variables are misclassified independently and if the key variables are independent. Under the independence of misclassification assumption, θ_{jj} may be expressed as a product of correct classification probabilities for the different key variables. This may need to be modified to allow for the possibility that the values of some key variables are missing.

6 Further research

The work in this paper requires development in a number of ways:

- alternative assumptions about the process underlying the probability definition merit consideration, in particular it seems desirable to weaken the assumption underlying the first component of this process considered here, i.e. that the pair (a,b) is drawn randomly with equal probability from $\tilde{s} = M \cup U$;

- alternative record linkage methods could be considered, in particular deterministic distance minimisation methods;
- numerical evaluation work is needed to assess the properties of the estimated probabilities and their dependence upon approximating assumptions in realistic settings, in particular to explore the potential underestimation of the false match probability, discussed in Belin and Rubin (1995).

References

Belin T.R., Rubin D.B. (1995) A method for calibrating false-match rates in record linkage, *Journal of American Statistical Association*, 90, 694-707.

Bethlehem J.G., Keller W.J., Pannekoek J. (1990) Disclosure control for microdata, *Journal of the American Statistical Association,* 85, 38-45.

Domingo-Ferrer J., Torra V. (2003) Disclosure risk assessment in statistical microdata protection via advanced record linkage, *Statistics and Computing*, 13, 343-354.

Duncan G., Lambert D. (1989) The risk of disclosure for microdata, *Journal of Business and Economic Statistics*, 7, 207-217.

Federal Committee on Statistical Methodology (2005), Statistical Policy Working Paper 22 (2nd Version): *Report on Statistical Disclosure Limitation Methodology*, Office of Management and Budget, Washington, D.C.

Fellegi I.P., Sunter A.B. (1969) A theory for record linkage, *Journal of American Statistical Association*, 64, 1183-1210.

Fienberg S.E. (2006) Privacy and confidentiality in an e-commerce world: data mining, data warehousing, matching and disclosure limitation, *Statistical Science*, 21, 143-154.

Herzog T.N., Scheuren F.J., Winkler W.E. (2007) *Data Quality and Record Linkage Techniques*. New York: Springer.

Jaro M.A. (1989) Advances in record linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of American Statistical Association,* 84, 414-420.

Jaro M.A. (1995) Probabilistic linkage of large public health data files. *Statistics in Medicine*, 14, 491-498.

Lambert D. (1993) Measures of disclosure risk and harm. *Journal of Official Statistics,* 9, 313-331.

Larsen M.D., Rubin D.B. (2001) Iterative automated record linkage using mixture models, *Journal of American Statistical Association,* 96, 32-41.

Reiter J. (2005) Estimating risks of identification disclosure in microdata, *Journal of the American Statistical Association*, 100, 1103-1112.

Skinner C.J. (2007) The probability of identification: applying ideas from forensic science to disclosure risk assessment. *Journal of the Royal Statistical Society, Series A*, 170, 195-212.

Skinner C.J., Shlomo, N. (2008) Assessing disclosure risk in survey microdata using log-linear models, *Journal of American Statistical Association,* to appear.

Willenborg L., De Waal T. (2001), *Elements of Statistical Disclosure Control*, Vol. 155, Lecture Notes in Statistics, New York: Springer.

Winkler W. (2004) Masking and re-identification methods for public use microdata: overview and research problems. In J. Domingo-Ferrer and V. Torra (eds.) *Privacy in Statistical Databases*. Lecture Notes in Computer Science 3050, Berlin: Springer, 231-246.

Model based record linkage: a Bayesian perspective

Brunero Liseo, Andrea Tancredi
Dip. di studi geoeconomici, linguistici, statistici e storici per l'analisi regionale,
Sapienza Università di Roma, Italy
e-mail: brunero.liseo,andrea.tancredi@uniroma1.it

Abstract Record linkage is a collection of techniques which aim at identifying data records on two different electronic files that contain information about the same "entity." There exists mainly two reasons to perform record linkage: data collation and list construction and both of them are among the most crucial task of National Institutes of Statistics, and other National and International organism and private users as well.

The use of record linkage techniques poses several interesting problems both from the methodological and the computational viewpoint. From the methodological perspective, the definition itself of a statistical model (to describe the way in which comparisons among records should be performed) is still debated Fellegi and Sunter (1969), Copas and Hilton 1991; Belin and Rubin 1995; Fortini et al. (2001). From the computational perspective, problems become formidable as soon as the sizes of the databases are large (more than 100 units); one of most popular solution is to perform comparisons only between those records which show the same values on some "blocking variables" which are assumed to be recorded without errors: the resolution, at least partial, of this problem seems crucial. In this paper we propose a Bayesian perspective for the construction of a record linkage statistical model. While it is definitely true that the result of a statistical analysis produced by an official organism "must" be objective (or - at least - it should be perceived as such by the users), it is also undeniable that Bayesian ideas and techniques can play an important role in official statistics (i) when important prior (or extra-experimental) information about the variables of interest exist and cannot be adequately exploited in a classical inference framework; (ii) even when prior information is lacking, a Bayesian analysis can be necessary simply because a classical approach cannot provide answers without introducing strong assumptions, not easily testable. In these situations a Bayesian analysis allows, at least, to perform a sensitivity analysis, with the aim of quantifying the influence of the assumptions on inferences.

Keywords: False Match Rate, Fellegi-Sunter method, Latent structure models, Linked data, MCMC algorithm.

1. Introduction

Record linkage refers to the use of an algorithmic technique to match records from different data sets that correspond to the same statistical unit, but lack unique personal identification code. The need of record linkage techniques is steadily increasing in various chapters of statistics. For example, in official statistics record linkage is a necessary preliminary step when the size of a population is estimated via capture-recapture techniques, especially when the target population is elusive (the estimation of the number of non regular immigrants in European Community is an example) and differences in identification variables in the two occasions are the rule rather than the exception. Another example, which is particularly important for Statistical Institutes, is the use of administrative data bases in order to integrate information obtained from a survey, relieving response burden. From a broader perspective, many Statistical Institutes and agencies use the methodology of file merging to create comprehensive files from multiple but incomplete data sources. The main scope of this endeavor is to perform statistical analyses on the synthetic data set, generated by file merging, which could not be performed by analyzing the incomplete data sets separately. In theory the validity and the efficacy of the file merging methodology could be assessed by means of statistical models which represents the mechanisms which generate the incomplete data sets. However there is no yet a complete and satisfactory theory of record linkage procedures.

In general, from a statistical methodology perspective, the merge of two (or more) data files can be important for two reasons

- *per sé*, to obtain a larger and integrated reference data set.
- to perform a subsequent statistical analysis based on the additional information which cannot be extracted from either of the two single data files.

Here we give a toy example of the latter: suppose we have two computer files $\mathcal{A}$ and $\mathcal{B}$ whose records relate respectively to units (e.g. individuals, firms) of partially overlapping populations $\mathcal{P}_{\mathcal{A}}$ and $\mathcal{P}_{\mathcal{B}}$. The two files consist of several fields, or variables, either quantitative or qualitative. For example, in a file of individuals, fields can be "surname", "age", "sex", etc. The goal of a record linkage procedure is to detect all the pairs of units (a, b), $a \in \mathcal{A}$ and $b \in \mathcal{B}$, such that a and b refer actually to the same unit. Suppose that the observed variables in $\mathcal{A}$ are denoted by

$$(Z, W_1, W_2, \cdots, W_k)$$

while we observe

$$(W_1, W_2, \cdots, W_k, X)$$

in file $\mathcal{B}$. Then we might be interested in studying a linear regression analysis (or any other more complex association model) between Z and X, restricted to those pairs of record which we declare as matches. The intrinsic difficulties which are present in such a simple problem are well documented and discussed in Scheuren and Winkler (1993) and Lahiri and Larsen (2005).

In the statistical practice it is quite common that the *linker* (the researcher who matches the two files) and the *analyst* (the statistician doing the subsequent analysis) are two different persons working separately. However, we agree with Scheuren and Winkler (1993), which say

> "*...it is important to conceptualize the linkage and analysis steps as part of a single statistical system and to devise appropriate strategies accordingly.*"

In a more general framework, suppose that file A contains the variables $(\mathbf{Z}, \mathbf{W}_A) = (Z_1, Z_2, \cdots Z_h, W_1, W_2, \cdots, W_k)$ observed on ν_A units, while X_B contains the variables $(\mathbf{W}_B, \mathbf{X}) = (W_1, W_2, \cdots, W_k, X_1, X_2, X_p)$. Our goal can be stated as follows:

1- to use the key variables $(W_1, W_2, \cdots, W_k)$ to detect the true links between X_A and X_B.

2- to perform a statistical analysis based on vectors of variables $\mathbf{Z}$ and $\mathbf{X}$ restricted to those records which have been defined matches.

To perform this task, we present a fully Bayesian analysis, which is particularly suitable to accomplish the above desideratum. The main point is that in our approach all the uncertainty about the matching process is automatically retained in the subsequent inferential steps. This paper generalizes and improve the Bayesian model for record linkage described in Fortini et al. (2001).

We present the general theory underlying the model and illustrate its performance via several examples related to various statistical analysis. The paper is organized as follows: Section 2 briefly recall the Bayesian approach to record linkage proposed by Fortini et al. (2001) and provides some computational improvements on it. Section 3 generalizes the method to include the inferential part. Section 4 concentrates on the special case of regression analysis, the only situation which has been already considered in literature: see Scheuren and Winkler (1993) and Lahiri and Larsen (2005).

2. Bayesian Record Linkage

2.1. The usual statistical model for record linkage

We first examine the classical approach to the record linkage problem. Consider two data files $\mathcal{A}$ and $\mathcal{B}$, with respectively ν_A and ν_B units. Let us call A and B the two sets (lists) of observed units, $a = 1, \cdots, \nu_A$, $b = 1, \cdots, \nu_B$. We assume that at least some units are present in both lists. The set of all ordered pairs

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

can be logically split into two non-overlapping sets, namely

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

the set of matches, and

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

the set of non-matches. In order to decide whether a specific pair (a, b) is actually a member of $\mathcal{M}$ or $\mathcal{U}$, we may compare variables observed in both the files (e.g. surname, name, sex, address, etc. for individuals): these variable are called *key* variables. Let us assume we have k key variables, $k \geq 1$, whose realizations in the two data lists are denoted by:

$$w_a = (w_{a,1},\, w_{a,2},\, ...,\, w_{a,k}), \qquad a \in A,$$

and

$$w_b = (w_{b,1},\, w_{b,2},\, ...,\, w_{b,k}), \qquad b \in B.$$

We denote by $Y_{ab}^{(j)}$, $j = 1, \ldots, k$, the result of the comparison among the values $w_{a,j}$ and $w_{b,j}$. The comparison $Y_{ab}^{(j)}$ may be, in general, any function of $w_{a,j}$ and $w_{b,j}$. The most commonly assumed comparison function takes the form of a vector of k elements, $Y_{ab} = (y_{ab}^{(1)}, ..., y_{ab}^{(k)})$ with:

$$y_{ab}^{(j)} = \begin{cases} 1 & \text{if } w_{a,j} = w_{b,j} \\ 0 & \text{otherwise.} \end{cases} \qquad j = 1, \cdots, k. \tag{1}$$

More general and sensible comparison functions can be used, especially in the case of continuous key variables. However, the $0/1$ comparisons are compatible with a reasonably fast and accurate matching process. A simple and not too expensive generalization of this dichotomy may be used by discretizing the observed values in a small number of classes. We will discuss more deeply this issue in the final section.

A more radically different approach would be based on the actual observations taken in the $\mathcal{A}$ and $\mathcal{B}$ files, rather than considering the comparisons. Copas and Hilton (1990) deal with this problem. We sketch a possible extension of their ideas in the final section.

In the $0/1$ case, the comparison vector y_{ab} can assume 2^k different values which we will indicate with $y_{\mathbf{i}}$ where $\mathbf{i} = 1, \ldots, 2^k$. In order to decide whether a pair (a, b) with comparison vector y_{ab} should be linked or not, Fellegi and Sunter (1969) suggest to consider the sampling distribution of the comparison vectors in $\mathcal{M}$, say $m(y)$, and the corresponding distribution in $\mathcal{U}$, $u(y)$. The decision rule for the pair (a, b) is based on the likelihood ratio

$$t(y_{ab}) = \frac{m(y_{ab})}{u(y_{ab})}. \tag{2}$$

Fellegi and Sunter (1969) discuss several frequentist optimality properties of such decision rule. Given that neither $m(y)$ nor $u(y)$ are known, most of the literature on record linkage concentrates on how to estimate them. Starting with Jaro (1989), a model based approach has been advocated for this task. The usual assumption is that the status of a pair (let's say C_{ab}, where $C_{ab} = 1$ when a pair (a, b) is a true match and 0 otherwise) is a non observable random variable, while the comparison vector Y represents the actual data. Also, a general latent structure is assumed via the configuration matrix $C = \{C_{ab}, a \in A, b \in B\}$, so that

1. C_{ab}, $(a, b) \in A \times B$, are assumed to be i.i.d. Bernoulli r.v. such that for all a, b, $P(C_{ab} = 1) = p$;
2. the comparison vectors Y_{ab}, $(a, b) \in A \times B$, are assumed to be i.i.d. replications of the r.v. Y whose marginal (with respect to **C**) distribution has the following mixture structure
$$\Pr(Y = y|p) = p\, m(y) + (1 - p)\, u(y);$$
3. for fixed p, the random vector (C_{ab}, Y_{ab}), $(a, b) \in A \times B$, are independent and identically distributed with distribution, for $c = 0, 1$,
$$\Pr(C = c,\, Y = y) = \left[p\, m(y)\right]^c \left[(1 - p)\, u(y)\right]^{1-c},$$

The independence assumption are quite unrealistic because if $c_{ab} = 1$ then all the other elements on the row a and on the column b must be 0. Notwithstanding, independence makes particularly easy the computation of the likelihood function given the $n_A \times n_B$ observations (c_{ab}, y_{ab}):

$$\prod_{(a,b) \in A \times B} \left(p\, m(y_{ab})\right)^{c_{ab}} \left((1 - p)\, u(y_{ab})\right)^{1-c_{ab}}. \tag{3}$$

Maximum likelihood estimates of the distributions $m(y)$ and $u(y)$ may consequently be obtained, using for instance the EM algorithm, where the matrix C plays the role of *missing data*. Jaro (1989) assumes that the components of the comparison vector Y are mutually independent, whereas Winkler (1993) and Larsen and Rubin (2001), among the others, consider the case of dependent key variables comparisons.

2.2. The Bayesian model

The Bayesian model should be expresses in terms of a prior distribution on the unknown parameters and in terms of the conditional distribution of the observed data given the unknown parameters. The observed data are lexicographically ordered in the vector $\mathbf{y} = (y_{11}, \ldots, y_{\nu_a \nu_b})$ while the parameters are represented by the configuration matrix $\mathbf{C}$, the vector $\mathbf{m} = (m_1, \ldots, m_{2^k})$ where $m_{\mathbf{i}} = P(Y_{ab} = y_{\mathbf{i}} | c_{ab} = 1)$ and the vector $\mathbf{u} = (u_1, \ldots, u_{2^k})$, where $u_{\mathbf{i}} = P(Y_{ab} = y_{\mathbf{i}} | c_{ab} = 0)$.

The conditional distribution of the $\mathbf{y}$ given the parameters $\mathbf{C}, \mathbf{m}, \mathbf{u}$ can be written as

$$\begin{aligned} f(\mathbf{y}|\mathbf{C}, \mathbf{m}, \mathbf{u}) &= \prod_{a=1}^{\nu_A} \prod_{b=1}^{\nu_B} f(y_{ab}|\mathbf{C}, \mathbf{m}, \mathbf{u}) \\ &= \prod_{a=1}^{\nu_A} \prod_{b=1}^{\nu_B} f(y_{ab}|c_{ab}, \mathbf{m}, \mathbf{u}) \qquad (4) \\ &= \prod_{a=1}^{\nu_A} \prod_{b=1}^{\nu_B} \left[\prod_{\mathbf{i}=1}^{2^k} m_{\mathbf{i}}^{d(y_{ab}, y_{\mathbf{i}})} \right]^{c_{ab}} \left[\prod_{\mathbf{i}=1}^{2^k} u_{\mathbf{i}}^{d(y_{ab}, y_{\mathbf{i}})} \right]^{1-c_{ab}} \end{aligned}$$

where

$$d(y_{ab}, y_{\mathbf{i}}) = \begin{cases} 1 & \text{if } y_{ab} = y_{\mathbf{i}} \\ 0 & \text{otherwise} \end{cases} .$$

In what follows, we will assume that $\mathbf{m}$ and $\mathbf{u}$ are a priori independent of $\mathbf{C}$. In absence of specific prior information on the vectors $\mathbf{m}$ and $\mathbf{u}$ it is reasonable to adopt a conjugate Dirichlet prior distribution both for $\mathbf{m}$ and for $\mathbf{u}$. In particular,

$$\mathbf{m} \sim \mathcal{D}(\alpha_1, \ldots, \alpha_{2^k}); \qquad \mathbf{u} \sim \mathcal{D}(\beta_1, \ldots, \beta_{2^k}).$$

Also, we need to introduce a hyper-structure over the vector $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in order to *unsaturate* the model. Following Fortini et al. (2001), we set

$$\log \alpha_{\mathbf{i}} = (\sum_{\mathbf{i}=1}^{k} y_{\mathbf{i}}^{k} - \phi) \log \theta, \qquad \log \beta_{\mathbf{i}} = (\phi - \sum_{\mathbf{i}=1}^{k} y_{\mathbf{i}}^{k}) \log \theta. \qquad (5)$$

The introduction of the hyperparameters θ and ϕ greatly simplifies the model and renders it non saturated. The rationale behind this reparameterization is that it is able to models our beliefs on the informative power of each comparison variable. In fact, the hyperparameters in the equation of (5) hierarchically order the possible values of the comparison vectors in such a way that, for example, the prior distribution for $\mathbf{m}$ puts more mass around “large values” of the $m_{\mathbf{i}}$’s for those i’s ($\mathbf{i} = 1, 2, \cdots, 2^k$) with a large number of 1’s in the observed comparison vector. The opposite argument holds for the hyperparameters in the second equation of (5). In particular Fortini et al. (2001) show that, by introducing the hyperparameters ϕ and θ, the marginal prior means of the $m_{\mathbf{i}}$'s and the $u_{\mathbf{i}}$'s are simple functions of θ only, whereas their variances depend on both θ and ϕ. The hyperparameters (5) have also direct effects on the statistical relationship among the comparison variables. For instance the linear correlation between two comparison variables has a null expected value for any θ and ψ, whereas their variance depends on both. These considerations can guide the elicitation process for the hyperparameters. It must be said that the the above elicitation of the prior structure makes a slight use of the data information and, only in this sense, our approach can be viewed as an empirical Bayes one.

To complete the model we need to elicit a prior distribution for the configuration matrix $\mathbf{C}$. We assume that each single record in $\mathcal{A}$ can match at most one record in $\mathcal{B}$; then $\mathbf{C}$ must satisfy the natural constraints

$$c_{ab} \in \{0,1\}\,(a=1,\cdots,A; b=1,\cdots B), \qquad \sum_{a=1}^{\nu_A} c_{ab} \leq 1, \quad \sum_{b=1}^{\nu_B} c_{ab} \leq 1. \tag{6}$$

Let $T = \sum_{ab} c_{ab}$ denote the number of true matches in $\mathbf{C}$; also, let $T_m = \min\{\nu_A, \nu_B\}$ be the maximum possible number of matches. We also denote by T_q = the quantity $\max\{\nu_A, \nu_B\}$. The prior distribution on $\mathbf{C}$ can be built up in two stages. First, we assume that T, the number of matches, follows a binomial distribution with parameters ξ and T_m, that is

$$P(T=t) = \binom{T_m}{t} \xi^t (1-\xi)^{T_m - t}, \qquad t = 0, 1, T_m.$$

At the second stage we assume that, conditional on $T = t$, the distribution over the space of all possible matrices $\mathbf{C}$ - satisfying the constraints (6) - is uniform. Then,

$$P(C|T=t) = \begin{cases} \left(\binom{T_m}{t}\binom{T_q}{t}t!\right)^{-1} & \text{if } \sum_{ab} C_{ab} = t \\ 0 & \text{otherwise} \end{cases}$$

Notice that the hyperparameter ξ has a precise interpretation, since it represents the probability that a generic unit in the smaller dataset will be present in the larger dataset also. We may consider ξ either known or unknown. In the latter case we will assume that ξ follows a Beta(δ_1, δ_2) distribution. It can also be proved that (see Appendix 5) that $E(C_{ab}) = p$, where $p = \xi / T_q$. Then the quantity p might be interpreted as the probability that a randomly chosen pair (a, b) is actually a match. Thenm our prior assumptions lead to the following posterior distribution for the parameters $(\mathbf{C}, \mathbf{m}, \mathbf{u}, \xi)$ (recall that $\mathbf{m} = \mathbf{m}(\phi, \theta)$ and $\mathbf{u} = \mathbf{u}(\phi, \theta)$)

$$\begin{aligned} p(\mathbf{C}, m, u, \xi | y) \quad \propto \quad & \prod_{a=1}^{\nu_A} \prod_{b=1}^{\nu_B} \left[\prod_{\mathbf{i}=1}^{2^k} m_{\mathbf{i}}^{d(y_{ab}, y_{\mathbf{i}})} \right]^{c_{ab}} \left[\prod_{\mathbf{i}=1}^{2^k} u_{\mathbf{i}}^{d(y_{ab}, y_{\mathbf{i}})} \right]^{1-c_{ab}} \\ \times \quad & \frac{\xi^{\sum c_{ab} + \delta_1 - 1} (1-\xi)^{T_m + \delta_2 - 1 - \sum c_{ab}}}{\binom{T_q}{\sum c_{ab}} \sum c_{ab}!} \prod_{\mathbf{i}=1}^{2^k} m_{\mathbf{i}}^{\alpha_{\mathbf{i}} - 1} u_{\mathbf{i}}^{\beta_{\mathbf{i}} - 1} \end{aligned}$$

2.3. MCMC implementation

The Bayesian model proposed in this paper is too complex to be amenable to analytical calculations. Hence, we turn to Monte Carlo Markov Chain methods, and in particular to a Gibbs sample algorithm. In fact, it is easy to show that

- the full conditional posterior distributions of the vector $\mathbf{m}$ and $\mathbf{u}$ are still Dirichlet distributed, while the full conditional of ξ is still a Beta distribution;
- each single entry of the matrix $\mathbf{C}$ has a full conditional distribution (given also the other entries of the matrix) which is either Bernoulli or degenerate.

To update each single element of the matrix $\mathbf{C}$ we need first to calculate the conditional prior probability that a couple (a, b) is a match given all the other elements of the matrix c. We will indicate with the symbol $\mathbf{C}^{-ab}$ the matrix $\mathbf{c}$ without the element c_{ab}. Of course we have that $\Pr\left(c_{ab} = 1|\mathbf{C}^{-ab}\right) = 0$ if a match is present in the row a or in the column b, i.e. if $\sum_{b' \neq b} c_{ab'} = 1$ or $\sum_{a' \neq a} c_{a'b} = 1$.
Let t^{-ab} be the number of matches of the matrix $\mathbf{C}^{-ab}$; when $t^{(-ab)} = t - 1$ and $\sum_{a' \neq a} c_{a'b} = 0$, $\sum_{b' \neq b} c_{ab'} = 0$ it can be shown that (Appendix 5)

$$P(c_{ab} = 1|c^{-ab}) = \left[1 + \frac{1 - pT_q}{pT_q}(T_q - t + 1)\right]^{-1}$$

The above formula allows to easily calculate the full conditional posterior distribution for each single $\{c_{ab}\}$. In fact

$$c^{ab}| \cdots \sim \text{Bernoulli}(P(c_{ab} = 1|y, c^{-ab}, m, u))$$

where $P(c_{ab} = 1|y, \mathbf{C}^{-ab}, m, u)$ can be written as

$$= \frac{P(y_{ab}|c_{ab} = 1)P(c_{ab} = 1|\mathbf{C}^{-ab})}{P(y_{ab}|c_{ab} = 1)P(c_{ab} = 1|\mathbf{C}^{-ab}) + P(y_{ab}|c_{ab} = 0)P(c_{ab} = 0|\mathbf{C}^{-ab})}$$
$$= \frac{\prod_{\mathbf{i}=1}^{2^k} m_{\mathbf{i}}^{d(y_{ab},y_{\mathbf{i}})} P(c_{ab} = 1|\mathbf{C}^{-ab})}{\prod_{\mathbf{i}=1}^{2^k} m_{\mathbf{i}}^{d(y_{ab},y_{\mathbf{i}})} P(c_{ab} = 1|\mathbf{C}^{-ab}) + \prod_{\mathbf{i}=1}^{2^k} u_{\mathbf{i}}^{d(y_{ab},y_{\mathbf{i}})} P(c_{ab} = 0|\mathbf{C}^{-ab})}.$$

The full conditional distributions of the other parameters can be obtained via lengthy but simple calculations: For $\mathbf{m}$ and $\mathbf{u}$ one has

$$m| \ldots \sim \mathcal{D}(\alpha_1 + \sum_{ab} d(y_{ab}, y_1)c_{ab}, \ldots, \alpha_{2^k} + \sum_{ab} d(y_{ab}, y_{2^k})c_{ab})$$

and

$$u| \ldots, \sim \mathcal{D}(\beta_1 + \sum_{ab} d(y_{ab}, y_1)(1 - c_{ab}), \ldots, \beta_{2^k} + \sum_{ab} d(y_{ab}, y_{2^k})(1 - c_{ab}))$$

whereas the hyperparameter ξ has a beta conditional distribution

$$\xi| \cdots \sim B(\delta_1 + t, \delta_2 + T_m - t).$$

For all these variables we can use a Gibbs sampling step.

3. A general method for modelling dependence

In this section we present the most innovative proposal of the paper: the construction and the calibration of a statistical model based on a data set which is the output of a record linkage procedure. As a simulation based Bayesian procedure, the final output provided by the procedure described in (2) will be a simulation from the joint posterior distribution of the parameters $(\mathbf{C}, \mathbf{m}, \mathbf{u}, \xi)$

This can be used according to two different strategies. In fact we can either

- compute a "point" estimate of the matrix $\mathbf{C}$ and then plug-in this estimate to establish which pairs are passed to the second stage of the statistical analysis. It must be noticed that, given the particular structure of the parameter matrix $\mathbf{C}$, no easy point estimates are available. The posterior mean of C is in fact useless since we need to estimate each single c_{ab} with 0 or 1! The posterior median is difficult to define as well, and the most natural candidate, the maximum a posteriori (MAP) estimate typically suffers from sensitivity (to the prior and to the Monte Carlo variability) problems: this last issue is particularly crucial in official statistics. For a deep discussion on these issues see Tancredi et al. (2005) and, for related problems in a different scenario, Green and Mardia (2006),

or

- transfer the "global" uncertainty relative to $\mathbf{C}$ (and to the other parameters), given by their posterior distribution, into the second step statistical analysis.

We argue that the latter approach is more coherent and, among other things, it avoids an over-estimate of the precision measures attached to the output of the second step analysis. However it is also possible to improve on the latter approach, by performing the second step analysis at the same time as the record linkage analysis, that is, including the second step analysis into the MCMC procedure. This will cause a feed-back propagation of the information between the record linkage parameters and the more specific quantities object of interest. Here we illustrate these ideas in a very general setting; in the next section we will consider the regression example in details.

Let $\mathcal{D} = [\mathbf{y}, \mathbf{z}, \mathbf{x}] = (\mathbf{y}_{11} \ldots, \mathbf{y}_{\nu_A \nu_B}, \mathbf{z}_1, \ldots, \mathbf{z}_{\nu_A}, \mathbf{x}_1, \ldots, \mathbf{x}_{\nu_B})$ the entire set of available data where, as in the Introduction, $\mathbf{y}_{ab}$ represents the vector of comparisons among variables which are present in both files, while $\mathbf{z}_a$ is the value of covariate Z observed on individual $a \in \mathcal{A}$ and $\mathbf{x}_b$ is the value of covariate X observed on individual $b \in \mathcal{B}$. The statistical model can then be writte as

$$p(\mathbf{y}, \mathbf{z}, \mathbf{x} | \mathbf{C}, \mathbf{m}, \mathbf{u}, \boldsymbol{\theta}), \tag{7}$$

where $(\mathbf{C}, \mathbf{m}, \mathbf{u})$ are the record linkage parameters while $\boldsymbol{\theta}$ is the parameter vector related to the joint distribution of (X, Z). La (7) can always be re-expressed as

$$p(\mathbf{y} | \mathbf{C}, \mathbf{m}, \mathbf{u}, \boldsymbol{\theta}) p(\mathbf{x}, \mathbf{z} | \mathbf{C}, \mathbf{y}, \mathbf{m}, \mathbf{u}, \boldsymbol{\theta}).$$

Also, it sounds reasonable to assime that, given $\mathbf{C}$, the vector of comparisons $\mathbf{Y}$ does not depend on $\boldsymbol{\theta}$; moreover, for given $\mathbf{C}$, the distribution of $(\mathbf{X}, \mathbf{Z})$ does not depend both on the comparison vector data $\mathbf{Y}$ and the parameters related to those comparisons. It follows that (7) can be simplified into the following general expression.

$$p(\mathbf{y}|\mathbf{C},\mathbf{m},\mathbf{u})p(\mathbf{x},\mathbf{z}|\mathbf{C},\boldsymbol{\theta}). \tag{8}$$

The first term in (8) is related to the record linkage analysis; the last term refers to the second step analyis and must be specified according to that. The presence of $\mathbf{C}$ in both the terms allows the feed-back phenomenon we have mentioned before.

4. Regression

Lahiri and Larsen (2005) consider the following scenario. Let assume that the two datasets consist of the same $n = \nu_A = \nu_B$ units. Let $\mathbf{Z}$ be a univariate response variable which is available on units in database $\mathcal{A}$ and let $\mathbf{X} = (X_1, \cdots, Xp)$ be the vector of covariates, available for units in database $\mathcal{B}$. Let us also define a matrix $\mathbf{P}$ where the generic element p_{ab} denotes the probability that the a-th unit of database $\mathcal{A}$ matches the b-th unit of database $\mathcal{B}$. Suppose we want to perform a linear regression for $\mathbf{Z}$ and $\mathbf{X}$, that is

$$\mathbf{Z} = \mathbf{X}\boldsymbol{\beta} + \varepsilon,$$

where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)$, under the usual regression assumptions. Since the information about the true links is missing, it is possible to restate the model in the following way;

$$Z_i = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i, \qquad i = 1, \cdots, n.$$

and introduce the new variables V_i, $i = 1, \cdots, n$,

$$V_i = \begin{cases} w_i & \text{c.p. } p_{ii} \\ w_j & \text{c.p. } p_{ij}, \quad j \neq i, j = 1, \cdots \nu_A \end{cases}.$$

Using our latent variable notation, Lahiri and Larsen's (2005) approach is equivalent to the introduction, for each unit in $\mathcal{A}$, a latent vector

$$\mathbf{S}_a = (S_{a,1}, \cdots, S_{a,n})$$

which consists of just *one* 1 and $n-1$ zeros. Also, $\mathbf{S}_1, \cdots, \mathbf{S}_n$ are assumed to be independent with $\mathbf{S}_j \sim$ Multinomial$(1, \mathbf{p}_j)$, with $\mathbf{p}_j = (p_{j,1}, \ldots, p_{j,n})$. Then, it is easy to show that $\mathbb{E}\left(\mathrm{W_j} \mid \mathrm{s}_1, \cdots \mathrm{s_n}\right) = \sum_{\mathrm{b=1}}^{\mathrm{n}} \mathrm{s_{j,b}}\mathbf{x}_{\mathrm{b}}'\boldsymbol{\beta}$, and

$$\mathbb{E}\left(\mathrm{W_j}\right) = \mathbb{E}\left[\mathbb{E}\left(\mathrm{W_j} \mid \mathbf{S}_1, \cdots \mathbf{S}_\mathrm{n}\right)\right] = \mathbb{E}\left(\sum_{\mathrm{b=1}}^{\mathrm{n}} \mathrm{S_{j,b}}\mathbf{z}_{\mathrm{b}}'\boldsymbol{\beta}\right) = \sum_{\mathrm{b=1}}^{\mathrm{n}} \mathrm{p_{j,b}}\mathbf{z}_{\mathrm{b}}'\boldsymbol{\beta},$$

that is $\mathbb{E}\left(\mathbf{W}\right) = \mathbf{PZ}\boldsymbol{\beta}$. The resulting *unbiased* estimator of $\boldsymbol{\beta}$ turns out to be

$$\hat{\boldsymbol{\beta}}_{LL} = (Z\mathbf{P}'\mathbf{PZ})^{-1}\mathbf{Z}'\mathbf{Pw}$$

In words, in order to account for the uncertainty related to the matching process, Lahiri and Larsen (2005) use a *weighted combination of covariates*, where the weights are *estimated* from the linkage model step. They also provide an estimate for the variance of

$\hat{\boldsymbol{\beta}}_{LL}$ via a parametric bootstrap approximation, in order to produce confidence interval for the components of vector $\boldsymbol{\beta}$. However this confidence intervals tends to be too optimistic since the uncertainty about probability of matching is accounted for only partially.

To illustrate our Bayesian approach to inference with linked data we consider a simple application based on two small real data sets having some *known* common units. Data are taken from the Italian Survey on Household Income and Wealth (SHIW). For the sake of brevity, we did not include data in the paper: they are available at the website `http://3w.eco.uniroma1.it/utenti/tancredi/datalink.txt` The survey is reapeted every two year; we consider, respectively, the 2000 and 1998 surveys. We have restricted our analysis to a small subsample, namely the data related to a single northern Italian region (Valle d'Aosta). The number of households interviewed in such a region in year 2000 was 25 (22 in 1998) and 13 of these (panel households) had been interviewed also in 1998 surveys. The panel households have the same questionnaire number across the surveys. For every household we have reported the following characteristics of the householder: year of birth, branch of activity, gender, marital status, level of education. Note that the householder is not suppose to change across the surveys. Moreover, we report, for the 2000 survey, the annual household consumption and the household annual net disposable income for the 1998 file; both the variables are expressed in thousands of liras, the previous Italian currency. Our goal is to study the possibly linear relationship between the 2000 consumption and the 1998 income, pretending to ignore which household match. Considering the common variables recorded on the householders as key variables, we have implemented the Bayesian approach described is section (2). We have assumed a Binomial distribution with parameters $\xi = 13/22$ and $T_m = 22$ as the prior for the number of matches; also, conditional on the value $T = t$, we assumed a uniform distribution on the space of all the possible matrices $\mathbf{C}$ with t matches. This way we try to be rather informative on the number of matches (an information which is often available in practice), but totally uninformative on *which pairs* are actually matches. As far as the hyperparameters in (5), we set $\phi = 0.5$ and $\theta = 2.0$; the resulting prior expected value for the probabilities $\mathbf{m}$ and $\mathbf{u}$ are summarized in the following table

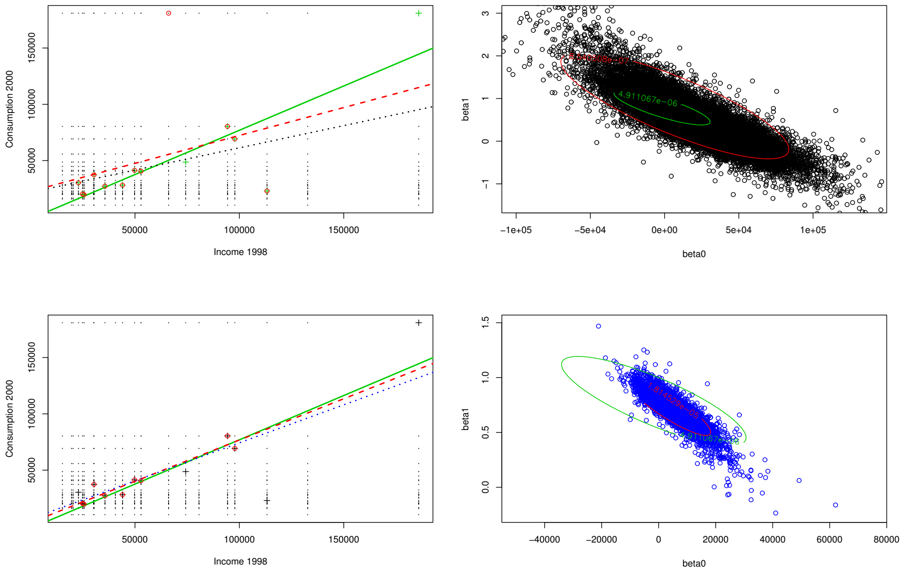
$\sum_{i=1}^{k} y_i$	0	1	2	3	4	5
$E(m_i)$	0.004	0.008	0.016	0.033	0.066	0.132
$E(u_i)$	0.132	0.066	0.033	0.016	0.008	0.004

Notice that, assuming that a couple is really a match, the average prior probability to observe a comparison vector with k components equal to 1 increases with k. On the other hand, given that a couple is not a match, the average prior probability to observe a comparison vector with k components equal to 1 decreases with k. We graphically report the results obtained with 50000 iterations of the algorithm described in section 2. Using a quadratic loss function (Tancredi et al., (2005) we have estimated C_{ab} equal to 0 if $\Pr(C_{ab} = 1 \mid \text{ data }) < 1/2$ and equal to 1 otherwise. This way we have obtained 11 matches. Ten of these couples are true matches, so we have 1 false match and 2 false negative matches.

Upper left corner of Figure 4 shows (with crosses) the 13 true matches, while the 12 estimated matches are represented with red circles.

The green line is the "true" regression line (obtained from the *crosses* points. The dashed red line is the regression line obtained from the red circles points; the dotted line is obtained using the means of the MCMC output for (β_0, β_1). Upper right corner

Figure 1: *Bayes estimates for the SHIW data*



of Figure 4 reports the posterior simulation for The inner ellipse is the 95% HPD region obtained with the 13 true matches, the largest ellipse is the 95% HPD region obtained with the estimated 12 matches. The presence of a false positive match dramatically increases the Bayesian standard error of the estimates.

In the bottom parts of Figure 4 we shows the results obtained with the new integrated method, where the *regression* part of the analysis is included into the Gibbs sampling. Estimates are dramatically better both in terms of bias and accuracy.

5. Conclusions

We have described a model based approach to record linkage analysis. We have showed through a simple example - but we think that the issue is quite general - that, when a record linkage output must be used for a subsequent statistical analysis, is by far better to merge the two parts of the analysis into a general statistical model.

However, it may happen that the same record linkage output is to be used for several different statistical analyses. In such cases it is probably better to perform record linkage first to produce an all-purposes list.

The model we have described here is somehow classical, in the sense that is is based on the vector of comparisons $\mathbf{Y}$ of the observed values of the key variables $\mathbf{X}$ on the two databases. One line of research, initiated by Copas and Hilton (1990) try to model directly the information provided by the $\mathbf{X}$'s.

These ideas can be captured into a hierarchical Bayesian model where, for each sample, the key variables are modelled conditionally on the latent true values. This way, it is possible to consider categorical key variables and, in order to take into account measurement errors, the hit-and-miss model (Copas and Hilton, 1990) can be used. Conditionally on the size N of the data generating population, standard Bayesian models for inference with categorical data drawn from finite population can then be used to model the unobserved true values. Finally, a non informative prior distribution can assumed for N. A specific version of these ideas can be implemented in such a way that a straight Gibbs sampler algorithm can be used to simulate from the posterior of all the model parameters. We stress the fact that, in this framework, the record linkage process can be performed on the base of the true unobserved values at each iteration of the MCMC algorithm. Also, treating the population size N as unknown, the uncertainty implicit in the record linkage process is automatically taken into account. We will explore these issues in details elsewhere.

Appendix

1. First, we prove that $E(C_{ab}) = p$, where $p = \xi/T_q$. In fact

$$\begin{aligned} P(C_{ab}=1|T=t) &= \sum_{c\in\mathcal{C}} P(C_{ab}=1, C=c|T=t) \\ &= \sum_{c\in\mathcal{C}:t(c)=t,c_{ab}=1} \frac{1}{\binom{T_m}{t}\binom{T_q}{t}t!} == \frac{\binom{T_m-1}{t-1}\binom{T_q-1}{t-1}(t-1)!}{\binom{T_m}{t}\binom{T_q}{t}t!}. \end{aligned}$$

Then

$$\begin{aligned} E(\mathbf{C}_{ab}) &= P(\mathbf{C}_{ab}=1) = \sum_{t=0}^{T_m} P(C_{ab}=1, T=t) \\ &= \sum_{t=1}^{T_m} P(C_{ab}=1|T=t)P(T=t) = \sum_{t=1}^{T_m} \frac{\binom{T_m-1}{t-1}\binom{T_q-1}{t-1}(t-1)!}{\binom{T_m}{t}\binom{T_q}{t}t!}P(T=t) \\ &= \sum_{t=1}^{T_m} \frac{t}{T_mT_q}P(T=t) = \frac{E(T)}{T_mT_q} = \frac{T_m\xi}{T_mT_q} = \frac{\xi}{T_q} = p. \end{aligned}$$

2. Conditional distribution of $C_{ab} = 1|\mathbf{C}^{(-ab)})$

$$\begin{aligned} P(C_{ab}=1|C^{(-ab)}) &= \left[1 + \frac{P(C_{ab}=0, C^{(-ab)})}{P(C_{ab}=1, C^{(-ab)})}\right]^{-1} \\ &= \left[1 + \frac{P(T=t-1)P(C_{ab}=0, C^{(-ab)}|T=t-1)}{P(T=t)P(C_{ab}=1, C^{(-ab)}|T=t)}\right]^{-1} \\ &= \left[1 + \frac{P(T=t-1)\frac{1}{\binom{T_m}{t-1}\binom{T_q}{t-1}(t-1)!}}{P(T=t)\frac{1}{\binom{T_m}{t}\binom{T_q}{t}t!}}\right]^{-1} \\ &= \left[1 + \frac{1-pT_q}{pT_q}(T_q-t+1)\right]^{-1} \end{aligned}$$

References

Belin, T.R. and Rubin, D.B. (1995) A method for calibrating false - match rates in record linkage Journal of the Amer. Stat. Assoc., vol. 90, 694–707.

Copas J.B., Hilton F. J. (1990) Record linkage: statistical models for matching computer records. Journal of the Royal Statistical Society Series A, 287–320.

Fortini M., Liseo B., Nuccitelli A., Scanu M. (2001) On Bayesian Record Linkage, Research in Official Statistics, 4, 185–198.

Green, P.J., Mardia, K. (2006) Bayesian alignment using hierarchical models, with applications in protein bioinformatics Biometrika, vol. 93, 235–254.

Jaro. M. (1989) Advance in record linkage methodology as applied to matching the 1985 census of Tampa, Florida. Journal of the American Statistical Association, 84, 414–420.

Lahiri P., Larsen M. (2005) Regression analysis with linked data, Journal of the American Statistical Association, 100, 222–230.

Scheuren F., Winkler W. (1993) Regression analysis of data files that are computer matched Survey Methodology, 19, 39–58.

Tancredi, A. Guagnano, G. Liseo, B. (2005) Inferenza statistica basata su dati prodotti mediante procedure di record linkage in: *L'Integrazione di dati di fonti diverse: tecniche e applicazioni del Record Linkage e metodi di stima basati sull'uso congiunto di fonti statistiche e amministrative*, Falorsi, P., Pallara, A. & Russo, A. (eds.), Franco Angeli, Roma, 41–59.

Theory and practice in developing a record linkage software

Nicoletta Cibella, Marco Fortini, Monica Scannapieco,
Laura Tosco, Tiziana Tuoto
Istituto Nazionale di Statistica (ISTAT), via Cesare Balbo 16, 00184 Roma, Italy,
e-mail: {firstname.surname}@istat.it

Abstract: The combined use of statistical survey and administrative data is largely widespread to maximize their respective usefulness: unfortunately data sources are often hard to integrate due to errors or lacking information. Record linkage techniques are a multidisciplinary set of methods and practices aiming to identify the same real world entity, differently represented in data sources. Record linkage is a complex process but it can be decomposed in separate phases, each of them requiring a specific technique. To deal with such a complexity, we propose RELAIS (REcord Linkage At IStat), an open source toolkit based on the idea of choosing the most appropriate technique for each phase and of dynamically combining them so as to build a *record linkage workflow*, given specific application constraints and input data features.

Keywords: record linkage, software for record linkage

1. Introduction

Record linkage is a process that essentially aims to quickly and accurately identify if two (or more) records represent or not the same real world entity. A record linkage project can be performed for different purposes and the variety of the uses makes it a powerful instrument to support decisions in large commercial organizations and government institutions.

In official statistics, the field in which this work is developed, the combined use of statistical survey and administrative data is largely widespread and strongly stimulates the investigation of new methodologies and instruments to deal with record linkage projects.

Since the earliest contributions to modern record linkage, dated back to Newcombe *et al*. (1959) and to Fellegi and Sunter (1969), there has been a proliferation of different approaches, that make use also of techniques based on data mining, machine learning, soft computing and others. However, despite this proliferation, no particular record linkage technique has emerged as the best solution for all cases. We believe that such a solution does not actually exist, and that an alternative strategy should be adopted. Specifically, record linkage can be seen as a complex process consisting of several distinct phases involving different knowledge areas, moreover for each phase several techniques can be selected. We consider that the choice of the most appropriate technique not only depends on the practitioner's skill but most of all it is application specific; moreover in some applications, there are not evidences to prefer a method to others. In addition, from the analyst's point of view, it is important to have the possibility to experiment alternative criteria and parameters in the same application scenario. These considerations led us to think that it could be reasonable to dynamically select the most appropriate technique for each phase and to combine the selected techniques for building a record linkage workflow of a given application.

In this paper we describe the RELAIS (Record Linkage At Istat) toolkit. This software allows combining techniques for each of the record linkage phases, so that the resulting workflow is actually built on the basis of application and data specific requirements. Moreover, this software aims to include not only a toolkit of techniques, but also a library of patterns that, given specific data and application requirements, could support the definition of the most appropriate record linkage workflow. The toolkit is being developed as an open source project. This is a choice motivated by the idea of re-using the several solutions already available for record linkage in the scientific community, and by the quite ambitious goal of providing, in the shortest possible time, a generalized toolkit for building dynamic record linkage workflows.

The paper is organized as follows. In Section 2, we outline the main phases in which a record linkage process can be decomposed, detailing three of them. In Section 3, we describe the idea, the design and the current state of implementation of RELAIS. Finally, in Section 4, some test scenarios on real data are described.

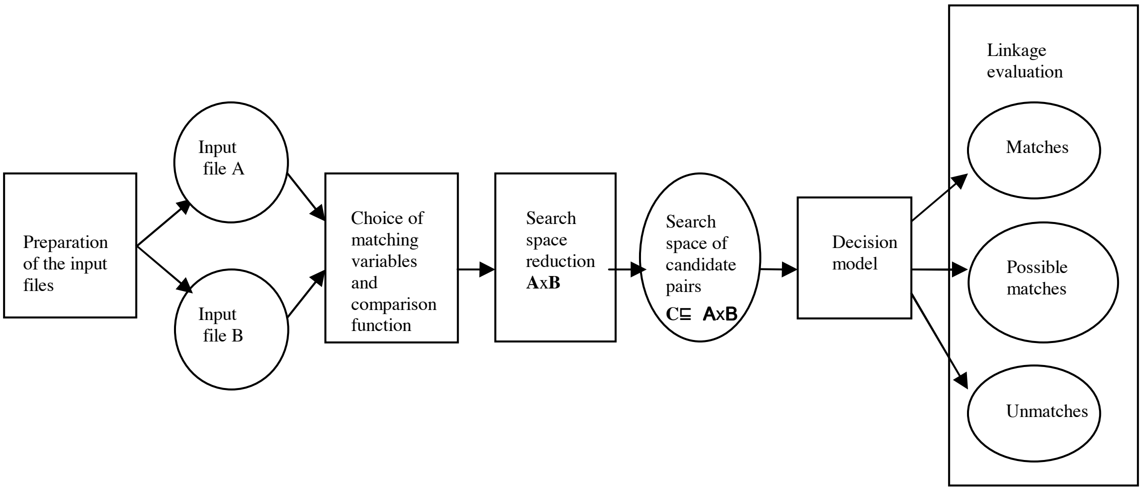
2. Phases of a record linkage project

The complexity of the whole linking process relies on several aspects; for example the lack of unique identifiers requires sophisticated statistical procedures, the huge amount of data to process involves complex IT solutions, constraints related to a specific application may require the solution of difficult linear programming problems. Due to such a complexity, it can be suitable to decompose a record linkage process into some main phases:

1. Pre-processing of the input files
2. Choice of the identifying attributes (matching variables)
3. Choice of the comparison function
4. Creation of the search space of link candidate pairs
5. Choice of the decision model
6. Selection of unique links
7. Record linkage evaluation

In Figure 1, the above listed phases are shown.

Figure 1: *Phases of a record linkage project*


Preparation of the input files
Input file A
Input file B
Choice of matching variables and comparison function
Search space reduction AxB
Search space of candidate pairs C⊆ AxB
Decision model
Linkage evaluation
Matches
Possible matches
Unmatches

Below we briefly report the main aspects of the record linkage, aiming at introducing and formalizing the problem.

Let be *A* and *B* two lists of size n_A and n_B. The goal of record linkage is to find all pairs of units (*a,b*), *a*∈*A*, *b*∈*B*, such that a and b refer to the same unit (*a*=*b*). Starting from the set $\Omega = \{(a,b); a \in A, b \in B\}$ containing all possible pairs of records from the lists *A* and *B*, with size $|\Omega| = N = n_A \times n_B$, a record linkage procedure is a decision rule based on the comparison of k matching variables; for each single pair of records, one of the following decisions can be taken: the pair is a link, the pair is a possible link or the pair is a non-link. Since the matching variables can be prone both to measurement errors and misreporting, the record linkage problem is far from being a trivial one. The comparison between the matching variables of the two units (*a,b*) is made by means of a suitable comparison function, depending on the kind of variables and their accuracy. For each pair of the set Ω, the result of the comparison of the matching variables is summarized in the vector γ, called comparison vector. For instance, when the comparison function applied to the k matching variables is the equality, the resulting k-dimensional comparison vector is composed by 1 or 0, depending on agreement or disagreement of the variables:

$$\gamma = (\gamma_1, ..., \gamma_j, ..., \gamma_k) \longrightarrow \gamma = (1,...,0,...,1).$$

According to a general point of view, including both the deterministic and the probabilistic approach to record linkage, an overall matching weight *r* is assigned to each pattern of the comparison vector. If the weight assignment follows a probabilistic mechanism, then the approach to record linkage is probabilistic too; on the contrary, if the weight is assigned according to deterministic rules, the approach to the linkage is deterministic or empirical. This paper deals explicitly with the probabilistic approach (detailed in Section 2.2), due to the fact that it allows to evaluate the quality of the linkage results and to identify a larger number of links when data are affected by errors.

Generally speaking, the pattern of a comparison vector belonging to pairs representing links is associated to an high value of the composite weight *r*; for instance the pattern $\gamma = (1,...,1,...,1)$ composed by all value equal to 1 is associated to the highest *r* value. On the contrary the pattern of a comparison vector belonging to pairs representing non-links is associated to a low value of the composite weight *r*; for instance, the pattern $\gamma = (0,...,0,...,0)$ composed by all value equal to 0 is associated to the lowest *r* value. Through to the composite weight *r*, each pair is classified as a link if the corresponding weight *r* is above a certain threshold T_m, and as a non-link if the weight lays below the threshold T_u; finally, for the pairs corresponding to weights falling into the range $I=(T_u , T_m)$, no-decision is made and the pair is assigned to a clerical review analysis. The threshold levels should be chosen in order to properly manage the trade off between the need of a small number of expected no-decisions and small misclassification error rates for the pairs.

The weight assignment (i.e. the choice of the decision model) is the core of a linkage project, but, as specified above, also other relevant steps can be identified and practitioners must usually tackle with them. According to Gill (2001), 75% of the whole effort for the implementation of a record linkage procedure is required by the first phase, the *preparation of files*. As a matter of fact, data can be stored in different formats and some items may be missing or with inconsistency or errors. The key job of this phase is to convert the input data in a well defined format, resolving the inconsistencies that may be present. Notice that many true matches may be erroneously classified as non-matches because of errors in comparing the matching variables. In this phase null string are cancelled, abbreviations, punctuation marks, upper/lower cases, etc. are cleaned and any necessary transformation is carried out so as to standardize variables. Furthermore the spelling variations are replaced with standard spelling for the common words. A parsing procedure which divides a free-form field into a set of strings, could be applied and a schema reconciliation can be performed to avoid possible conflicts (i.e. description, semantic and structural conflicts) among data source schemas so as to have standardized data fields.

The next sections are devoted to the definition of other record linkage phases: the *reduction of the search space of the candidate pairs*, the *probabilistic model* and the *selection of unique links*, respectively.

2.1 Search space reduction

Generally speaking, the pairs needed to be classified as matches, non-matches and possible matches are those in the cross product *A x B* of the record stored in each input file, A and B. When dealing with large datasets, the creation, the store and the management of the cross product *A x B* could be almost impracticable; as a matter of fact, while the number of possible matches increases linearly, the computational problem raises quadratically, that is the complexity is $\boldsymbol{O}(n^2)$ (Christen and Goiser, 2005). To reduce this complexity, which is an obvious cause of problems for large databases, it is necessary to reduce the number of pairs (*a; b*). Blocking and sorted neighbourhood are the two main methods which aim to reduce the number of comparison between records. *Blocking* sets out to remove pairs of records that are no matches; it consists of partitioning the two sets into blocks and of considering linkable only records within each block. The partition is made through blocking keys: two records belong to the same block if all the blocking keys are equal or if a hash function applied to the blocking keys of the two records gives the same result.

Sorted neighbourhood sorts by the same variable the two record sets and searches possible matching records only inside a window of a fixed dimension which slides on the two ordered record sets.

2.2 Probabilistic decision model

Fellegi and Sunter (1969) firstly defined record linkage as a decision problem, where each pair of the comparison space Ω must be assigned to the set of matches *M* or to the set of non-matches *U*. The distribution of the comparison vector $\gamma = (\gamma_1,\dots,\gamma_j,\dots,\gamma_k)$, calculated as the result of a comparison function on the *k* matching variables, for all the pairs in the space Ω, is supposed to come from a mixture of two different (unobserved) distributions: the first one comes from the pairs (*a,b*) which actually are the same unit, called distribution *m*; the other one comes from the pairs (*a,b*) which actually represent different units, called distribution *u*. The estimation of these two distributions requires the use of iterative methods, generally the EM algorithm or its generalizations, due to the latent unknown random variable "link status", that assigns each pair to the set *M* or to the set *U*.

Once the two distributions *m*(γ) and *u*(γ) are estimated, it is possible to define the composite matching weight, that can be read as a likelihood ratio:

$$r = \frac{m(\gamma)}{u(\gamma)} = \frac{\Pr(\gamma \mid M)}{\Pr(\gamma \mid U)}$$

where *M* is the set of the pairs which actually are links and *U* is the set of the pairs corresponding to non-links, with $M \cup U = \Omega$ and $M \cap U = \varnothing$.

2.3 Reduction from multiple linkage to unique linkage

In several applications, the record linkage target is to recognize exactly and univocally the same units and to establish only unique or "1 to 1" links. In other words, the linkage result must satisfy the constraint that one record on file A can be assigned to one and only one record on file B, and vice-versa. This kind of application requires several constraints and is a difficult problem of optimization, for which different algorithms have been proposed.

For instance Jaro (1989), suggested to formulate it as a linear programming problem: once the matching weight is assigned to each pair, the identification of 1 to 1 links can be solved maximizing the objective function given by the sum of weights for the link pairs, under the constraints given by the fact that each unit of A can be linked almost with one unit of B and vice-versa. According to Jaro (1989), this is a degenerate transportation problem, and the use of such a linear programming model to provide the assignments represents an advance with respect to other *ad hoc* assignment methods. In order to formulate the problem, let C_{ij} be the matrix containing the composite weights for all pairs, the maximizing function is:

$$Z = \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} C_{ij} X_{ij}$$

under the $\mathrm{n_A} + \mathrm{n_B}$ constrains:

$$\sum_{i=1}^{n_A} X_{ij} = 1, \qquad j=1,2, \ldots, n_B$$

$$\sum_{j=1}^{n_B} X_{ij} = 1, \qquad i=1,2, \ldots, n_A$$

where C_{ij} is the gain (in terms of the log transformation of the weight *r*) of matching record *i* on file A with record *j* on file B, X_{ij} is an indicator variable that is 1 if record *i* is assigned to record *j* and 0 if *i* is not assigned to record *j* .

3. RELAIS: a software for record linkage

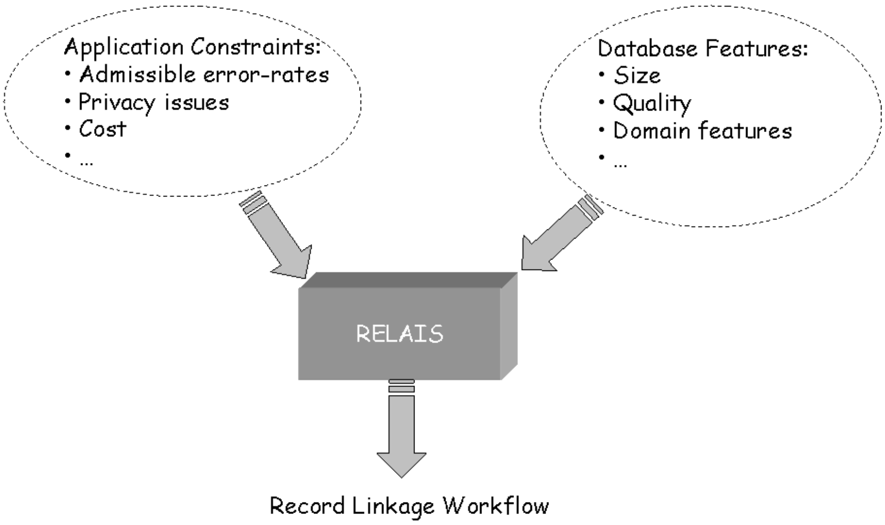
Due to the great attention to the integration data matters and the complexity of the problems, several record linkage systems and tools have been proposed, in both the academic and private sectors. Such tools include, for example, Big Match (Yancey, 2007), CANLINK (Fair, 2001), Febrl (http://www.sourceforge.net/projects/febrl), Link Plus (http://www.cdc.gov/cancer/npcr/tools/registryplus/lp.htm), Tailor (Elfeky *et al.* 2002), The Link King (http://www.the-link-king.com). The first two systems have been developed at the U.S. Bureau of the Census and the Statistics Canada respectively, the other systems have been developed at medical-epidemiological centres or at universities. Some of the systems provide a certain degree of flexibility for the user; for instance, Febrl allows to choose which comparison function can be more appropriately applied. However, any of these tools provides the flexibility of multiple choices for *each* of the record linkage phase.

In order to deal with both the modularity of the record linkage problem and the need of flexible choices, we propose the RELAIS toolkit (Fortini *et al.*, 2006) (Tuoto *et al.*, 2007).

The inspiring principle is to allow combining the most convenient techniques for each of the record linkage phase and also to provide a library of *patterns* that could support the definition of the most appropriate workflow, in both cases taking into account the specific features of the data and the requirements of the current application. In such a way, the toolkit not only provides a set of different techniques to face each phase of the linkage problem, but it could also be seen as a compass to solve the linkage problem as better as possible, given the problem constrains. In addition, RELAIS aims at joining specifically the statistical and computational essences of the matching issue.

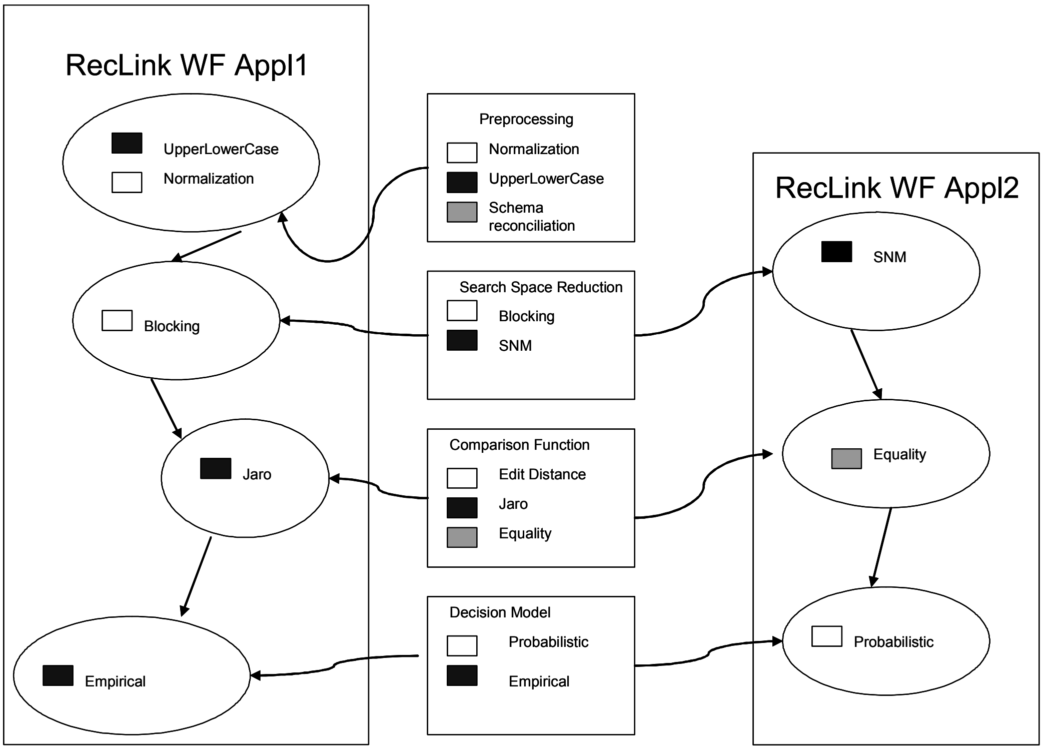
The RELAIS toolkit idea is based on the consideration that the record linkage process is application dependent. Indeed, available tools do not provide a satisfying answer to the various requirements that different applications can exhibit. As seen in the previous section, the record linkage process consists of different phases; the implementation of each phase can be performed according to a specific technique or on the basis of a specific decision model. For instance, choosing which decision model to apply is not immediate: the usage of a probabilistic decision model can be more appropriate for some applications but it can be less appropriate for others, for which an empirical decision model could prove more successful. Furthermore, even using the same decision model in different application scenarios, a comparison function could fit better than others. Therefore, we claim that no record linkage process, deriving from the choice and combination of a specific technique for each phase, is the best for all applications.

Figure 2: *The RELAIS's input-output*



Therefore, the RELAIS toolkit is composed by a collection of techniques for each record linkage phase that can be dynamically combined in order to build the *best record linkage workflow*, given a set of application constraints and data features provided as input (see Figure 2). As an example, if it is known that the datasets to compare have poor quality, it is suitable the usage of comparison functions ensuring high precision (e.g. Jaro distance); as a further example, if no specific error-rates are required by the application, it can be appropriate the usage of an empirical decision model. Some phases of the record linkage process can be missing: for instance the search space reduction phase makes sense only for huge data volumes, or for applications that have time constraints. In Figure 3, examples of possible workflows that may be built with the RELAIS toolkit are shown.

Figure 3: *Examples of RELAIS's workflows*



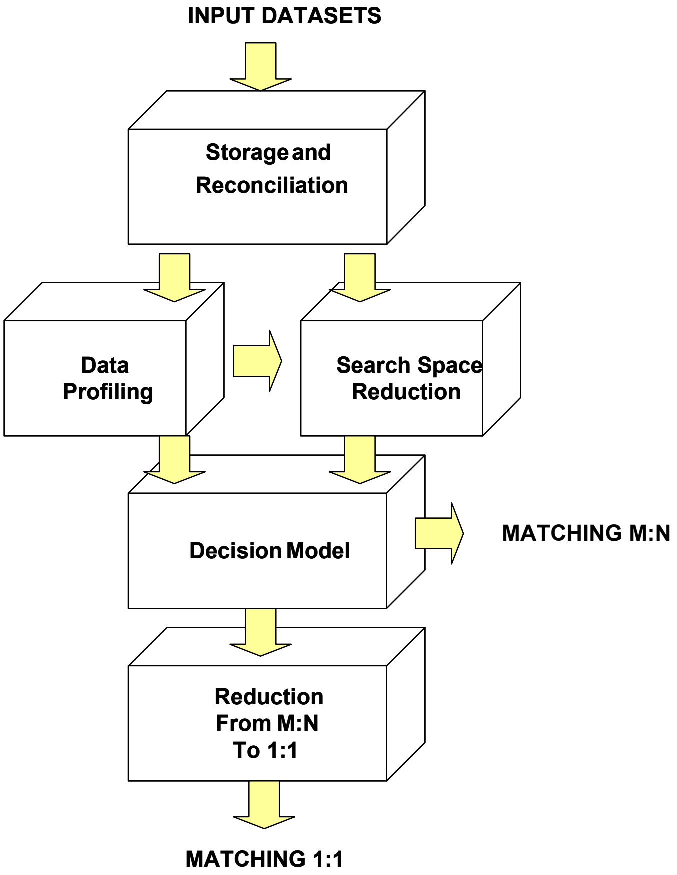
RELAIS is configured as an open source project. There are at least two reasons for this choice. First, as often highlighted above, there are many possible techniques that can be implemented for each of the record linkage phases: relying on a community of developers such set can be increased and maintained very rapidly. Second, we do believe that there have been, in the last years, several independent efforts towards the definition of a record linkage project and that such efforts have not led to the best for all solution. An open source record linkage project could instead give the possibility of gathering together the efforts already done, according to the idea described above, in order to make them available to the community for the most appropriate usage.

RELAIS has been implemented using two languages based on different paradigms: Java, an object oriented language and R (*http://www.r-project.org*), a functional language. This choice depend on our belief that a record linkage process is composed of techniques for manipulating data, for which Java is more appropriate, and of computation-oriented techniques for which R is more appropriate. Java and R have been chosen because they reflect the open source philosophy of the RELAIS project.

3.1 Design and implementation choices of RELAIS

RELAIS gives the opportunity to design different record linkage workflows. As shown in Figure 4, the principal phases of the record linkage process have already been implemented (with one or more techniques): (i) storage and reconciliation, (ii) data profiling, (iii) search space reduction, (iv) decision model and (v) reduction from matching M:N to matching 1:1.

Figure 4: *Status of implementation of RELAIS*



The phases (ii) and (iv) will be respectively detailed in Sections 3.2 and 3.3. In the following, we provide some design and implementation details for the remaining phases.

In the storage and reconciliation phase, the input datasets are internally stored and a partial schema reconciliation is performed. Specifically, common variables between the two datasets are identified, and though the dataset can be different in the number of variables they have, the system will keep just the common ones for subsequent processing.

As far as the search space reduction phase, we have implemented both blocking and sorted neighbourhood, in addition to the creation of the search space by means of the cross product of the two files. Both methods are implemented by means of an external memory sort, in order to be able to process huge amount of data without having main memory as a bottleneck.

With respect to the decision model phase, we have implemented the Fellegi-Sunter probabilistic model by using the EM algorithm for the estimation of the model parameters. The method takes as input a contingency table, which reports the frequencies of the agreement patterns resulting from the application of the comparison function, and the output is a many to many linkage of the datasets records. Starting from this output, we can propose to the user the clusters of matches, non-matches and possible matches.

The storage and reconciliation phase and the search space reduction one are implemented in Java, while the decision model phase is implemented in R.

3.2. A focus on data profiling for selecting matching and blocking variables

RELAIS includes a data profiling phase in which a set of quality metadata are calculated starting from real data; these metadata help the user in the critical phase of choosing the best blocking or matching variables.

Generally speaking, the matching variables determine if a pair of records identifies or not the same unit. If unique identifiers are available in the data sources, the easiest and most efficient way is to use these ones as link variables; but very strict controls are needed when using just numeric identifiers. Otherwise, if unique identifiers are lacking, the choice of the common identifying attributes is more difficult, it is typically not automatic and is generally done by a domain expert. In any case, the choice can be supported by some helpful information, deriving from metadata description and simple statistics on the variables distribution, in order to select those with a high identification power and low error and missing rates.

In general, the correct identification of the links depends on the number of matching variables but, at the same time, if strongly correlated variables or variables with correlated errors are included in the model, the estimates could be not reliable, thus increasing the values of the matching weights without improving the identification of the links. The identification power of a variable increases according to its different values and depends on the distribution of these values among the units: when a variable has a large number of categories, but few of these are much more frequent than others, it would be useless to select them as matching variables. The larger the number of categories of a variable is, the higher is its discriminative power.

As regards the blocking variables, in order to reduce the search space of the candidate pairs, the most suitable variables are again those most discriminating and accurate, i.e., not affected by errors or missing. In fact, links are searched only within the blocks, assuming that there are no matches out of them; so, if the blocking variable is error affected, some true links could be missed. Furthermore, it is useful to avoid blocking variables which create too small groups (i.e. blocking variable with a large amount of values) in order to reduce risk of errors; in addition, also blocking variables which create too large groups must be avoided, generally because they do not allow to reduce enough the search space. It is suitable to create blocks of the same size, selecting one or more variables, which present a consistent number of values uniformly distributed among the units.

In selecting the matching and blocking variables, a data profiling phase could be very useful to guide the practitioners. In this paper some indicators are proposed: (1) variable completeness, (2) identification power, (3) accuracy, (4) internal consistency, (5) consistency, (6) correlation, (7) entropy. All these metadata can be evaluated for each variable, and are merged together into a quality vector associated to the variable itself. The first five indicators are useful to select matching and blocking variables as well, while the sixth is more relevant to select matching variables and the last is more suitable to select blocking variables.

Given a dataset *A* of size *N* with variables $(X_1,\ldots,X_k)$, we have:

1. **Completeness**. Let $V_i=\{v_{i1},\ldots,v_{iN}\}$ be the set of values of the variable X_i and $\underline{V_i}=\{v_{ij} \in V_i \mid v_{ij} \neq \text{NULL}, j \in \{1,\ldots,N\}\}$ the set of non missing values for the variable X_i, the completeness of X_i is defined as:

$$Compl(X_i) = \frac{|V_i|}{N}.$$

The completeness of a variable is defined as the proportion of non-missing values for the variable on the total number of records.

2. **Identification power**. Let n_i be the number of different values of the variable X_i. The identification power of X_i is defined as:

$$PI(X_i) = \frac{n_i}{N}.$$

The identification power is given by the ratio between the number of the different values recorded for the variable and the total number of records. This indicator is useful for selecting matching and blocking variables, but when creating blocks is important to avoid variables with very high value of identification power, due to the risk of creating too much small blocks, implying loss of matches.

3. **Accuracy.** We measure accuracy with respect to reference dictionaries of values that are known to be correct. Let $V_i^a=\{v_{ij} \in V_i \mid v_{ij}$ is labelled as "accurate"$\}$ be the set of values known to be accurate; the accuracy of X_i is defined as:

$$ACC(X_i) = \frac{|V_i^a|}{N}.$$

The accuracy indicator implies the comparison of the values of a variable with a dictionary or a set of reference values. The measure provides the number of correct values on the overall. If used to select the matching variables, a low value of this indicator can suggest to adopt a suitable comparison function.

4. **Internal consistency.** It is a specific type of accuracy computed on pairs of variables X_i and X_k, with respect to a list of paired values known to be correct. Let $V_{ik}^a=\{< v_{ij}, v_{kj} >$, with $v_{ij} \in V_{i,}\ v_{kj} \in V_k \mid <v_{ij,}\ v_{kj>,}\ j \in \{1,\ldots,N\}$, is labelled as "accurate"$\}$ be the set of values known to be accurate; the internal consistency of X_i and X_k is defined as:

$$Cons(X_i, X_k) = \frac{|V_{ik}^a|}{N}.$$

5. **Consistency.** It takes into account the number of internal consistency relationships in which a variable X_i is involved. Being X_{ij}^c the set of variables X_j such that each X_j is involved in a consistency check with $X_{i,}$ consistency is defined as:

$$Cons(X_i) = \frac{\sum_j Cons(X_i, X_{ij}^c)}{|X_{ij}^c|}.$$

The consistency indicator represents how well each item of the considered variable relates independently to the rest of the items on a scale.

6. **Correlation.** A simply measure of the correlation between two variables is given by the chi-squared coefficient. Let n_i be the number of different values of the variable X_i and n_j be the number of different values of the variable $X_{j;}$ let $f_{i.}$ *(i=1,…, n_i)* be the marginal relative frequency of the i-th value of the variable X_i, $f_{.j}$ *(j=1,…, n_j)* be the marginal relative frequency of the j-th value of the variable X_j and $f_{ij.}$ be the relative frequency of the double distribution of the variables X_i and X_j. For each variable X_i, a correlation index with the other variables can be calculated as:

$$\chi^2(X_i, X_j) = \sum_i \sum_j \frac{f_{ij}^2}{f_{i.}f_{.j}} - 1\,.$$

In order to build a suitable probabilistic linkage model, it is appropriate to select as matching variables only one among those with high correlation coefficient.

7. **Entropy.** Let n_i be the number of different values of the variable X_i and $V_i=\{v_{i1},\dots,v_{iN}\}$ be the set of values of the variable X_i , the entropy, or heterogeneity relative index, is defined as:

$$e(X_i) = -\sum_{j=1}^{n_i} |V_{ij}| \log_{n_i}\left(|V_{ij}|\right).$$

The entropy index helps to select blocking variables, representing how the different values are concentrated, so permitting to get blocks of the similar sizes.

Referring to this set of quality metadata as quality vector q, a ranking within the quality vectors can be performed in order to suggest which variable is more suitable for blocking or matching. As far as the procedure adopted to rank the quality vectors obtained by metadata evaluation is concerned, two distinct solutions can be adopted:

1. **One-step ranking.** The first option takes into account the fact that some elements of the quality vector associated with the X_i variable may have missing values, i.e. the evaluation of some specific metadata is not possible or not required. Some "dummy values" can be used in place of such missing values that can be set by the user; these dummy values can be the mean, the maximum or the minimum of the metadata values computed for the other attributes different from X_i. Vectors are then compared each other by means of a Euclidean weighted norm, i.e.:

$$\|q\| = \left(\sum_{i=1}^{n} w_i q_i^2\right)^{1/2},$$

where $q=(q_1,\dots q_n)$ is a vector of quality metadata and $w=(w_1,\dots w_n)$ is a vector of weights. The quality vectors are then ordered on the basis of such a norm.

2. **Two-step ranking.** The second option is performed in two steps: (i) evaluation of the norm on the present metadata, i.e. ignoring missing ones, and sorting on the basis of such a norm; (ii) refinement of the ranking performed at step (i) by using an *insertion sort,* i.e. by comparing only the values of common metadata for each couple of vectors.

3.3 A focus on reduction from multiple linkage to unique linkage

The principal output of the Fellegi-Sunter procedure is a set of matching pairs. Being A and B the two dataset which undergo the record linkage process, in the general case, each record of A can be matched to M (where M>=0) records of B and each record of B can be matched to N (where N>=0) records of A.

However, several applications require that records of the two data sets are matched "uniquely", that is each record of A with at most one of B and vice-versa. In other words, there may be the problem of reducing an M:N matching to a 1:1 matching. In the case of a probabilistic record linkage procedure, in which for each pairs of records belonging to A and B the ratio *r* is computed, this problem can be formulated as an optimization problem.

As described in Section 2.3, a possible formulation of the 1:1 matching is as a linear programming (LP) problem. In this case, the simplex algorithm can be used for the solution. The worst case complexity of the simplex method is exponential, though in practice it exhibits polynomial time complexity, hence it is quite efficient.

In order to solve the cardinality reduction problem formulated as an LP problem we have used a package of the "R" language named "lpSolve".

Such a formulation has proven not very efficient in practice, especially for the need of a "high" dimensionality of the data structures to be used. Indeed, a memory overflow in a PC environment was caused even by small instances of the problem. In Table 1, we report the execution times for different input sizes, with test performed with a PC with 756MB RAM and with the R version 2.5.1 (with an extension of the virtual memory up to 3 GB).

Table 1: *Performance of LP*

Dataset Dimensions (#records)	LP TIME (sec)
10 960	2.95
32 636	42.13
76 720	Out of memory

The input datasets for the tests shown in Table 1 are actual outputs of the Fellegi and Sunter procedure and the number of records corresponds to the pairs with the associated value of *r* found by the procedure. As we can see, for 76 720 records, i.e. for 280 distinct record in A associated to 274 distinct records in B, the solution to the 1:1 matching problem implemented by the simplex implementation of the lpSolve package already fails because of memory problems.

An alternative formulation of the 1:1 matching problem can be done by relying onto combinatorial optimization, and specifically the problem can be formulated as a "maximum bipartite matching". A "matching" on a graph G is a set of edges of G such that no two of them share a vertex in common. Bipartite matching gives as a result of the matching a bipartite graph. By constraining the bipartite matching to be a 1:1 matching with weights represented by *r* coefficients, we get exactly our 1:1 matching problem. In order to solve the problem formulated as a maximum bipartite matching, we have used the R package "optmatch". Such a formulation has proven very efficient in practice too, especially with respect to the size of the needed data structures when compared to the LP formulation.

We have also experimented a modification to the LP formulation by means of an optimization that selects only the pairs of records with values of *r* greater than one. In practice, given the specific distribution of *log(r)* values, such values are associated to non-matching pairs with a high probability and do not contribute to the maximization of the objective function. While still maintaining good level of optimality of the result, the optimized LP formulation also proved very efficient in practice.

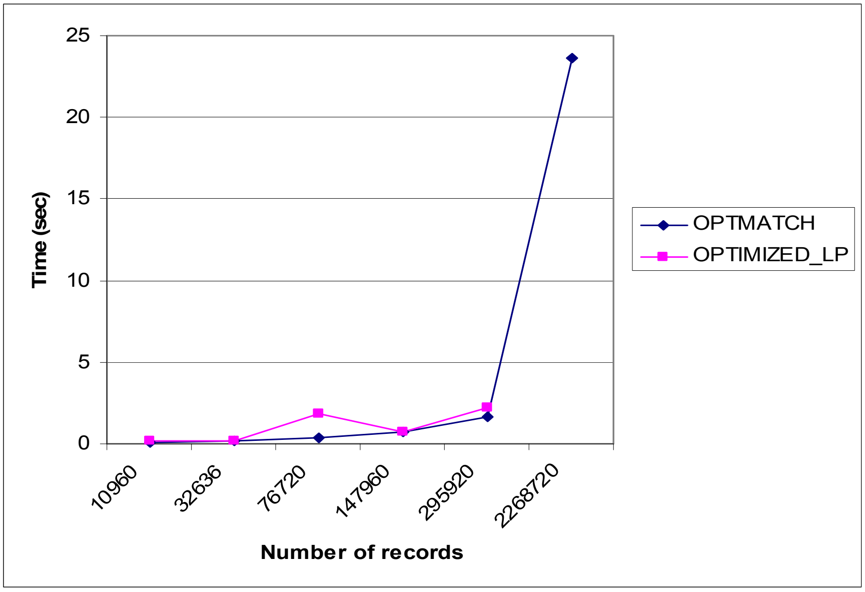
In Figure 5, we show the time performance of the optmatch solution and of the optimized LP one, for different input sizes, we recall that the number of records corresponds to the pairs with the associated value of *r* found by the procedure.

Looking at the graph, we can see:

- Both solutions (optmatch and optimized LP) produce a result in a few seconds and with greater input sizes when compared to the LP solution
- Optmatch has a polynomial time behaviour and still produces an output for more than 2 000 000 of records.

The solution to the 1:1 matching problem implemented in RELAIS 1.0 is however the optimized LP one. Indeed, the optmatch package contains an implementation of the Bertsekas and Tseng’s relax-iv algorithm (Bertsekas D. P. and Tseng, 1994) which has a licence with some restrictions for governmental usage. Nevertheless, we do not exclude to consider it in future versions of RELAIS, depending on the results of other solutions we are experimenting.

Figure 5: *Performances of optmatch and optimized LP solution*



4. Test scenarios

The below described tests referred to data from the 2001 Italian Population Census and its Post Enumeration Survey (PES). The main goal of the Census was to enumerate the resident population at the Census date, 21/10/2001. The PES instead had the objective of estimating the coverage rate of the Census; it was carried out on a sample of enumeration areas (called *EA* in the following), which are the smallest territorial level considered by the Census. The size of the PES's sample was about 70 000 households and 180 000 individuals while the variables stored in the files are name, surname, gender, date and place of birth, marital status, etc. Correspondingly, comparable amounts of households and people were selected from the Census database with respect to the same EAs. The PES was based on the replication of the Census process inside the sampled EAs and on the use of a capture-recapture model (Wolter K., 2006) for estimating the hidden amount of the population. In order to apply the capture-recapture model, after the PES enumeration of the statistical units (households and people), a record linkage between the two lists of people built up by the Census and the PES was performed. In this way the rate of coverage, consisting of the ratio between the people enumerated at the Census day and the hidden amount of the population, was obtained.

The estimates of the Census coverage rate through capture-recapture model has required to match Census and PES records, assuming no errors in matching operations. Therefore the linkage between the two sources was both deterministic and probabilistic and the results was checked manually; all the linkage operations lasted several working days. Due to the accuracy of the matching procedures adopted, we know the true linkage status of all candidate pairs, in this way we can evaluate the performances of the linkage developed with the techniques described in Section 3 and implemented in RELAIS.
The RELAIS performances are tested on three different data sets sizes:

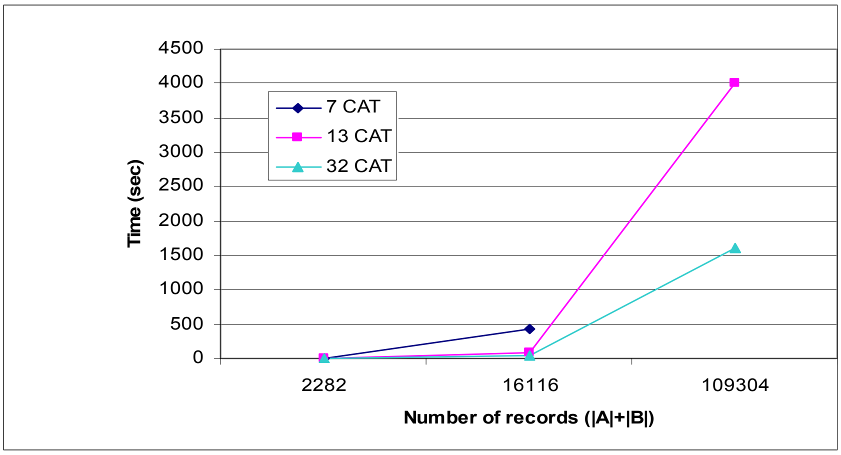
1. 1 000 individuals;
2. 8 000 individuals;
3. 50 000 individuals.

The efficiency performances of the blocking methods are tested on all the data sets, while the effectiveness performances of the overall blocking procedure are evaluated on the 8 000 record size data set.

4.1 Efficiency performances of the probabilistic linkage

We performed efficiency experiments to test the blocking method. In Figure 6, we show results with a PC environment of 756MB RAM. Specifically, the execution time for creating blocks is drawn against the number of records obtained as the sum of the records of the two input datasets. Different behaviours are shown, while varying the blocking variable. In particular, we report three different trends according to three distinct number of modalities for the chosen blocking variables.

Figure 6: *Performance of blocking method*



As we can see, there is a bottleneck for the blocking method when the number of modalities is not high, and the number of records is more than about 15 000. In fact, in this case, the disk space acts as a bottleneck, because we write on it a distinct file for each block. This is a specific design choice in order to give the user the possibility of analyzing the created blocks; in the next future, we plan to permit the user to choose the option of writing or not the blocks. As a further improvement, we plan to introduce a database management system for optimize this phase, as well as some other phases (such as sorted neighbourhood) which exhibit the same problems.

When increasing the number of modalities of the blocking variable, the blocking method achieves a final result, this is because smaller files are created in correspondence of the blocks.

As we can see, in general execution times are quite low: they are always below one hour, except for datasets of more than 100 000 records, for which the total execution time is of about 1 hour and 10 minutes.

4.2 Effectiveness performances of the probabilistic linkage

The effectiveness of the linkage performances were evaluated in terms of match rate, false match rate and false non-match rate. The match rate is defined as the number of linked record pairs divided by the total number of true match record pairs. The false match rate and the false non-match rate correspond to the well-known type II and type I errors in a one-tail hypothesis test context. The false non-match rate indicates the ratio between the number of incorrectly non matched records and the whole number of the true matched records. False non-matches are the most common and occur when records which should have been assigned to the same unit are instead not matched. The false match rate denotes the ratio between the records incorrectly matched and the whole number of matched pairs. False matches are less common but potentially more serious because of further analyses on erroneously linked data could lead to biased statistics. Other authors consider performance measures in terms of positive predicted value and sensitivity, that consist of the algebraic transformation of the false match rate and the false non-match rate.

The efficacy performances were tested applying the linkage procedure, as described in Section 3 using the RELAIS software, on the two data set of size 8 000 records, ignoring the known true matching status. As matching variables all the strongest identifiers were used: name and surname, gender, day, month, and year of birth. The equality were applied as comparison function. The parameters of the Fellegi-Sunter probabilistic model were estimated via the EM algorithm. Two thresholds were fixed in order to individuate the tree sets of Matches, of Unmatches and of Possible Links. The upper threshold was fixed assigning to the set of Matches all the pairs with the composed matching weights correspondent to estimated matching probability higher than 0.99; the set of the possible links were created fixing the lower threshold level with the composite matching weight correspondent to the estimated matching probability lower than 0.50. The pairs falling into the set of the Possible Links were assigned to the set of Matches without a clerical supervision of the results.

A blocking phase were performed considering as blocking variable the month of birth of the household header. In this way 12 blocks were created, plus a residual block formed by the units with missing information about the month of birth of the household header. The resulting blocking size are quite similar and homogeneous. The overall match rate is equal to 82%, the false match rate is 0.5% and the false non-match rate is 12%, as resulting form Table 2. Those results are comfortable and quite optimistic if compared with those coming from the scientific community, when a record linkage is performed in analogous conditions in terms of identification variables, number of matched records, kind of matched units. The results have to be regarded also more optimistic considering the unsupervised possible link data processing. Anyway, when the linkage is finalized to evaluate coverage rate, as in Census Post Enumeration Survey, the value of the false non-match rate has to be as small as possible and the resulting 12% false non-match rate is too high. In this situation, a further linkage procedure should be applied to the records non-linked at the first time, if it is possible without using blocking phase, so to minimize the risk of loosing matches.

Table 2: *Linkage results*

		True Linkage Status		
		Matched	*Not Matched*	
Results of the Linkage Procedure	*Matched*	6 016	30	6 046
	Not Matched	856	689	
		6 872		

Table 3 allows to analyze the results of the linkage procedure in more detail. For each block the table reports the amount of pairs linked by the procedure, the number of pairs that the procedure identifies as possible links and for which a manual review or a more in-depth analysis is suggested, and finally the matches missed by the software procedure (i.e. the false non-matches); moreover both true matches and false matches are specified respectively for the linked pairs and the possible-link ones. Looking at Table 3, it is possible to note that, as expected, the false matches introduced in the possible linked pairs are more considerable than the number of false matches introduced in the pairs linked with certainty, which are quite ignorable. Regarding the missed matches, the most of them is introduced by the blocking procedure itself, because true links cannot be individuated due to the fact that the records do not agree on the blocking variable. In particular, few categories of the blocking variable, corresponding to months 'June' and

'September', are specially affected by errors that cause a higher amount of false non-matches.

Table 3: *Linkage results in blocks*

Block	**Linked Pairs**	***True Matches in Linked Pairs***	***False Matches in Linked Pairs***	**Possible Link Pairs**	***True Matches in Possible Link Pairs***	***False Matches in Possible Link Pairs***	**Missed Matches**
1	506	*506*	*0*	40	*39*	*1*	49
2	470	*470*	*0*	20	*17*	*3*	51
3	489	*489*	*0*	41	*39*	*2*	67
4	473	*473*	*0*	10	*10*	*0*	55
5	499	*498*	*1*	39	*35*	*4*	68
6	413	*412*	*1*	27	*26*	*1*	104
7	504	*503*	*1*	33	*30*	*3*	50
8	513	*513*	*0*	33	*32*	*1*	73
9	473	*470*	*3*	53	*48*	*5*	81
10	492	*492*	*0*	42	*42*	*0*	55
11	419	*419*	*0*	33	*32*	*1*	46
12	397	*396*	*1*	27	*25*	*2*	49
-	-	-	-	-	-	-	108
Tot	5 648	*5 641*	*7*	398	*375*	*23*	856

Another relevant point regards the time and the efforts consumed in performing the linkage. With respect to the data considered in this experiment, the complex linkage procedure applied for obtaining the Post Enumeration Survey estimates required several days of work and more than one devoted person. On the contrary, the linkage performed by RELAIS was obtained in less than one day by only one person.

5. Concluding remarks

In official statistics, data integration is of major interest as a mean of using available information more efficiently. Record linkage is among the principal activities of data integration by taking into account errors on data and fostering reconciliation of data values. In this paper, we have illustrated the RELAIS project an open source toolkit for building record linkage workflows. The idea of this project has been developed keeping in mind: (i) the complexity of a record linkage problem, which involves different techniques and sciences; (ii) the opportunity of treating the linkage with modularity, identifying several phases which can occur, even iteratively; and (iii) the different suitable approaches depending on both the data features (e.g. type of data, amount of data) and the application requirements (e.g. efficiency, efficacy, accuracy). The toolkit aims to offer multiple techniques for record linkage, both deterministic and probabilistic, and also the possibility of building ad-hoc solution combining each modules. This approach allows to overcome the question on which method is better than others, being convinced that actually there is not a technique dominating all the others.

In the paper, we have described the main phases of a record linkage process detailing the search space reduction phase and the probabilistic decision model, implemented in the actual version of RELAIS. We have also described the data profiling phase that helps the user in the difficult phases of choosing matching and blocking variables. Moreover, we detailed the reduction from M:N linkage to 1:1 linkage phase illustrating the different options we have explored and their performances. Finally we have described some test scenarios on real data, presenting performance and effectiveness results.

In future work, we plan to extend the current functionalities of RELAIS and to optimize its performances. First, we plan to add a deterministic decision model. Then, we will remove some restricting hypothesis of the probabilistic model such as the conditional independency of matching variables. In order to optimize the performances, we plan to introduce a DBMS in the RELAIS architecture and to experiment the usage of different algorithms to speed up and enhance some phases such as the reduction to 1:1 linkage phase.

References

Ananthakrishna R., Chaudhuri C., and Ganti V. (2002) Eliminating Fuzzy Duplicates in Data Warehouses. In Proceedings of VLDB 2002, Hong Kong, China.

Bertolazzi P., Santis L.D., and Scannapieco M. (2003)Automatic Record Matching in Cooperative Information Systems. In Proceedings of the ICDT'03 International Workshop on Data Quality in Cooperative Information Systems (DQCIS'03), Siena, Italy.

Bertsekas D. P. and Tseng P. (1994) “Relax-iv: a Faster Version of the relax Code for Solving Minimum Cost Flow Problems,” Tech. rep., M.I.T., report P-2276.

Chauduri S., Ganti V., and Motwani R. (2005) Robust identification of fuzzy duplicates. In Proceedings of ICDE 2005, Tokyo, Japan.

Christen P. and Goiser K. (2005) Assessing duplication and data linkage quality: what to measure?, Proceedings of the fourth Australasian Data Mining Conference, Sydney.

Ding Y. and Fienberg S.E. (1994) Dual system estimation of Census undercount in the presence of matching error, *Survey Methodology*, 20, 149-158.

Elfeky M., Verykios V., and Elmagarmid A. K. (2002) Tailor: A Record Linkage Toolbox. In Proceedings of the 18th International Conference on Data Engineering. IEEE Computer Society, San Jose, CA, USA.

Fair M. (2001) Recent developments at statistics canada in the linking of complex health files. In Federal Committee on Statistical Methodology, Washington D.C.

Febrl. http://www.sourceforge.net/projects/febrl.

Fellegi I. and Sunter A. (1969) A Theory for Record Linkage. *Journal of the American Statistical Association*, 64.

Fortini M., Liseo B., Nuccitelli A., and Scanu M. (2001) On Bayesian record linkage. *Research in Official Statistics*, 4:185-198.

Fortini M., Scannapieco M., Tosco L. and Tuoto T. (2006) Towards an Open Source Toolkit for Building Record Linkage Workflows, In Proc. of SIGMOD 2006 Workshop on Information Quality in Information Systems (IQIS'06), Chicago, USA.

Gill L. (2001) Methods for Automatic Record Matching and Linkage and their Use in National Statistics. National Statistics Methodological Series no. 25, HMSO Norwich, UK.

Gu L., Baxter R., Vickers D. and Rainsford C. (2003) Record linkage: Current practice and future directions. Technical Report 03/83, CSIRO Mathematical and Information Sciences, Canberra, Australia.

Gu L. and Baxter R. (2004) Adaptive filtering for efficient record linkage. In Proceedings of the Fourth SIAM International Conference on Data Mining.

Hernandez M. and Stolfo S. (1998) Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem. *Journal of Data Mining and Knowledge Discovery*, 1(2).

Jaro M. (1985) Advances in Record Linkage Methodologies as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of American Statistical Society*, 84(406):414-420.

Koudas N. and Srivastava D. (2005) Approximate joins: Concepts and techniques. In Proceedings of VLDB.

Monge A. and Elkan C. (1997) An Efficient Domain Independent Algorithm for Detecting Approximate Duplicate Database Records. In Proceedings of SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'97), Tucson, AZ, USA.

Newcombe H., Kennedy J., Axford S. and James A. (1959) Automatic Linkage of Vital Records, Science, Vol.130 pp. 954-959.

The-Link-King. http://www.the-link-king.com.

The-R-Project for Statistical Computing. http://www.r-project.org/.

Tuoto T., Cibella N., Fortini M., Scannapieco M. and Tosco L. (2007) RELAIS: Don't Get Lost in a Record Linkage Project, In Proc. of the Federal Committee on Statistical Methodologies (FCSM 2007) Research Conference, Arlington, VA, USA.

Winkler W. (2000) Frequency-based matching in Fellegi-Sunter model of record linkage. Technical report, U.S. Bureau of the Census - Washington D.C. Technical Report RR/2000/06, Statistical Research Report Series.

Winkler W. (2001) Record Linkage Software and Methods for Merging Administrative Lists. Technical report, U.S. Bureau of the Census - Washington D.C. Technical Report RR/2001/03, Statistical Research Report Series.

Winkler W. (2004) Methods for Evaluating and Creating Data Quality. *Information Systems*, 29(7).

Wolter K. (1986) Some coverage error models for census data. *Journal of the American Statistical Association*, 81:338-346.

Yancey W. (2007) BigMatch: A Program for Extracting Probable Matches from a Large File. Technical report, Statistical Research Division U.S. Bureau of the Census - Washington D.C. Research Report Series - Computing n. 2007-01.

2

Statistical matching and forecasting

The validity of data fusion

Hans Kiesl, Susanne Rässler
Institute for Employment Research, Federal Employment Services, Regensburger Straße 104, 9047Nürnberg, Germany
University of Bamberg, Feldkirchenstraße 21, 96045 Bamberg, Germany
e-mail: hans.kiesl@iab.de, susanne.raessler@uni-bamberg.de

1. Introduction

Statistical matching techniques typically aim to achieve a complete data file from different sources that do not contain the same units. On the contrary, if samples are exactly matched using identifiers such as social security numbers or name and address, this is called record linkage. Traditionally, statistical matching is done on the basis of variables common to all files. Statistical twins, i.e., donor and recipient units that are similar according to their common variables, are usually found by means of nearest neighbor or hot deck procedures. The specific variables of a donor unit which are observed only in one file are added to the record of the recipient unit to finally create the matched sample. We like to note that in our sense statistical matching is not restricted to the case of merging different samples without overlap. Also one single file may contain some records with observations on more variables than others, then, these records can be matched with those containing less information based on the variables common to all units.

In this paper we refer to the situation of data fusion which means there are groups of variables that are *never jointly observed*, say X and Y. In all other cases of statistical matching we assume that, at least, every pair of variables has been jointly observed in one or the other data set. The fusion of data sets with the aim of analyzing the unobserved relationship between X and Y and addressing quality of data fusion is done, e.g., by National Statistical Institutes such as Statistics Canada or the Italian National Institute of Statistics, see, e.g., Liu and Kovacevic (1997) or D'Orazio et al. (2003). The focus often is on analyzing consumers' expenditures and income, which are in detail only available from different surveys. In the U.S., e.g., data fusion is used for microsimulation modeling, where "what if" analyses of alternative policy options are carried out using matched data sets, see Moriarity and Scheuren (2001, 2003). Especially in Europe and among marketing research companies, data fusion has become a powerful tool for media planning, see, e.g., Wendt (1986). Often surveys concerning the purchasing behavior of individuals or households are matched to those containing valuable information about print, radio and television consumption.

2. Data Fusion and its Identification Problem

2.1 Traditional Fusion Algorithms

The general benefit of data fusion is the creation of one complete data source containing information about all variables. Without loss of generality, let the (X,Z) sample be the recipient sample *B* of size n_B and the *(Y,Z)* sample the donor sample *A* of size n_A. The traditional matching procedures determine for every unit *i, i* = 1,…, n_B, of the recipient sample with the observations (x_i,z_i) a value *y* from the observations of the donor sample. Thus, a composite data set $(x_1,\tilde{y}_1,z_1),\ldots,(x_{n_B},\tilde{y}_{n_B},z_{n_B})$ with n_B elements of the recipient sample is constructed. The main idea is to search for a statistical match, i.e., for a donor unit *j* with $(y_j,z_j)\in\{(y_1,z_1),\ldots,(y_{n_A},z_{n_A})\}$ whose observed data values of the common variables z_j are identical to those z_i of the recipient unit *i* for *i* = 1,…,n_B. Notice that $\tilde{y}_i$ is not the true *y*-value of the *i*-th recipient unit but the *y*-value of the matched statistical twin. In the following, all density functions (joint, marginal, or conditional) and their parameters produced by the fusion algorithm are marked by the symbol ~. Notice that $\tilde{Y}$ is called fusion or imputed variable herein.

A typical matching algorithm chooses randomly among all possible statistical matches for each recipient unit *i* (i.e. among all (y_j, z_j) with $z_j = z_i$); we shall call this the ideal case thereafter. In reality, not every recipient allows for an exact match in the common variables; therefore some nearest neighbor rules are usually imposed. There are very sophisticated fusion techniques in practice; for an overview see Rässler (2002).

In order to judge the quality of any data fusion procedure, it is essential to study how the true (only partially known) distribution $f(x,y,z)$ and the fusion distribution $\tilde{f}(x,y,z)$ are related. In the ideal case, it can be shown that the joint distributions of *X* and *Z* and of *Y* and *Z* are unaltered by the matching algorithm. The overall joint distribution satisfies

$$\tilde{f}_{X,Y,Z}(x,y,z) = f_{X,Z}(x,z)\cdot f_{Y|Z}(y\,|\,z);$$

see Rässler (2002) for technical details. Obviously, the fusion distribution equals the true distribution if and only if $f_{Y|X,Z} = f_{Y|Z}$, i.e., if *Y* and *X* are conditionally independent given *Z*. This implicit assumption of traditional algorithms was first pointed out by Sims (1972); see also Rodgers (1984) for an enlightening discussion.

Rässler and Fleischer (1998) show that in the ideal case, the fusion covariance between *X* and *Y* is given by

$$\tilde{\mathrm{cov}}(X,Y) = \mathrm{cov}(\mathrm{E}(X\,|\,Z),\mathrm{E}(Y\,|\,Z)).$$

Because in general,

$$\mathrm{cov}(X,Y) = \mathrm{E}(\mathrm{cov}(X,Y\,|\,Z)) + \mathrm{cov}(\mathrm{E}(X\,|\,Z),\mathrm{E}(Y\,|\,Z))$$

holds, the fusion covariance $\tilde{\mathrm{cov}}(X,Y)$ equals the true covariance, if and only if $\mathrm{E}(\mathrm{cov}(X,Y\,|\,Z)) = 0$, i.e., if *X* and *Y* are on the average conditionally uncorrelated given *Z*. Notice that variables which are conditionally independent are also conditionally uncorrelated and, of course, on the average conditionally uncorrelated, but not vice versa in general. If *f* is multinormally distributed, however, these concepts coincide, since in this case the conditional covariance $\operatorname{cov}(X,Y \mid Z=z)$ is given by $\operatorname{cov}(X,Y) - \operatorname{cov}(X,Z)\operatorname{var}(Z)^{-1}\operatorname{cov}(Z,Y)$, which is independent of *z*.

With small sample sizes, the ideal case is seldom observed. However, simulation studies have shown that these derivations are even approximately valid, if nearest neighbour algorithms are applied (see Rässler 2002).

Summing it up: Traditional algorithms produce fusion data sets which reflect the true joint distribution only in the case of conditional independence of *X* and *Y* given *Z*. The true covariance structure is retained in the fused file only in the case of *X* and *Y* being on the average conditionally uncorrelated given *Z*. The question that naturally arises is: can we learn from the data, whether these assumptions are met?

2.2 The Identification Problem of Data Fusion

2.2.1. Joint Distributions

Data fusion initially is connected to an identification problem concerning the joint distribution and the association of the specific variables that are never jointly observed. For every pair of specific variables (X_i,Y_j), the marginal joint cumulative distribution function $F_{X_i,Y_j}(x,y)$ is bounded by the Fréchet-Hoeffding inequality, although it is usually not very informative:

$$\max\{F_{X_i}(x) + F_{Y_j}(y) - 1,\ 0\} \le F_{X_i,Y_j}(x,y) \le \min\{F_{X_i}(x), F_{Y_j}(y)\}. \qquad (1)$$

With common variables *Z* these bounds can be slightly improved, since the same inequalities are valid for the conditional distributions either (Ridder and Moffitt 2006):

$$\begin{aligned}\max\{F_{X_i|Z=z}(x \mid Z=z) + F_{Y_{j|Z=z}}(y \mid Z=z) - 1,\ 0\} \quad &\le \quad F_{X_i,Y_j|Z=z}(x,y \mid Z=z)\\ \le \min\{F_{X_i|Z=z}(x \mid Z=z), F_{Y_j|Z=z}(y \mid Z=z)\}.&\end{aligned}$$

Taking expectations over *Z*, we have

$$\begin{aligned}\mathrm{E}\big(\max\{F_{X_i|Z=z}(x \mid Z=z) + F_{Y_{j|Z=z}}(y \mid Z=z) - 1,\ 0\}\big) \quad &\le \quad F_{X_i,Y_j}(x,y)\\ \le \mathrm{E}\big(\min\{F_{X_i|Z}(x \mid Z), F_{Y_j|Z}(y \mid Z)\}\big)&\end{aligned} \qquad (2)$$

While F_{X_i} and F_{Y_j} might be estimated with sufficient accuracy from the samples, this is probably not always true for the expectations in (2), especially in the case of continuous *Z*. Thus, in practice the unconditional bounds might be the more reliable choice, although the lower and upper bounds are usually quite far apart and therefore rather useless in reality. The lesson to be learned is, by means of the observed data we are not able to decide which joint distribution (given that it lies within the Fréchet-Hoeffding bounds) could have generated the data.

2.2.2 Correlation Structure

Consider, for example, a univariate common variable *Z* determining another variable *X* which is only observed in one file. Suppose first that *X* and *Z* be linearly dependent, i.e., let the correlation $\rho_{ZX} = 1$, and thus *X* = *a*+*bZ* for some real-valued *a* and *b* ($b \neq 0$). The correlation between this common variable *Z* and a variable *Y* in a second file may be $\rho_{ZY} = 0.8$. It is easy to see that the unconditional correlation of the two variables *X* and *Y* which are not jointly observed is determined by *Z* with $\rho_{XY} = \rho_{a+bZ\,Y} = \rho_{ZY} = 0.8$. If the correlation between *X* and *Z* is less than one, say 0.9, we can easily calculate the possible range of the unconditional association between *X* and *Y* by means of the determinant of the covariance matrix which has to be positive semidefinite; i.e., the determinant of the covariance matrix cov(*Z,Y,X*) must be positive or at least zero, see, e.g., Cox and Wermuth (1996).

Given the above values and setting the variances to one without loss of generality, the covariance matrix of (*Z,Y,X*) is

$$\operatorname{cov}(Z,Y,X) = \begin{pmatrix} 1 & 0.9 & 0.8 \\ 0.9 & 1 & \operatorname{cov}(X,Y) \\ 0.8 & \operatorname{cov}(X,Y) & 1 \end{pmatrix}$$

with

$$\det(\operatorname{cov}(Z,Y,X)) = -\operatorname{cov}(X,Y)^2 + 2\cdot 0.72\operatorname{cov}(X,Y) - 0.45\,.$$

Calculating the roots of det(cov(*Z,Y,X*)) = 0, we get the two solutions cov(*X,Y*) = $0.72 \pm \sqrt{0.0684}$. Hence we find the correlation bounded between [0.4585, 0.9815]; i.e., every value of the unknown covariance cov(*X,Y*) greater than 0.4585 and less than 0.9815 leads to a valid and thus feasible covariance structure for (*Z,Y,X*). By means of the observed data we are not able to decide which covariance matrix could have generated the data, provided that it is positive semidefinite.

Bearing these identification problems in mind, note that traditional data fusion algorithms make specific implicit assumptions (conditional independence or at least conditional uncorrelatedness on average) about the data. The need for alternative approaches that overcome these assumptions is obvious, although little research has been done in the literature so far.

Only few approaches, basically three different procedures, have been published to assess the effect of alternative assumptions about the inestimable correlation structure. One approach is due to Kadane (2001; reprinted from 1978), generalized by Moriarity and Scheuren (2001). The next approach dates back to Rubin and Thayer (1978), it is used to address data fusion explicitly by Rubin (1986), and generalizations are presented by Moriarity and Scheuren (2003). Both approaches use regression based procedures to produce synthetic data sets under various assumptions on this unknown association.

Finally, a full Bayesian regression approach using multiple imputations is first given by Rubin (1987, p. 188), and then generalized by Rässler (2002).

3. Calculation of Feasible Correlations

To ease notation, we again set all variances equal to 1. Consider again the correlation matrix $\Sigma := \mathrm{cov}(Z, Y, X)$ of all observed variables. Recall that Z is the vector of variables observed in both samples; Y and X are the vectors of variables which are only observed in sample A and B, respectively. The matrix Σ and its inverse can be partitioned corresponding to the partition of the complete data vector (Z,Y,X), to give

$$\Sigma = \begin{pmatrix} \Sigma_{ZZ} & \Sigma_{ZY} & \Sigma_{ZX} \\ \Sigma_{YZ} & \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XZ} & \Sigma_{XY} & \Sigma_{XX} \end{pmatrix} \qquad \Sigma^{-1} = \begin{pmatrix} \Sigma^{ZZ} & \Sigma^{ZY} & \Sigma^{ZX} \\ \Sigma^{YZ} & \Sigma^{YY} & \Sigma^{YX} \\ \Sigma^{XZ} & \Sigma^{XY} & \Sigma^{XX} \end{pmatrix}$$

In the case of data fusion, Σ_{YX} consists of the correlations between variables that are never jointly observed and may therefore not be directly estimated from the data. However, as we will discuss below, there is information in the data about their feasible values.

Correlation matrices have to be positive semidefinite; apart from the case of exact linear dependence they are positive definite. We will ignore this distinction and assume positive definiteness, since an exact linear relationship never occurs in sample data (or can be easily detected and removed).

All other submatrices of Σ apart from Σ_{YX} can be estimated from the two samples. Therefore, Σ is only partially determined; since we know that it has to be positive definite, Σ is called a partial positive definite matrix. Finding the set of feasible correlation matrices in this case is a special application of what is called matrix completion problems in matrix theory; we are interested in positive definite completions of Σ.

Due to the special structure of Σ, a positive definite completion of Σ always exists. Moreover, there is a unique positive definite completion, whose determinant is maximal, and this matrix is the unique one whose inverses has zeros in those positions corresponding to the unspecified entries in Σ, i.e. $\Sigma^{YX} = 0$ (see Grone et al. 1984). Consider now the matrix $\Sigma^{*}_{YX|Z}$ of partial covariances of X and Y given Z, i.e. the covariance matrix of the residuals of linear least squares regression of every component of X and Y on all components of Z. (Notice that partial covariances and conditional covariances are different concepts. In case of multivariate normality these matrices coincide, whereas in general the two concepts produce different results.)

$\Sigma^{*}_{YX|Z}$ can be easily derived from the simple correlation matrix as the Schur complement of Σ_{ZZ} in Σ (see e.g. Whittaker 1990, p.135):

$$\Sigma^{*}_{YX|Z} = \begin{pmatrix} \Sigma_{YY|Z} & \Sigma_{YX|Z} \\ \Sigma_{XY|Z} & \Sigma_{XX|Z} \end{pmatrix} = \begin{pmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{pmatrix} - \begin{pmatrix} \Sigma_{YZ} \\ \Sigma_{XZ} \end{pmatrix} \Sigma_{ZZ}^{-1} \begin{pmatrix} \Sigma_{ZY} & \Sigma_{ZX} \end{pmatrix} \tag{3}$$

There is an interesting relationship between the partitioned inverse of Σ and the partial covariance matrix: The term $\Sigma^{YX} = 0$ if and only if the partial correlations between X and Y given Z vanish, i.e. $\Sigma_{YX|Z} = 0$ (Whittaker 1990, p. 144). Hence zero partial correlations given Z maximize the determinant of Σ among all feasible correlation matrices; the corresponding simple correlations being $\Sigma_{YX} = \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX}$. Notice that in case of normality, this is the correlation matrix of the fused data set that traditional algorithms create.

Positive definiteness places restrictions on the feasible correlations between X and Y. In general it is a difficult task to describe the set of feasible values in closed form. Kadane (2001) and Moriarity and Scheuren (2001) provide formulae for univariate X and univariate Y with multivariate Z. For multivariate X or multivariate Y, no closed form yet exists in the literature. One way to numerically tackle this problem is via grid search over all possible completions of Σ and deciding for every value if the completion is positive definite; see Rässler (2002) for an example of this approach.

In the following, we show that even in case of either multivariate X or multivariate Y (though not both), one can derive the range of all feasible solutions analytically.

Let (without loss of generality) X be univariate, i.e. $\Sigma_{XX} = 1$, so that Σ_{ZX} and Σ_{YX} are column vectors. Since all leading principal submatrices of Σ are fully specified and (by assumption of consistency) positive definite, the positive definiteness of Σ is equivalent to the determinant of Σ being positive, i.e. $\det(\Sigma) > 0$. Partitioning Σ and using a standard argument on the determinant of a partitioned matrix leads to the following condition:

$$\begin{pmatrix} \Sigma'_{ZX} & \Sigma'_{YX} \end{pmatrix} \begin{pmatrix} \Sigma_{ZZ} & \Sigma_{ZY} \\ \Sigma'_{ZY} & \Sigma_{YY} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{ZX} \\ \Sigma_{YX} \end{pmatrix} < 1 \tag{4}$$

The inverse can be written in closed form:

$$\begin{pmatrix} \Sigma_{ZZ} & \Sigma_{ZY} \\ \Sigma'_{ZY} & \Sigma_{YY} \end{pmatrix}^{-1} = \begin{pmatrix} \Sigma_{ZZ}^{-1}(\mathrm{I} + \Sigma_{ZY}C\Sigma'_{ZY}\Sigma_{ZZ}^{-1}) & -\Sigma_{ZZ}^{-1}\Sigma_{ZY}C \\ -C\Sigma'_{ZY}\Sigma_{ZZ}^{-1} & C \end{pmatrix} =: \begin{pmatrix} A & B \\ B' & C \end{pmatrix}$$

with $C := (\Sigma_{YY} - \Sigma'_{ZY}\Sigma_{ZZ}^{-1}\Sigma_{ZY})^{-1}$.

After straightforward calculation (4) evolves into

$$\Sigma'_{YX}C\Sigma_{YX} + 2\Sigma'_{ZX}B\Sigma_{YX} + \Sigma'_{ZX}A\Sigma_{ZX} < 1. \tag{5}$$

From this inequality, the geometric shape of the set of feasible correlations can be determined. Since C is positive definite, the set of possible vectors Σ_{YX} satisfying (5) is the interior of an n-dimensional ellipsoid (n being the dimension of vector Y). Transforming (5) into the normal form of an ellipsoid in order to be able to calculate its centre and axes, we get

$$(\Sigma_{YX} + C^{-1}B'\Sigma_{ZX})' \cdot \tilde{C} \cdot (\Sigma_{YX} + C^{-1}B'\Sigma_{ZX}) < 1$$

with $\tilde{C} := (1 + \Sigma'_{ZX}(BC^{-1}B' - A)\Sigma_{ZX})^{-1}C$.

Thus, the centre of the ellipsoid is $-C^{-1}B'\Sigma_{ZX}$. Plugging in the formulae for B and C yields

$$-C^{-1}B'\Sigma_{ZX} = \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX};$$

from this it can be seen that the correlation vector providing zero partial correlation (which maximizes the determinant) is the centre of the ellipsoid.

Final calculations give $1 + \Sigma'_{ZX}(BC^{-1}B' - A)\Sigma_{ZX} = 1 - \Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX}$, from which $\tilde{C}$ can be computed:

$$\tilde{C} = (1 - \Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX})^{-1} \cdot (\Sigma_{YY} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY})^{-1}.$$

The semi-axes of the ellipsoid are in the direction of the eigenvectors of $\tilde{C}$ (or C), the lengths of the semi-axes are given by $1/\sqrt{\lambda_i}$, where λ_i is the i-th eigenvalue of $\tilde{C}$ (i = 1,…, n).

The volume of the ellipsoid of feasible correlations (which is proportional to the product of the lengths of its semi-axes) might be considered as a new quality index for a data fusion process: the less volume the ellipsoid has, the greater is the explanatory power of the common variables and the less uncertainty remains for creating the fused data set.

In some cases, the marginal distributions might restrict the set of feasible correlation matrices even further. To see this, consider again the Fréchet-Hoeffding inequality (1). The upper and lower bounds are valid bivariate distributions, whose correlation coefficients are upper and lower bounds of possible correlations given the marginals (Tchen 1980). Thus, for every pair (X_i,Y_j) of specific variables, this inequality might place an additional restriction to the feasible correlations (in case of normality every correlation can be achieved with any marginal distributions, therefore no further restriction can be imposed).

If there are lots of ordinal variables in the samples, it is appropriate not to consider Bravais-Pearson correlation coefficients but to use association measures based on ranks. Frequently Spearman's ρ or Kendall's τ are measures of interest, even in metric settings. Since correlation matrices based on these measures also have to be positive definite (note that they can be expressed as Bravais-Pearson correlations for recoded variables), the results of this section remain valid, if consideration is upon matrices of Spearman or Kendall correlations rather than upon Bravais-Pearson correlation coefficients.

4. Summary and Outlook

In this paper we derived bounds for the correlations between variables not jointly observed, provided that one of the vectors of specific variables is univariate, and suggest a new quality index of data fusion which is built upon these bounds. Using our results, multiply imputed datasets can be produced according to different admissible correlation structures between X and Y by using appropriate algorithms (e.g. NIBAS, see Rässler 2002; notice that since data fusion can be viewed as a problem of missing data, multiple imputation procedures are applicable in general). Analyzing the different fused data sets can then reveal sensitivity to the different assumptions about the correlation structure between the variables that have never been jointly observed.

References

Box G.E.P., Tiao G.C. (1992) *Bayesian Inference in Statistical Analysis*, New York, Wiley.

Cox D.R., Wermuth N. (1996) *Multivariate Dependencies*, London, Chapman and Hall.

Grone R., Johnson C.R., Sá E.M., Wolkowicz H. (1984) Positive Definite Completions of Partial Hermitian Matrices, *Linear Algebra and its Applications*, 58, 109-124.

Kadane J.B. (2001) Some Statistical Problems in Merging Data Files, *Journal of Official Statistics*, 17, 423-433.

Liu T.P., Kovacevic M.S. (1997) An Empirical Study on Categorically Constrained Matching, *Proceedings of the Survey Methods Section, Statistical Society of Canada*, 167-178.

Moriarity C., Scheuren F. (2001) Statistical Matching: A Paradigm for Assessing the Uncertainty in the Procedure, *Journal of Official Statistics*, 17, 407-422.

Moriarity C., Scheuren F. (2003) A Note on Rubin's Statistical Matching Using File Concatenation With Adjusted Weights and Multiple Imputations, *Journal of Business and Educational Studies*, 21, 65-73.

D'Orazio M., Di Zio M., Scanu M. (2004) Statistical matching and the likelihood principle: uncertainty and logical constraints, *ISTAT Technical Report* 1/2004.

Rässler S. (2002) *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*, Lecture Notes in Statistics, 168, New York, Springer.

Rässler S., Fleischer K. (1998), Aspects Concerning Data Fusion Techniques, *ZUMA Nachrichten Spezial*, 4, 317-333.

Ridder G., Moffitt R. (2006) The Econometrics of Data Combination, in Heckman, J.J., and E.E. Leamer (eds.), *Handbook of Econometrics Volume 6*, Amsterdam: North Holland (to appear).

Rodgers W.L. (1984) An Evaluation of Statistical Matching, *Journal of Business and Economic Statistics*, 2, 91-102.

Rubin D.B. (1986) Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations, *Journal of Business and Economic Statistics*, 4, 87-95.

Rubin D.B. (1987) *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

Rubin D.B., Thayer D. (1978) Relating Tests Given to Different Samples, *Psychometrika*, 43, 3-10.

Sims C.A. (1972) Comments, *Annals of Economic and Social Measurement*, 1, 343-345.

Tchen A.H. (1980) Inequalities for Distributions with Given Marginals, *Annals of Probability*, 8, 814-827.

Wendt F. (1986) Einige Gedanken zur Fusion, in Arbeitsgemeinschaft Media-Analyse e.V. (eds.), *Auf dem Wege zum Partnerschaftsmodell*, Frankfurt, Media-Micro-Census GmbH, 109-140. [In German]

Whittaker J. (1990) *Graphical Models in Applied Multivariate Statistics*, Chichester, Wiley.

Statistical matching of two surveys with a non randomly selected common subset [(⋆)]

Nicola Torelli[1], Marco Ballin[2], Marcello D'Orazio[2], Marco di Zio[2]
Mauro Scanu[2], Gianni Corsetti[3]

[1] Dipartimento di Scienze Economiche e Statistiche, Università di Trieste, Italy
e-mail: nicola.torelli@econ.units.it
[2] Istat, Rome, Italy
[3] Inea, Rome, Italy

Abstract Statistical matching techniques are aimed to combine information available in two distinct datasets. Usually it is assumed that records in the two datasets refer to different units. When the two datasets contain data collected in surveys it is often the case that a not negligible number of units are included in both the surveys. Information collected on these units can convey information useful to asses crucial assumptions commonly adopted for statistical matching.

Keywords: file concatention, sampling weights, uncertainty, incomplete and syntetic two way stratification, agricoltural enterprises surveys

1. Introduction

Statistical matching techniques (D'Orazio et al, 2006) are aimed to combine information available in two distinct datasets. It is often the case that the two datasets, A and B, contain data collected in two independent sample surveys of size n_A and n_B respectively and such that (i) the two samples contain distinct units (the samples do not overlap); (ii) the two samples contain information on some variables X (common variables), while other variables are observed distinctly in one of the two samples, say, Y in A and Z in B. Common variables X can be used to create synthetic records containing information on joint ditribution (X, Y, Z) never jointly observed, but properties of the synthetic archive obtained need a careful examination. Without external information, one has often to rely upon the assumption of conditional independence between Y and Z given X. Strategies to alleviate, or to control for, conditional independence assumption are based on (a) availability of auxiliary information, (b) development of methodologies that study parameter uncertainty due to lack of joint information on the variables of interest.

When data are collected according to complex survey designs it can happen that a (often small) set of units is included in both the samples. This is, for instance, the case of enterprise surveys where fine stratification and pps selection of the units is usually adopted. For the units in the common subset data on (X, Y, Z) are collected and these can provide valuable auxiliary information to avoid conditional independence assumption, or at least to evaluate its appropriateness.

A problem that arises when combining survey data is that one has to consider which units are included in the synthetic file in order to appropriately use survey weights to

[(⋆)] Work supported by Italian Ministry of University and Research, Prin 2005: "Innovative Methods, organization and contents in sample surveys for agricolture and environment"

estimate population parameters from the combined archive. This problem is even more relevant when analysing also the data from the common subset.

We will consider alternative approaches for dealing with the problem of statistical matching of data collected in two sample surveys with complex survey desegn with a common subset. Each approach has its merits and integrating the approaches can help inference on the (Y, Z) distribution.

The first approch follows Rubin (1986). He proposed a new procedure for statistical matching, called file concatenation, where all the units of the two archives can be merged into a single archive using appropriate (possibly multiple) imputation techniques to fill in the missing information. The procedure involves computation of weights for the units of the two samples merged into the concatenated archive, under the reasonable assumption that each sample is selected according to a different survey design.

The second approach, proposed by Rennsen (1998), takes esplicitly into account the availability of data collected in a small supplementary survey specifically designed to collect data on (X, Y, Z) and it is based on calibration of survey weigths in order to exploit this information.

The third approach is analysis of uncertainty (D'Orazio et *al.*, 2006a) where properties of the unobserved (Y, Z) distribution (in terms of interval of plausible values) are inferred by marginal and conditional distributions actually estimable from surveys A and B.

The paper is motivated by application of statistical matching techniques to data collected in two important Italian surveys on agricoltural enterprises: the Farm Structural Survey (hereafter FSS) and the Farm Accountancy Data Network Survey (FADN). The two surveys are designed to investigate separate phenomena: simply stated, structure of the farm is the focus of the first, economic accounts of the second. Combining information from the two sources is potentially of great interest. For these surveys the design is complex and design variables, although available for both the surveys, are not the same in the two cases. Moreover, the sampling units are farms and probability that some units (large farms, for instance) are included in both the samples is far from being negligible. This implies that for a small fraction of units the variables X, Y and Z are jointly observed.

The paper is organized as follows: section 2 describes main features of the two surveys, section 3 introduces main ideas of statistical matching techniques and the three approaches to combine FSS and FADN taking into account data in the common subset are presented. Some preliminary results are presented in section 4.

2. The case study: the FADN and FSS Surveys

The Farm structural survey (FSS) is a survey carried out on farms every two years. Its main objective is to investigate the principal phenomena like crops, livestock, machinery and equipment, labour force, holder's family characteristics.

More specifically the FSS used for this study, has been carried out at the end of the agricultural year 2003 (November 2002 - October 2003)

The target population of the survey is defined as the set of farms which in the agricultural year 2003 have the following characteristics:

- the agricultural area utilized for farming is one hectare or more, or;

- the agricultural area utilized for farming is less than one hectare if they produce a certain proportion for sale (2500 euro) or if their production unit has exceeded certain physical threshold.

The sampling units have been selected according to a stratified sample design with a take all stratum containing the largest farms. The total sample size is 55,030. 53000 units have been selected from the target population and about 2,000 units have been selected from the set of other small units enumerated by census. Furthermore all farms resulting from a splitting or a merging of a sampling unit have been added to the sample by the interviewers.

The stratification of units has been carried out as follows:

First - the take all stratum has been defined using the size of the farms expressed in terms of Utilized agricultural area (UAA), Livestock size unit (LSU) and Economic size unit (ESU) of each unit;

Second - the reference population has been stratified according to location (region or province), dimension (UAA, LSU and ESU) and typology of the agricultural holdings.

Third - the remaining units of the population list have been stratified using the region code.

The Farm Accountancy Data Network Survey (FADN - also known as REA in Italian) collects data on the economic structure and results of the farms, as costs, added value, employment labour cost, household income etc. The target population consists of those farms satisfying the following definition: “The UAA is at least one hectare or, if the UAA is less than one hectare, its economic dimension is large enough (more than 2066 euro; of the production is sold)”.

According to this definition, the reference population consists of 2,1 millions of units. The sample is selected according to a stratified random sample with a take all stratum containing the largest farms, in terms of ESU. Stratification has been defined with respect to region or province code, typology classification (first digit), ESU classes, working days classes. The sample consists of 20317 units.

It should be noted that the target population coincides for both the surveys. The selection of the units in the two surveys is negatively coordinated, in order to reduce the response burden. This negative coordination is made possible by attaching to each units of the population frame the same random number. This procedure is described below, where for the sake of simplicity, we refer to the intersection of stratum r in FADN and stratum s in FSS.

1. Assign a permanent random number between zero and one to each unit in the census frame of the Italian farms.
2. Order these units according to the permanent random number.
3. Select nr units with the lowest random numbers among the N_r units in stratum r. These units correspond to the r-th stratum of the FADN survey.
4. Add 0.5 to each permanent number (those units with new random number greater than 1 are shifted by subtracting 1).

5. Select the ns units with the lowest random numbers among the N_s units in stratum s. These units correspond to the s-th stratum of the FSS survey.

The overlap between the two surveys resulted in 1624 farms. Among these farms, 1593 belongs to the take all stratum of FSS.

3. Statistical matching of survey data with a common subset

3.1. General issues

The statistical matching problem in its basic form can be considered as an inferential problem with partial information. It is usually assumed that the two samples to match, denoted as A and B, do not overlap on the observed units and in this case the observed common set of variables (X) is the only available information for drawing inferences on the relationship between two other set of variables, Y and Z, observed in A and B respectively, and never jointly observed. This framework allows a pointwise estimation of parameters on X, $(Y|X)$ and $(Z|X)$, while anything related to the distribution of (Y, Z) or of $(Y, Z|X)$ can be estimated pointwise only under some simplifying assumptions, which are untestable for the data at hand. Usually, the conditional independence assumption is assumed, i.e. Y and Z are considered as independent variables given the common set of observed variables X. If there are not enough clues in order to assume such assumption, different procedures have been proposed in the statistical matching literature.

In the context we have considered, i.e. the statistical matching of FADN (A) and FSS (B), it is possible to exploit the fact that the two surveys do overlap among the observed units. This characteristic allows the definition of a subset of units, say C, where all the variables of interest (X, Y, Z) are jointly observed. This is a precious source of information that should be included in the statistical matching process of the two samples.

In the next section, we focus on some alternatives procedures (file concatenation, incomplete two-way stratification and synthetic two-way stratification) which have been defined in order to take into account that samples A and B are drawn according to complex survey designs. Note that the presence of complex survey designs makes the statistical matching problem more difficult, because there is also the problem of the treatment and harmonization of survey weights. File concatenation solves this problem creating a unique sample by the union of the two samples, and managing the survey weights of the two surveys so that the resulting weights of the concatenated sample can be considered as representative of the population of interest. Incomplete and synthetic two-way stratification retain the two different samples and tackle the problem of harmonising the two samples by calibrating the two sample weights on the common information in the two surveys.

Finally, to correctly evaluate results obtained from the above mentioned statistical matching startegies, it will be given some considerations on what could be the result of the statistical matching procedure without any assumptions on the statistical model of (X, Y, Z) and without the use of the file of completely observed observations C (in case this file is not considered as representative of the population of interest).

3.2. File concatenation of FSS and FADN

The original proposal of Rubin (1986) consisted in modifying the sample weights of the two surveys A and B in order to get a unique sample given by the union of A and B ($A \cup B$) with survey weights representative of the population of interest. The basic idea is that new sampling weights can be derived from the concatenated files by using the simplifying assumption that the probability of including a unit in both the samples is negligible. Rubin's procedure has been thoroughly reviewed by Moriarity and Scheuren (2003) who noted "The notion of file concatenation is appealing. However on a close examination it seems to have limited applicability" (p.71). In fact, in many cases, when different and complex survey designs are adopted in the two surveys and when probability that a unit belongs to both the samples is far from being negligible, computation of weights is unfeasible and a simplifying assumption, like the one suggested by Rubin, could be inappropriate.

In fact, this assumption, stating that the probability that any unit is selected in both tha samples is negligible, generally holds for two independent sample surveys, and allows to compute the inclusion probability of a record i in $A \cup B$ simply as:

$$\pi_i = \pi_{i,A} + \pi_{i,B} \tag{1}$$

Note however that for each unit in $A \cup B$ the inclusion probability of the records in A under the survey design in B, as well as the inclusion probability of the records in B under the survey design in A, must be computed. It is worth noting that design variables of a survey are not necessarily available in other surveys and for this reason the approach proposed by Rubin has been seldom applied.

As illustrated in section 2, survey design variables are known for all the units in the target population for both the FADN and the FSS, and consequently they represent a natural framework for the application of concatenated weights proposed by Rubin. Anyway, since the two surveys have been designed to allow a not negligible intersection the simplified concatenated weights in (1) cannot be used. For this reason, the inclusion probability of a generic unit i in $A \cup B$ becomes:

$$\pi_i = \pi_{i,FADN} + \pi_{i,FSS} - \pi_{i,FADN \cap FSS} \tag{2}$$

where $\pi_{i,FADN \cap FSS}$ is the probability that the intersection of the two samples include unit i. While $\pi_{i,FADN}$ and $\pi_{i,FSS}$ are known by design, it is more difficult to compute the probability that a population unit is included in both the samples, and its exact computation for the FADN and FSS surveys is unfeasible.

In order to compute these weights, we have considered a Monte Carlo simulation, following the approach suggested by Fattorini (2006). Let $\mathcal{P}$ be a with-frame population of size N. Fattorini suggests to estimate the first and second order inclusion probabilities according to the following steps.

1. Draw M independent samples $(S_1, ..., S_M)$ from the population $\mathcal{P}$ according to the survey design.
2. Estimate the first order inclusion probabilities π_j ($\pi_j > 0$) through the empirical inclusion probabilities

$$p_j = \frac{X_j + 1}{M + 1}, \qquad j = 1, 2, ..., N,$$

where X_j is the number of times unit j is included in the M samples.

3. Estimate the second order inclusion probabilities π_{jh} ($\pi_{jh} > 0$) through the empirical inclusion probabilities

$$p_{jh} = \frac{X_{jh}+1}{M+1}, \qquad h > j = 1, 2, ..., N,$$

where X_{jh} is the number of times units j and h are jointly included in the M samples.

The empirical inclusion probabilities are based on a Monte Carlo approach that is generally used to approximate the expected value of a function g of a random variable Y through the computation of g in a finite number of points. Let Y be a categorical variable with probability mass function $f_Y(y)$, the Monte Carlo estimate of

$$E[g(Y)] = \sum_y g(y) f_Y(y)$$

is given by

$$\widetilde{g}(Y) = \frac{1}{n}\sum_{i=1}^{n} g(y_i)$$

where y_i, $i = 1, \ldots, n$, are n observations drawn independently from $f_Y(y)$. The strong law of large numbers implies that $\widetilde{g}_n(Y)$ converges almost surely to $E[g(Y)]$ as n increases.

When computing inclusion probabilities, let δ_j be the indicator of inclusion of unit j in a sample $\mathcal{S}$

$$\delta_j = \begin{cases} 1 & \text{if } j \in \mathcal{S} \\ 0 & \text{otherwise.} \end{cases}$$

Hence

$$\pi_{i,FADN \cap FSS} = \sum \delta_i p(s^{(FADN)}, s^{(FSS)}) = E(\delta)$$

where the sum is over all the samples that can be drawn according to the FADN and FSS survey designs, and $p(s^{(FADN)}, s^{(FSS)})$ is the probability of drawing the two samples. Following the approach suggested in Fattorini (2006), this probability can be approximately given by computing the number of times that the ith observation is included in the two samples. More precisely we estimate the probability that the ith observation belongs to both the FADN and FSS samples by applying the following steps:

1. draw M independent samples $(s_j^{(FADN)}, s_j^{(FSS)})$, $j = 1, ..., M$, from the population $\mathcal{P}$ according to the two survey designs;
2. estimate the probabilities $\pi_{iFADN \cap FSS}$ ($\pi_{iFADN \cap FSS} > 0$) through the empirical inclusion probabilities

$$p_{j,FADN \cap FSS} = \frac{X_{j,FADN \cap FSS}+1}{M+1}, \qquad j = 1, 2, ..., N,$$

where $X_{j,FADN \cap FSS}$ is the number of times unit j is included in the M samples.

We notice that FADN and FSS samples are obtained through a stratified sampling design, but with different stratification variables. The final sample, obtained by the concatenation of the two samples, will have a stratification that is composed of the variables related to the FADN and FSS sample designs, and within these strata, the inclusion probabilities will take the same value (however the number of units in the strata is random).

A final comment on the use of empirical inclusion probabilities is about their impact on the estimates. As analysed in Fattorini (2006) in the case of the Horvitz-Thompson estimator, the use of empirical inclusion probabilities instead of inclusion probabilities, implies a further source of variability that must be taken into account when computing the reliability of the estimator.

3.3. Incomplete and synthetic two way stratification

Instead of creating a unique sample by the union of FADN and FSS, Renssen (1998) suggests to preserve the two different samples, but harmonise their statistical content by calibrating the two systems of survey weights. The result is the creation of two samples that are able to estimate consistently totals of the common information, and to obtain estimates of the parameters of interest by combining estimates from the two samples. When there is the presence of a third file containing complete observations on the variables of interest (the C file) also the weights of the units in this file should be appropriately modified. Note that Renssen considers C as an additional (autonomous and independent) sample survey, while in our context C is the intersection of the FADN and FSS surveys. Hence some slight modifications of the Renssen approach will be considered. From now on, we will consider only continuous variables. the case of categorical values is discussed in detail in Renssen (1998).

At first, some preliminary steps are necessary. Split the set of common variables X in two subsets: the common variables for which population totals are known (V) and common variables for which population totals are unknown (U). Furthermore, let $\omega_a^{(FADN)}$, $a = 1, \ldots, n_{FADN}$, be the survey weights of the units in the FADN sample, and $\omega_b^{(FSS)}$, $b = 1, \ldots, n_{FSS}$, be the survey weights of the units in the FSS sample.

1. The FADN and FSS survey weights are calibrated a first time to the new weights $\omega_a^{1,(FADN)}$ and $\omega_b^{1,(FSS)}$ that should fulfill the constraints

$$\sum_a \omega_a^{1,(FADN)} v_a = \sum_b \omega_b^{1,(FSS)} v_b = T_V$$

$$\sum_a \omega_a^{1,(FADN)} = \sum_b \omega_b^{1,(FSS)} = N$$

   where T_V is the (known) population total for V, and N is the population size.
2. Compute preliminary estimates for the U totals according to the previously computed survey weights:

$$\widehat{T}_U^{FADN} = \sum_a \omega_a^{1,(FADN)} u_a, \qquad \widehat{T}_U^{FSS} = \sum_b \omega_b^{1,(FSS)} u_b.$$

3. Pool together the two estimates $\widehat{T}_U^{FADN}$ and $\widehat{T}_U^{FSS}$ by a linear combination:

$$\widehat{T}_U = \alpha \widehat{T}_U^{FADN} + (1-\alpha)\widehat{T}_U^{FSS}, \qquad 0 < \alpha < 1.$$

4. Calibrate $\omega_a^{1,(FADN)}$ and $\omega_b^{1,(FSS)}$ to new weights $\omega_a^{2,(FADN)}$ and $\omega_b^{2,(FSS)}$ according to the constraints:

$$\sum_a \omega_a^{2,(FADN)} = \sum_b \omega_b^{2,(FSS)} = N$$

$$\sum_a \omega_a^{2,(FADN)} v_a = \sum_b \omega_b^{2,(FSS)} v_b = T_V \tag{3}$$

$$\sum_a \omega_a^{2,(FADN)} u_a = \sum_b \omega_b^{2,(FSS)} u_b = \widehat{T}_U. \tag{4}$$

5. The calibrated survey weights $\omega_a^{2,(FADN)}$ and $\omega_b^{2,(FSS)}$ can be used for estimating any parameter of interest from the FADN and FSS surveys under the conditional independence assumption. For instance, the regression coefficients of Y (Z) on U, V can be estimated on the FADN (FSS), while estimates on the common variables can be estimated in either one of the two surveys.
6. Alternatively one can start from the weights of the concatenated file for calibration of the common variables X and U along the lines suggested above.

When auxiliary information in terms of an additional complete file C is available, the calibration approach should be applied also on the survey weights associated with the units in C. The units in the intersection between FADN and FSS are associated with both the survey weights $\omega_a^{(FADN)}$ and $\omega_b^{(FSS)}$. We suggest to consider the average of these weights:

$$\omega_i = \frac{1}{2}(\omega_i^{(FADN)} + \omega_i^{(FSS)}), \qquad i \in \mathcal{S}^{(FADN)} \cap \mathcal{S}^{(FSS)}.$$

Sometimes these preliminary weights do not work properly, because the calibration algorithm does not converge. A suggestion is to use weights (even constant weights) such that the sum of the weights of the units in the intersection of the FADN and FSS surveys is equal to N;

The calibration approach for the weights in this file is different, according to the procedure taken into consideration.

Incomplete two-way stratification

This approach consists in estimating the parameters of interest on the joint distribution of Y and Z only on C. To this purpose, Renssen suggests to calibrate the weights in C so that the totals of the estimated totals of Y and Z are reproduced:

$$\sum_{i\in\mathcal{S}^{(FADN)}\cap\mathcal{S}^{(FSS)}} \omega_i^3 y_i = \sum_{a\in\mathcal{S}^{(FADN)}} \omega_a^{2,(FADN)} y_a, \tag{5}$$

$$\sum_{i\in\mathcal{S}^{(FADN)}\cap\mathcal{S}^{(FSS)}} \omega_i^3 z_i = \sum_{b\in\mathcal{S}^{(FSS)}} \omega_b^{2,(FSS)} z_b, \tag{6}$$

as well as

$$\sum_{i\in\mathcal{S}^{(FADN)}\cap\mathcal{S}^{(FSS)}} \omega_i^3 = N. \tag{7}$$

Note that it is not important to calibrate these weights with respect to the common variables, because they are not used in the computations. For instance, the correlation coefficient of Y and Z estimator is:

$$\rho_{YZ} = \frac{\sum_i \omega_i^3 (y_i - \bar{y}^{(FADN\cap FSS)})(z_i - \bar{z}^{(FADN\cap FSS)})}{\sqrt{\sum_a (y_a - \bar{y}^{(FADN)})^2 \sum_b (z_b - \bar{z}^{(FSS)})^2}}$$

where $\bar{y}^{(D)}$ is the estimated average on sample D.

Synthetic two-way stratification

This approach consists in estimating at first the parameter of interest under the conditional independence assumption by means of the weights ω^2 on the FADN and FSS samples. The residual with respect to the conditional independence assumption is estimated on the calibrated file C. This residual should be in explicit form. Renssen (1998) gave the result when Y and Z are categorical.

In this case, computations usually involve also the common variables X. Hence, the calibration approach should include constraints (5), (6), (7), as well as an additional constraint. In order to construct this constraint, at first estimate the regression parameters of Y on the common variables $X = (U, V)$ on the FADN, call this vector $\hat{\beta}_{FADN}$, and the regression parameters of Z on the common variables $X = (U, V)$ on the FSS, call this vector $\hat{\beta}_{FSS}$. Let

$$\Phi = \gamma \sum_a \omega_a^{2,(FADN)} x_a x_a^t + (1-\gamma) \sum_b \omega_b^{2,(FSS)} x_b x_b^t.$$

Then, the additional constraint to consider is

$$\sum_{i \in \mathcal{S}^{(FADN)} \cap \mathcal{S}^{(FSS)}} \omega_i^3 y_i z_i^t = \\ = \hat{\beta}_{FADN}^t \Phi \hat{\beta}_{FSS} + \sum_{i \in \mathcal{S}^{(FADN)} \cap \mathcal{S}^{(FSS)}} \omega_i^3 (y_i - \hat{\beta}_{FADN}^t x_i)(z_i - \hat{\beta}_{FSS}^t x_i)^t.$$

3.4. Analysis of uncertainty

The use of a third completely observed file C relies on an assumption: C is representative of the population of interest. If this is not the case, the only conclusions that can be considered in the statistical matching problem consists in assessing which parameters are compatible with the (X, Y) parameters (estimated in A) and the (X, Z) parameters (estimated on B). The statistical literature focuses mainly on the case (X, Y, Z) is a trivariate normal distribution (Kadane, 1978, Rubin, 1986, Moriarity and Scheuren, 2001, 2003, Raessler, 2002). The categorical case has been treated in D'Orazio et al (2006a, 2006b). The result of this approach consists in determining the interval of plausible estimates for the unestimable parameters of the (Y, Z) distribution. The categorical case deals also with the possibility of reducing this interval by introducing suitable constraints on the unobserved variables, such as structural zeros.

According to the procedures outlined in sections 3.2 and 3.3, it is possible to estimate uncertainty in many different ways. One of the possibilities to estimate uncertainty is the following: estimate the marginal distribution of the common variables X on the concatenated file obtained in Section 3.2. Consider the conditional distributions computed from the original data sets A and B.

In this paper we will discuss the case X, Y and Z are categorical.

According to the Fréchet bounds, the joint probability for Y and Z $\theta_{.jk}$ once the marginal distribution on X ($\theta_{i..}$) and the conditional distributions for Y and Z given X ($\theta_{j|i}$ and $\theta_{k|i}$ respectively) are known (estimated), lie in the following interval:

$$\sum_i \theta_{i..} \max\{0; \theta_{j|i} + \theta_{k|i} - 1\} \leq \theta_{.jk} \leq \sum_i \theta_{i..} \min\{\theta_{j|i}; \theta_{k|i}\}.$$

This is a narrower interval than the one we would have obtained by means of the marginal distributions on Y and Z respectively:

$$\max\{0; \theta_{.j.} + \theta_{..k} - 1\} \leq \theta_{.jk} \leq \min\{\theta_{.j.}; \theta_{..k}\}.$$

This results from the fact that in the former interval we have exploited information on the common variables X.

4. Some results on statistical matching of FADN and FSS samples

First the concatenated file for the FADN and FSS samples is obtained. Sampling weights are computed by using te Monte Carlo approach described in Rubin (1986). 3000 samples have been selected according to sampling design of FSS and FADN repectively and for each unit the probability of being included inboth the samples is estimated.

We obtained a single archive comprising Furthermore, we have considered the categorization of the following variables:

- Utilized agricultural area in hectares (X_1), with categories:
  $X_1 = 1$ [0-1), $X_1 = 2$ [1-2), $X_1 = 3$ [2-3), $X_1 = 4$ [3-5), $X_1 = 5$ [5-10), $X_1 = 6$ [10-20), $X_1 = 7$ [20-30), $X_1 = 8$ [30-50), $X_1 = 9$ [50-100), $X_1 = 10$ (more than 100)
- European size units (X_2), with categories:
  $X_2 = 1$ [0-1), $X_2 = 2$ [1-2), $X_2 = 3$ [2-4), $X_2 = 4$ [4-6), $X_2 = 5$ [6-8), $X_2 = 6$ [8-12), $X_2 = 7$ [12-16), $X_2 = 8$ [16-40), $X_2 = 9$ [40-100), $X_2 = 10$ [100-250), $X_2 = 11$ (more than 250)
- Livestock unit coefficient (X_3) with categories:
  $X_3 = 1$ (0, $X_3 = 2$ [1-2), $X_3 = 3$ [2-16), $X_3 = 4$ [16-40), $X_3 = 5$ [40-100), $X_3 = 6$ [100-500), $X_3 = 7$ (more than 500)
- Total number of cattle (Y) with categories:
  $Y = 1$ (one or more bovines) and $Y = 2$ (no cattle)
- Intermediate consumption (Z) with categories:
  $Z = 1$ (up to 4999), $Z = 2$ (5000-24999), $Z = 3$ (25000-99999), $Z = 4$ (100000-499999), $Z = 5$ (over 500000).

The intervals of uncertainty (i.e. the values that can be plausible according to the estimates of the marginal distribution of X_1, X_2 and X_3 in the concatenated file, and of the conditional distributions of Y and Z given the common variables on the respective files) are represented in Table 1.

As a matter of fact, the intervals in Table 1 are extremely informative, because they are very narrow. In some cases the difference between the lower and upper bounds are at the third decimal point. The informative power of the common variables is high, and can be seen comparing the intervals in Table 1 with those we obtain knwign only the marginal distributions for Y and Z (Table 2).

Looking at Table 3 it is possible to see that the uncertainty intervals when information on the common variables is included are one fourth the width of the intervals that do not take information on X into account.

As expected, the estimates of the joint distribution for Y and Z under the conditional independence assumption are inside the uncertainty intervals described in Table 1. These estimates are reported in Table 4.

Table 1: *Lower ($\underline{\theta}_{.jk}$) and upper ($\overline{\theta}_{.jk}$) bounds for $\theta_{.jk}$ when $\theta_{i..}$, $\theta_{j|i}$ and $\theta_{k|i}$ are known*

	$Y = 1$		$Y = 2$	
	$\underline{\theta}_{.1k}$	$\overline{\theta}_{.1k}$	$\underline{\theta}_{.2k}$	$\overline{\theta}_{.1k}$
$Z = 1$	0.02959	0.04903	0.75830	0.77774
$Z = 2$	0.02302	0.04686	0.10060	0.12444
$Z = 3$	0.00715	0.01511	0.02037	0.02833
$Z = 4$	0.00183	0.00420	0.00329	0.00566
$Z = 5$	0.00018	0.00063	0.00035	0.00080

Table 2: *Lower ($\underline{\theta}_{.jk}$) and upper ($\overline{\theta}_{.jk}$) bounds for $\theta_{.jk}$ when $\theta_{.j.}$ and $\theta_{..k}$ are known*

	$Y = 1$		$Y = 2$	
	$\underline{\theta}_{.1k}$	$\overline{\theta}_{.1k}$	$\underline{\theta}_{.2k}$	$\overline{\theta}_{.1k}$
$Z = 1$	0.00000	0.08861	0.71869	0.80730
$Z = 2$	0.00000	0.08861	0.05935	0.14797
$Z = 3$	0.00000	0.03650	0.00000	0.03650
$Z = 4$	0.00000	0.00729	0.00000	0.00729
$Z = 5$	0.00000	0.00093	0.00000	0.00093

Table 3: *Width of the uncertainty intervals $\overline{\theta}_{.jk} - \underline{\theta}_{.jk}$ for the two cases of Table 1 and Table 3*

	case Table 1	case Table 2
$Z = 1$	0.0194	0.0886
$Z = 2$	0.0238	0.0886
$Z = 3$	0.0080	0.0365
$Z = 4$	0.0024	0.0073
$Z = 5$	0.0004	0.0009

Table 4: *Estimates of $\theta_{.jk}$ under the conditional independence assumption*

	$\theta_{.1k}$	$\theta_{.2k}$
$Z = 1$	0.0403	0.7670
$Z = 2$	0.0341	0.1133
$Z = 3$	0.0119	0.0236
$Z = 4$	0.0032	0.0043
$Z = 4$	0.0004	0.0005

References

D'Orazio M., Di Zio M. Scanu M. (2006a) *Statistical Matching: Theory and Practice*, Wiley, New York.

D'Orazio M., Di Zio M., Scanu M. (2006b) Statistical matching for categorical data: displaying uncertainty and using logical constraints, *Journal of Official Statistics*, 22, 137–157.

Fattorini L. (2006) Applying the Horvitz-Thompson criterion in complex designs: a computer-intensive perspective for estimating inclusion probabilities, *Biometrika*, 93, 269–278.

Kadane J.B. (1978) Some statistical problems in merging data files. In Department of Treasury, Compendium of Tax Research, pp. 159–179. Washington, DC: US Government Printing Office. Reprinted in 2001: *Journal of Official Statistics*, 17, 423–433.

Moriarity C., Scheuren F. (2001) Statistical matching: a paradigm for assessing the uncertainty in the procedure, *Journal of Official Statistics*, 17, 407–422.

Moriarity C., Scheuren F. (2003) A note on Rubin's statistical matching using file concatenation with adjusted weights and multiple imputation, *Journal of Business and Economic Statistics*, 21, 65–73.

Raessler, S. (2002) *Statistical Matching: A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches*, New York: Springer-Verlag.

Renssen R.H. (1998) Use of statistical matching techniques in calibration estimation, *Survey Methodology*, 24, 171–183.

Rubin D. (1986) Statistical matching using file concatenation with adjusted weights and multiple imputations, *Journal of Business and Economic Statistics*, 4, 87–94.

Combining sources: a reprise

Kevin Moore, Gary Brown, Tullio Buccellato
Office for National Statistics, Cardiff Road, Newport, Wales, UK
e-mail: Kevin.Moore@ons.gsi.gov.uk

Abstract: A researcher is faced with two data sources to answer their question of interest. Their survey data are expensive, imprecise but timely. Their administrative data are cheap, precise, but untimely. To combine their two sources, they simply extrapolate the administrative source forwards, and hope to reap the benefits. However, simple extrapolation will rarely be the best option for forecasting future values in a time series. Additionally, by visualising a time series as a sample realization from an infinite population, it is clear that forecasting will decrease the precision of the administrative source and reduce the benefits from the combination of the two sources. What should the researcher do? This paper presents generic scenarios for combining sources, and weighs the impact of increasing the standard error of the administrative data against its added benefits.

Keywords: Forecasting, Administrative Data, Error

1. Introduction

1.1 Background

The use of administrative data to augment or even replace survey data in sampling designs or estimation methodologies is a key area of research in National Statistical Institutes. Significant resource has been devoted to discovering how best to reap the benefits from the rich mine of administrative information. There are a wide range of methods: from simply using the data “as is” to form sample strata, to more complicated modelling techniques for improving the quality of estimates for publication. However, are the benefits from including administrative data as real as they seem? It is unlikely the definition of the administrative source is identical to the survey source, nor the timeliness, nor the frequency (not to mention the coverage), so definitional and/or chronological alignment is needed – and the estimation involved ultimately leads to the inclusion of more error in an administrative series that was once only subject to non-sampling error.

The UK Office for National Statistics (ONS) is considering the use of administrative data when any of its surveys is redesigned. The importance of this has increased due to three contributing factors:

- a review of economic statistics in the UK (Allsopp 2004)
- the increasing demand for more detailed statistics
- ongoing efficiencies required to be found within the ONS

1.2 Aim of paper

This paper is aimed at participants of the CENEX-ISAD workshop "Combination of surveys and administrative data", Vienna, 29-30 May 2008.

We reprise the costs and limitations that alignment with survey data imposes on administrative data. To do this we present generic scenarios for including administrative data in survey estimates. The scenarios describe how the administrative data need to be transformed to enable combination with the survey data. Theoretical costs and generic optimal solutions for each scenario are provided. The scope is limited to aggregate data sources – ie data matching is not considered.

Although this paper includes some empirical investigation of a single scenario, the aim is not to provide perfect solution in each case. Rather, the paper aims to provide a springboard for discussion amongst workshop participants.

2. Scenarios for combining administrative and survey data

This section outlines the possible scenarios for combining an administrate source with results obtained from a survey. The aspects of the administrative data, relative to the survey data, that define the possibilities are as follows:

- timeliness – as timely or less timely
- frequency – as frequent or less frequent
- definition (including coverage) – same definition or different definition

As each of these aspects is two dimensional, there are eight possible scenarios – however these generalize to five: two for forecasting, two for temporal disaggregation (ie interpolation), and one where only the combination of sources is required. After introducing some notation, the generic scenarios are outlined.

2.1 Notation

- Survey estimate at time t: Y_t (assumed unbiased)
- Administrative return at time t: Z_t
- Composite estimate at time t: C_t
- Alignment function: F (where $E[F(Z)_t] = Y_t$)
- Forecast function: G
- Temporal disaggregation function: H
- Shrinkage parameter: a
- Time periods where administrative source is not produced or unavailable: m
- Mean squared error: mse
- Standard error: se
- Variance: var
- Covariance: cov
- Correlation: corr

2.2 Generic scenario: forecasting

When the administrative source is not available at time t, ie it is less timely than the survey source or it is as timely but not as frequent (and t is a non-published time point), it needs to be forecast (extrapolated) to be able to be combined with the survey source. The forecast function is defined generically as G – in practice the function would be chosen that minimised the mse of the composite estimate. (If the estimate is unbiased, or at least assumed unbiased, this equates to minimising the se.)

The generic form for the composite estimator with a forecast administrative source is

$$C_t = aY_t + (1-a)G\left[F(Z)_t\right]$$

with

$$\operatorname{var}(C_t) = a^2\operatorname{var}(Y_t) + (1-a)^2 \operatorname{var}\left\{G\left[F(Z)_t\right]\right\} + 2a(1-a)\operatorname{cov}\left(Y_t, G\left[F(Z)_t\right]\right)$$

where

$$a = \operatorname{mse}\left\{G\left[F(Z)_t\right]\right\} / \left(\operatorname{mse}(Y_t) + \operatorname{mse}\left\{G\left[F(Z)_t\right]\right\}\right)$$

and the alignment function F is simply the identity function I( ) if the two sources have the same definition.

Although not explicitly considered, backcasting (when the administrative source didn't exist at all beyond a certain date in the past) is simply a special case of the forecasting equations above, with t replaced by t-n.

2.3 Generic scenario: Temporal disaggregation

When the administrative source is available regularly in the past, but not at t-m, ie it is less frequent than the survey source or is as frequent but less timely, the series needs to be temporally disaggregated. For example, the survey series is available monthly, but the administrative series is only available quarterly. The temporal disaggregation function is defined here generically as H – in practice the function would be chosen that minimised the mse of the composite estimate. (If the estimate is unbiased, or at least assumed unbiased, this equates to minimising the se.)

The generic form of the composite estimator with a disaggregated administrative source is

$$C_{t-m} = aY_{t-m} + (1-a)H\left[F(Z)_{t-m}\right]$$

with

$$\operatorname{var}(C_{t-m}) = a^2\operatorname{var}(Y_{t-m}) + (1-a)^2 \operatorname{var}\left\{H\left[F(Z)_{t-m}\right]\right\} + 2a(1-a)\operatorname{cov}\left(Y_{t-m}, H\left[F(Z)_{t-m}\right]\right)$$

where

$$a = \operatorname{mse}\left\{H\left[F(Z)_{t-m}\right]\right\} / \left(\operatorname{mse}(Y_{t-m}) + \operatorname{mse}\left\{H\left[F(Z)_{t-m}\right]\right\}\right)$$

and the alignment function F is simply the identity I( ) if the two sources have the same definition.

2.4 Generic scenario: Series combination

When the administrative source is as timely and as frequent, neither forecasting nor temporal disaggregation is required. If the two sources have the same definition, the survey source should be discarded and the administrative source used in its entirety. However, if the definitions are different, a function of the administrative data needs to be derived that has the same definition. The alignment function is defined here generically as F – in practice the function would be chosen that minimised the mse of the composite estimate. (If the estimate is unbiased, or at least assumed unbiased, this equates to minimising the se.)

The generic form for the composite estimator with an aligned administrative source is

$$C_t = aY_t + (1-a)F(Z)_t$$

with

$$\operatorname{var}(C_t) = a^2\operatorname{var}(Y_t) + (1-a)^2 \operatorname{var}\left[\mathrm{F}(\mathrm{Z})_t\right] + 2a(1-a)\operatorname{cov}\left(Y_t, F(Z)_t\right)$$

where

$$a = \operatorname{mse}\left\{F(Z)_t\right\} / \left(\operatorname{mse}(Y_t) + \operatorname{mse}\left\{F(Z)_t\right\}\right)$$

3. Optimisation of combined data

Now that the required alignment adjustments between survey and administrative data have been made, we can ask the question: "How useful is your administrative source now?" The answer lies in the shrinkage parameter "a".

- If a = 0, then the administrative source has no mse, and so all the weight in the composite estimator should be given to the administrative data. This is exactly what happens with a = 0: the survey source is discarded.
- If a = 1, then the administrative source has infinite standard error, and so all the weight in the composite estimator should be given to the survey data. This is exactly what happens with a = 1: the administrative source is discarded.

The further that "a" moves away from 0, the less benefit there is from combining the administrative data with the survey data. An interesting question is which source is the more important in the composite estimator – as expected; it is the one with the smallest mse. Hence, the smaller the mse of the transformed administrative source, the more benefit it will bring.

1. If we assume the alignment function F is fixed for each administrative source, it cannot be optimized in terms of mse.
2. The only optimization is thus possible in terms of the forecasting function G (or interpolation function H).

Determining the mse of forecasting/temporal disaggregation techniques, and minimizing them, will optimize the benefits from combining administrative with survey data. However, determining the mse of forecasts is only possible once a researcher knows the point in the series they are forecasting. Furthermore, once we have the forecast mse, it is very difficult to assess whether this is an accurate measure of the mse. It is possible that the variance of the error of a forecast can even be infinite under certain conditions, Sweet (1985).

4. Empirical testing using ONS data

This section uses aggregated survey data from short-term employment surveys, collected by the ONS, and from administrative data collected by Her Majesty's Revenue and Customs (HMRC). Both sources measure jobs in Great Britain. The surveys collect the number of jobs undertaken by employees of British companies, the data from HMRC detail the number of jobs performed by employees (who earn above a certain threshold wage). The short-term surveys collect data every month. However, the HMRC series is quarterly. Therefore this problem is one of temporal disaggregation of the HMRC data in the months between the quarterly points, and of forecasting the endpoints.

The interpolation can be performed using many statistical packages. For this paper we used PROC EXPAND in SAS to turn our quarterly administrative series into a monthly one. This procedure produces a spline of the required order polynomial. We ensured that the resulting spline passed through the actual quarterly points in the administrative series by fixing them as knots. However, although the interpolated points along a spline should have error associated with them, there are no quality measures provided with spline coefficients from this package. The lack of quality measures meant that our scenarios could not be used directly in this case. It is part of further work in this study to determine how to calculate such errors. There are alternative interpolation methods such as kriging that provide an associated variance estimate.

For the purposes of this paper, instead of disaggregating the administrative series we assume that the survey series was quarterly, and consider only forecasting techniques. We used over five years of quarterly data, both administrative and survey. To provide a simple test, we dropped the most recent 1, 2 and 3 quarters of the administrative series, and then tried to forecast them to simulate untimeliness.

In doing this, we then computed forecasts using three different techniques. As we have the actual administrative returns that we're forecasting, we could compare the error the techniques produced. The first method employed was by producing a simple forecast using only the information found in the administrative series via PROC FORECAST in SAS, from now on called the univariate procedure. This SAS procedure uses a number of forecasting techniques. However, we simply used the default settings which meant that the procedure used a stepwise autoregressive method.

In an attempt to improve on the forecasting power we tried to add to the information from the administrative series by using what we had in the survey series. In our example, the survey source is simulated as being timelier and therefore we may be able to use information from that series in our forecast of the administrative data.

We start with a very simple approach considering the weight as the correlation between the two series along the span of time observed for both. Using the correlation between two series in order to interpolate one of them is not a new idea, see Friedman (1962). However, this procedure implies the adjustment of the predicted rate of growth of the univariate series by the observed rate of growth of the survey series. We can write this growth rate, R, as a linear combination of the growth rates of the survey and (forecast) administrative series ie

$$R_Y = \left(Y_t - Y_{t-1}\right) / Y_{t-1} \qquad R_{G(Z)} = \left(G\left(Z\right)_t - Z_{t-1}\right) / Z_{t-1}$$

then

$$R = \rho R_Y + \left(1-\rho\right) R_{G(Z)}$$

where

$$\rho = \operatorname{corr}\left(Y_t, G\left(Z\right)_t\right)$$

Once we have obtained estimates for the combined rate of growth we can use it to produce forecasts that borrow strength from the survey source:

$$L\left[G\left(Z\right),Y\right]_t = RG\left(Z\right)_t$$

Our third method incorporates the measure of the degree of quality of the estimate in the survey source - the coefficient of variation (cv) – which we want to include in our combined forecast for the administrative source. Intuitively, we want to add a correction factor for the amount of information derived from the survey source, which decreases as the coefficient of variation of the series increases. To do this, we can amend the correlation between the two series, correcting it for the coefficient of variation in the survey source using the following adjusted correlation measure.

$$\rho' = \left(1-\mathrm{cv}\right)\rho$$

We then replace the standard correlation measure ρ with its coefficient of variation adjusted measure ρ' in R as defined above. We recognise that this method may only be used with series that are of reasonable quality to ensure that this adjusted correlation value doesn't become too small or even negative. However, if you have a survey source with cv $\gtrsim$ 1, it contains very little useful information anyway. The data used in our case study are national estimates of jobs and all have a cv of 10% or less.

Table 1 displays results obtained through the different forecast methodologies described above using quarterly data for both the administrative and survey series from quarter 1 2001 until quarter 1 2007. We report results obtained with three different procedures to assess and compare the different forecast methodologies that were implemented. The methodology which appears to provide the best forecasts in this case is that using both the correlation between the administrative and the survey series and the coefficient of variation associated with the survey source. The forecasts provided through this measure are a slight improvement on those using only the correlation. However there is a marked improvement compared to forecasting using the univariate procedure alone.

Table 1: *Comparing results through different estimated forecasts*.

	Observed values of the admin series	**Univariate method**	**Combined (correlation)**	**Combined (correlation and CV)**
Forecast t + 1	25,470,535	24,912,657	25,350,158	25,348,434
Forecast t + 2	25,969,942	26,152,936	25,825,497	25,826,867
Forecast t + 3	26,352,906	26,379,261	26,220,383	26,221,038
FMSE		**1.15E+11**	**1.76E+10**	**1.76E+10**
FMAE		**255,742**	**132,448**	**132,348**
FRMSE		**339,318**	**132,812**	**132,625**

FMSE, FMAE and FRMSE are the forecast mean squared error, forecast mean absolute error and the forecast root mean squared error respectively, from Kendall and Ord (1990), which we can take as the indicative of the standard error of our forecast function – the eventual aim of such analysis.

5. Concluding remarks

5.1 A discussion of the basics

The actions within each scenario are those that are intuitively obvious, to a seasoned researcher at least. However, there are pitfalls in the methods that should not be overlooked. With the use of any standard statistical computer package it is easy to forecast or temporally disaggregate points of a series. However, the mechanism of performing these tasks provides an answer that is subject to error. Indeed, since the series that you are forecasting or temporally disaggregating often includes error itself, any point in the series is subject to variation and hence so are the points you are calculating.

Even the choice of method a researcher uses to forecast or temporally disaggregate will introduce some variation in the results. It was the purpose of this paper to provide a background for discussion of the effect the different choices that the researcher may make has on the quality of the final combined series.

For instance, assume that we have a quarterly administrative series and a monthly survey series. In this example, the administrative series will need to be temporally disaggregated between quarter months in order to obtain a value to combine with the survey values. How should the researcher do this? Simple linear interpolation can be easy to implement in a production system. Splines, using cubic or other polynomials, or even other methods, can be more complicated. Are there any advantages to using these more complicated methods in terms of quality of the final product ie the combined series? Surely yes.

Similar considerations should be made when we need to forecast the administrative source. For example, is there a distinct advantage of forecasting a series using Holt-Winters' method or ARIMA modelling? Once again, the forecasts from these methods have associated variances. To use the scenarios as described, combining the two series requires the error of the forecast to be evaluated. The larger the error of the temporal disaggregation or forecast, depending on the situation the researcher is faced with, the larger the weight given to the survey series when combining the two.

5.2 Further discussion

This paper aims to provide a basis for discussion of issues and problems surrounding combining survey and administrative sources – the outcome of which should be beneficial to many organizations. Administrative data can be a rich source of information and, if used correctly, should improve the quality of estimates obtained from surveys alone and potentially improve efficiency in survey design. However, there are issues, such as the simple ones posed in this paper, that need to be addressed before this improvement can be measured. For example, to enable combination of series that are of differing frequency we need to be able to obtain the error of temporally disaggregated points.

We hope this paper provides participants with a useful introduction to the workshop topic of combining survey and administrative data, and is successful in provoking further discussion.

References

Allsopp C (2004) Review of statistics for economic policymaking, Chancellor's report, Her Majestey's Stationery Office.

Kendall M., Ord J. K. (1990) *Time Series,* Edward Arnold.

Sweet A. L (1985) Computing the Variance of the Forecast Error for the Holt-Winters Seasonal Models, *Journal of Forecasting*, Vol 4, 235-243

Friedman M (1962) The interpolation of a time series by related series, *Journal of the American Statistical Association,* Vol 57, No 300 729-757

3

Conceptual aspect for integration

A framework for statistical entity identification to enhance data quality

Michaela Denk
Statistics Austria, Dir. Population Statistics, A-1110 Wien, Guglg. 13
University of Vienna, Dept. of Statistics, A-1010 Wien, Universitaetsstr. 5,
e-mail: michaela.denk@statistik.gv.at, michaela.denk@univie.ac.at

Abstract: Ensuring data quality is a crucial challenge in scientific and statistical database management aiming at an improved usability and reliability of the data. Entity identification deals with matching records from different data sets or within a single data set that represent the same real-world entity and, thus, enables data integration at record level as well as the detection of duplicates. Both can be regarded as a means of improving data quality, the former by completing data sets through adding supplementary variables, replacing missing or invalid values, and appending records for additional real-world entities, the latter by resolving data inconsistencies. This paper suggests a framework for statistical entity identification particularly focusing on probabilistic record linkage with an implementation in R.

Keywords: record linkage, record matching, data integration, deduplication

1. Introduction

Data quality is commonly defined as the ‘fitness for use’ of the data, i.e. data quality can merely be measured relative to its intended usage. Syntactic and semantic correctness of the data, format and value consistency, accuracy, completeness and precision, as well as various temporal aspects, for instance timeliness, are regarded as classical criteria of data quality (e.g. Tayi and Ballou 1998, or Missier et al. 2003). The importance of metadata for data quality is also frequently emphasized (cf. Denk and Froeschl 2000, Denk 2002). Poor data quality is mainly due to missing, invalid (i.e. contradictory or out-of-range) or incorrect entries.

The ongoing trend towards multiple uses of data, in official statistics due to guidelines stipulating the reduction of responder burden, especially requires the combination of data that was never meant to be combined and poses problems of multi-source data quality (Tayi and Ballou 1998, Wang and Madnick 1989). Schema-level and record-level multi-source data quality problems are discerned, viz. heterogeneous data models and schema designs, as well as overlapping, contradicting and inconsistent data, respectively. Schema-level issues include structural conflicts (e.g. incompatible formats) and semantic conflicts (such as homonyms or synonyms), whereas record-level problems involve inconsistencies with respect to measurement units or aggregation levels. One of the most crucial issues is the joining of overlapping data, in particular of records representing the same real-world entity, which requires entity identification (Wang and Madnick 1989, Bilenko et al. 2003).

Besides data integration from multiple sources, the second field of application of entity identification is the detection and elimination of duplicate records in a single data set. Thus, entity identification vitally contributes to data quality improvement (Missier et al. 2003, Cao et al. 2004, Winkler 2004). The explosive growth of available data sources, especially on the WWW, has made entity identification even one of the most important issues in data warehousing where it plays a major role in the ETL process (e.g. Aizawa and Oyama 2005, Bilenko et al. 2003, Cochinwala et al. 2001, Rahm and Do 2000, or Jarke et al. 2000). Hence, it constitutes a crucial preparatory step in data mining projects as well (Dasu and Johnson 2003).

Section 2 provides a brief introduction to entity identification, related quality measures, and conventional approaches. Section 3 discusses a framework for statistical entity identification consisting of a multi-stage model that covers the entire process, including preparatory steps, the selection and comparison of candidate records, the estimation of statistical scoring and classification models that establish the basis for the final decision between 'match' ('duplicate') and 'unmatch' ('distinct entity') as well as the evaluation of these results. Section 4 succinctly introduces the implementation of the framework in R that has already been presented in more detail at the GfKl'07 (Denk 2007a). Finally, section 5 concludes with a short summary and outlook.

2. Background and motivation

Entity identification is defined as the detection and merging of two or more records representing the same real-world identity in a single data set or a collection of data sets, which is relevant in duplicate detection and elimination as well as data integration. Entity identification is also known as object identification, instance identification, object consolidation, entity matching, entity reconciliation, entity resolution, record matching, record linkage, data reconciliation, merge/purge problem (prevalent particularly in business contexts), or instance integration. However, it is defined slightly deviating in several contexts, for instance only referring to the single or the multiple database problem, respectively. Related notions merely corresponding to subprocesses of entity integration are field matching, approximate matching, approximate string join, string matching, name matching or clustering, and the key equivalence problem. Cf. for example Lim et al. 1996, Galhardas et al. 2000, Wang and Madnick 1989, Elfeky, Verykios, and Elmagarmid 2002, Fellegi and Sunter 1969, Aizawa and Oyama 2005, Missier et al. 2003, Hernandez and Stolfo 1995 and 1998, Cao et al. 2004, Minton et al. 2005, or Pu 1991. Apart from data cleaning, data integration and data warehousing, entity identification is closely related to information retrieval, pattern recognition and classification, machine learning, and data mining as well, thus, making use of ideas from several research areas (e.g. Bilenko et al. 2003).

In the data integration scenario, there are two data sets A and B with records a_i, $i=1,\ldots,I$, and b_j, $j=1,\ldots,J$, representing real-world entities ω_k, $k=1,\ldots,K$. The set of record pairs A x $B = \{(a_i,b_j); a_i \in A; b_j \in B\}$ is a union of the two disjoint sets of true matches M (true duplicates) and true non-matches (true distinct entities) U (Fellegi and Sunter, 1969). M contains all record pairs representing identical real-world entities ω_k (representations may differ) and U all pairs that represent distinct real-world entities. In the deduplication scenario, only one data set is processed; i.e. A and B are identical, and, thus, M contains at least I record pairs, viz. the record pairs containing the same record twice. M and U are defined as follows:

$$M = \{(a_i,b_j);\ a_i=r_A(\omega_k);\ b_j = r_B(\omega_k);\ a_i\in A;\ b_j\in B\} \qquad (1)$$
$$U = \{(a_i,b_j);\ a_i=r_A(\omega_k);\ b_j=r_B(\omega_p);\ \omega_k\neq\omega_p;\ a_i\in A;\ b_j\in B\} \qquad (2)$$

The entity identification process aims at finding a classification rule assigning each record pair to the set of links (*L*, identical entities, or duplicates in the deduplication scenario) or the set of non-links (*NL*, distinct entities), respectively. *M* and *U* are defined by the true matching status of record pairs, whereas *L* and *NL* are based on the estimated linkage (or duplication) status. Frequently, a third class *P* is introduced containing undetermined record pairs (possible links/duplicates) for which the final linkage status can only be set by using supplementary information (usually obtained via clerical review). Comparison vectors are determined consisting of the measured similarities of a record pair with respect to the matching variables available in both data sets. Based on the comparison vectors classification rules are specified. In statistical entity identification, matching scores for each record pair are estimated first, and thresholds are then determined to establish a classification rule, often based on pre-specified error levels.

2.1. Data quality

Apart from the assessment of the quality of source data, which plays an important role in the entity identification process, measures of the variability, reliability and accuracy of the results of entity identification as well as measures for the quality of specific stages of the process (see section 3.6 below) are required to evaluate the quality of the resulting data set.

Concerning the quality of matching variables, primarily completeness, reliability, and discriminating power are relevant. Suitable measures are provided by Cochinwala et al. (2001), Newcombe et al. (1959), and Jamieson, Roberts, and Browne (1995). For the assessment of the quality of the entire entity identification process, misclassifications and correct classifications as well as the number of possible links are of particular interest. The accuracy of the estimation of error rates mainly depends on the availability of training data with known true matching status. Essentially, two types of misclassification error are discerned, viz. (i) the false non-link *FN* (type I error) corresponding to the failure to link records that represent the same real-world entity; and (ii) the false link *FL* (type II error) corresponding to the linking of records that represent different real-world entities. Error frequencies are typically presented in a confusion matrix (e.g. Missier et al. 2003, Christen and Goiser 2005) as depicted in Table 1 for an entity identification scenario without possible links.

Table 1: *Confusion matrix*

estimated link status true matching status	*L*	*NL*
M	$\lvert TL \rvert$	$\lvert FN \rvert$
U	$\lvert FL \rvert$	$\lvert TN \rvert$

Based on the elements of the confusion matrix, different quality measures for the entity identification process are specified:

1. false match rate (Fellegi and Sunter 1969) or false positive rate (Christen and Goiser 2005) defined as the ratio of the number of false links with respect to all true non-matches, i.e.
   $|FL| / |U| = |FL| / (|FL| + |TN|)$
   Belin and Rubin (1995) define the false match rate with respect to the number of all linked pairs, i.e.
   $|FL| / |L| = |FL| / (|FL| + |TL|)$
2. false non-match rate (Fellegi and Sunter 1969) defined as the ratio of the number of false non-links with respect to all true matches, i.e.
   $|FN| / |M| = |FN| / (|FN| + |TL|)$
   A variant of this false non-match rate can be defined in analogy to the deviating definition of the false match rate by Belin and Rubin (1995), viz.
   $|FN| / |NL| = |FN| / (|FN| + |TN|)$
3. accuracy (Christen and Goiser 2005) defined as the proportion of the number of accurate classifications with respect to the total comparison space (i.e. the number of compared (or classified) record pairs), i.e.
   $(|TL| + |TN|) / (|TL| + |TN| + |FL| + |FN|)$
4. precision (Christen and Goiser 2005, Lee, Ling, and Low 2000) defined as the proportion of the number of accurate links with respect to the total number of links, which is equal to 1 – Belin/Rubin false match rate, i.e.
   $|TL| / |L| = |TL| / (|TL| + |FL|)$
5. recall (Christen and Goiser 2005, Lee, Ling, and Low 2000) or sensitivity (Jamieson, Roberts, and Browne, 1995) defined as the proportion of the number of accurate links with respect to the total number of true matches, which is equal to 1 – Fellegi/Sunter false non-match rate, i.e.
   $|TL| / |M| = |TL| / (|TL| + |FN|)$
6. f-measure (Christen and Goiser 2005) derived from precision and recall in order to take into account the trade-off between these two measures (Lee, Ling, and Low 2000), namely
   *2 * precision * recall / (precision + recall)*
7. specificity (Jamieson, Roberts, and Browne, 1995) defined as the proportion of the number of correctly unlinked pairs with respect to the number of true non-matches, i.e. 1 – Fellegi/Sunter false match rate.

Another important quality criterion is the number of possible links – comparing procedures with equivalent error rates, those minimizing the grey zone of record pairs without a decision about linkage requiring clerical follow-up are preferable (see also Winkler 1985).

Quality measures should neither be presented individually nor for a single threshold value (i.e. for only one particular classification set-up) in order to enable a comprehensive and thorough analysis of the quality of the entity identification process. Plots of quality measures over a range of threshold values as well as precision-recall-plots provide invaluable insights.

The number of correct non-links depends on whether correctly non-linked entities or correctly non-linked record pairs are counted. If the frequency is calculated for record pairs, it will highly dominate all other counts presented in the confusion matrix, potentially resulting in misleading accuracy and false positive rate (Christen and Goiser 2005).

2.2. Conventional approaches

According to Bilenko et al. (2003), entity identification approaches can be categorized by how much human expertise they require and the extent to which they use probabilistic or machine learning methods in order to automate (partially at least) the identification process. One end of the spectrum is occupied by rule-based methods based on equational theory. These methods are also called knowledge-based, as they require human experts who specify equivalence rules for records in a declarative rule language, building the 'knowledge base' for the derivation of whether two records are regarded as equivalent or not. The definition of rules may involve string similarity metrics (see subsection 3.3) as well as domain-specific comparisons (such as equality conditions for nicknames and first names). Examples for rule- or knowledge-base approaches are:

1. the entity join (Kent 1979), where rules have to be specified to establish the 'matching part' of the entity join operator that refers to nomenclatures, simple correspondence tables (or ontologies), and the knowledge base of rules to decide on the equivalence of records;
2. the IntelliClean 'knowledge-based' framework (Lee, Ling, and Low 2000) for effective data cleaning, where again if-then rules have to be specified to build the framework. IntelliClean furthermore requires clerical intervention during validation;
3. ad-hoc compliance or equivalence classes defined via constraints on matching variables and/or comparison outcomes (e.g. Denk and Oropallo 2002);
4. the usage of heuristic rules (Wang and Madnick 1989) from which inferencing algorithms derive additional information on records to find out whether they refer to the same real-world entity;
5. the usage of extra semantic information as well as identity and distinctness rules (i.e. semantic constraints on real-world entities, Lim et al. 1996) to match records that do not share common keys but for which an extended key can be generated from common key variables and additional variables;
6. the BOYS Algorithm (Yesilcay 1993). Based on a training sample with known matching status a sequence of classification rules made up of a selection of matching variables as well as the optimum tolerance levels for the variables that declares record pairs as links or non-links is derived, yielding matching errors within specified limits (Bias of the Order You Specify).

Obviously, rule- or knowledge-based methods offer a high degree of flexibility and tuning possibilities, resulting in good performance for specific applications, yet always at the risk of overfitting. However, such 'systems' are not scalable, development costs are high, and their maintenance is rather laborious (see also Minton 2005).

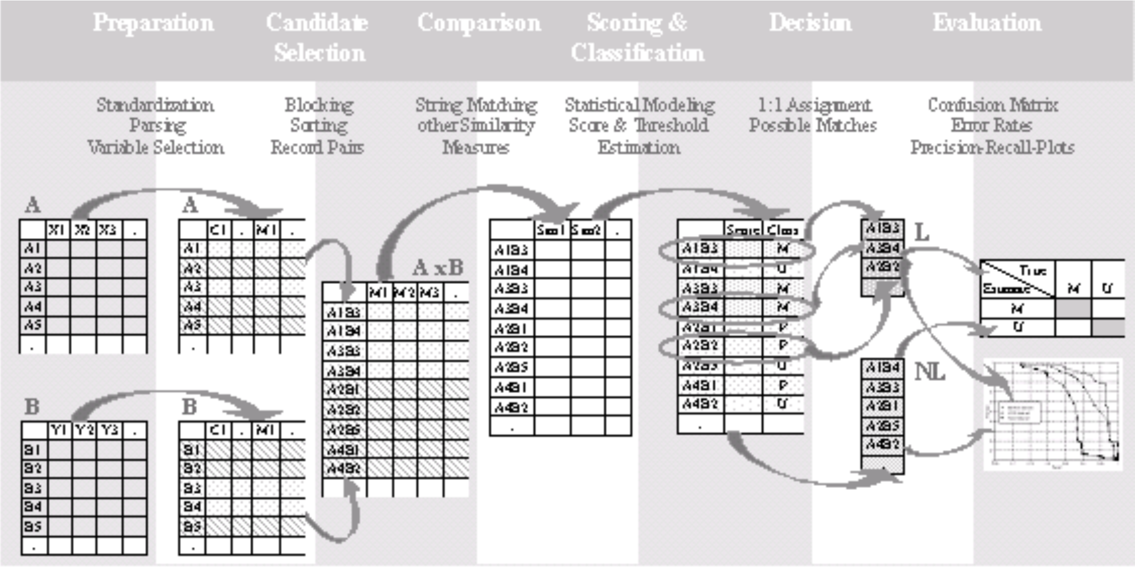
Distance-based approaches are more suitable to automating the entity identification process and less domain-dependent. In the approach of Dey, Sarkar, and De (2002) the weighted sum of the similarities of matching variables between records quantifies the total similarity of records which are then used in a simple assignment model; weights often correspond to the discriminating power of the variable. A comparable approach uses the concatenated values of all matching variables as if they were only one variable and then calculates the similarity between records via simple string matching algorithms (Monge and Elkan 1996). Another example is the AJAX declarative framework that extends SQL to allow the specification of standardization, duplicate elimination, and matching steps (Galhardas et al. 2000). Records with high similarity values with respect to matching criteria are grouped together in clusters which are then collapsed into one record of the resulting data set.

The main shortcomings of most of these approaches are the requirement of domain-specific knowledge coded in equivalence rules and/or the usage of only several of the phases the whole entity identification process is composed of.

3. The SEI framework

It is of vital importance for an entity identification framework to comprise the entire process from data preparation to the evaluation of the results. Conventional entity identification approaches are often limited to searching and matching, sometimes preceded by a pre-processing or preparation phase (e.g. Cochinwala et al. 2001, Missier et al. 2003). For example, an equivalence class approach usually amounts to the first two SEI-phases directly followed by a (frequently clerical) decision phase. Besides, it is rather common to assume that schema-level heterogeneity has been resolved prior to entity identification (which is merely relevant in the data integration scenario). Consequently, the SEI framework does not include schema-level integration, either.
Figure 1 illustrates the multi-phase process model of the suggested statistical entity identification framework (cf. Denk 2006). The fourth phase ('scoring & classification') may be split up into two separate phases ('scoring' and 'classification') as well (cf. Denk 2007a).

Figure 1: *Process model of the SEI framework*



3.1. Preparation

The first phase of entity identification in the SEI framework is the data preparation phase, encompassing different transformations of common variables to obtain comparable variables suitable for usage in the further identification process. In particular, string variables, such as names and addresses have to be pre-processed to be comparable among data sets, but also simple calculations, for example age to be determined from date of birth, can be necessary to derive matching variables. Typically, standardization, parsing and/or phonetic coding are required in case of string variables. Standardization is tantamount to the conversion of values to a consistent format. Parsing deals with the decomposition of a string variable into a common set of components that are better comparable, as for instance splitting up a general address variable into postcode, city, street address and number. Coding eliminates common mistakes, e.g. due to similar-sounding consonants, interchanging of vowels, or errors in repeating letters, and retains only the characteristic part of a string such that like-sounding strings end up as the same code. Widespread coding algorithms are the Russel Soundex Code (Odell and Russel 1918 and 1922), NYSIIS (Lynch and Arends 1977) and ONCA (Oxford Name Compression Algorithm, Gill 1997, 2001). In general, the parsing and standardization of free-form strings in combination with advanced string comparators (cf. section 3.3) is more effective than the use of coding methods. For a more detailed discussion of standardization, parsing and coding see Winkler 1995, Cochinwala et al. 2001, or Rahm and Do 2000.

3.2. Candidate selection

The second phase comprises a collection of methods for a fast and thus computationally cheap filtering of record pairs with a negligible chance of containing records representing the same real-world entity. In general, a detailed comparison with respect to all available matching variables is extremely time-consuming, if accomplishable at all. Especially for large data sets, the selection of candidate record pairs with higher likelihood of belonging to the set of true matches is necessary to reduce the number of pairs that undergo the subsequent detailed comparison of matching variables as input for the scoring and classification step. However, this restriction affects the error levels established for the entire process: the false match rate is reduced, whereas the false non-match rate is increased. An overview of the most recent advances in the selection of candidate pairs is provided by Aizawa and Oyama (2005).

Phases one and two usually also precede conventional entity identification approaches; yet, they are hardly ever integrated in the EI framework. Phase two is sometimes labelled searching phase.

Blocking is the classical approach: the set of all possible record pairs is subdivided into blocks agreeing on a specified blocking key. Only record pairs within these blocks are further analysed, whereas the (usually larger) residual set of pairs are classified as non-links directly. The best blocking variables have a high number of categories, high reliability and low error rates. Variables often used for blocking are regional classifications, e.g. postcode, ‘Soundexed’ string variables (mostly names), or initials. For details see Fellegi and Sunter (1969).

In the sorted neighbourhood method (Hernandez and Stolfo 1995; also discussed by Bertolazzi, De Santis, and Scannapieco 2003 or Missier et al. 2003), records from both data sets are put into one list and then sorted by the blocking variables. A record from the first data set is only compared to its k nearest neighbours in the sorted list of all records. The choice of the parameter k is subject to the available data and may contribute to the success of the blocking strategy.

The bigram indexing method as implemented in the Febrl record linkage system (ANU Data Mining Group 2005) allows for ‘fuzzy’ blocking. The basic idea is that the candidate selection variable values are converted into a list of bigrams and sub-lists of all possible combinations of a particular number of bigrams (smaller than the total number of bigrams) are built. Each bigram list represents one block, which increases the number of candidate record pairs compared to classical blocking.

Canopy clustering forms blocks of records based on those records placed in the same canopy cluster. A canopy cluster is formed by choosing a record at random from a candidate set of records (initially, all records) and then putting in this cluster all the records within a certain loose threshold distance of it. The record chosen at random and any records within a certain tight threshold distance are then removed from the candidate set of records. This method is heavily dependent on the distance/similarity measure used. The algorithm and details can be found in McCallum, Nigam, and Ungar (2000).

A simple ranking algorithm for candidate pair selection considers records as plain text and generates clusters of similar records by applying conventional string comparator metrics (as discussed in the next section).

A strategy to minimize classification errors introduced by candidate selection are multi-pass algorithms: during each pass different blocking/sorting keys are applied to the record pairs that have not been selected in the previous pass(es). Multiple passes help decrease the false nonmatch rate introduced by candidate selection, yet may (slightly) increase the false match rate. The transitive closures of the results of the different passes are combined to give the final set of records to be further compared. According to Baxter, Christen, and Churches (2003), multiple passes improve overall linkage accuracy, though efficient implementation and tuning of multiple blocks and multiple sets of record comparisons can be difficult to achieve.

3.3. Comparison

In phase three, the comparison phase, comparison vectors are determined as agreement or similarity patterns from matching variables for candidate pairs selected in phase two. Similarity measures for various types of variables, including string comparator metrics for variables such as names or addresses, are provided. Simple binary outcomes discerning agreement and disagreement or tolerance limits allowing for 'approximate' agreement of numeric variables, such as age differences of plus or minus one or two years, are possible as well. String comparators (e.g. Gill 2001, Cohen, Ravikumar, and Fienberg 2003) are mappings from a pair of strings to the closed interval [0, 1] measuring the degree of compliance of the compared strings (Winkler 1990). The resultant value is either directly used for the classification of record pairs or for the adjustment of matching scores used in probabilistic record linkage.

An early string comparator is the edit distance. Its basic idea is that any string can be transformed into another string through a sequence of changes via substitutions, deletions, insertions, and possibly reversals. The smallest number of such operations required to change one string into another is a measure of the difference between them. For reasons of comparability, the edit distance is rescaled to the interval [0, 1] and then converted to a string similarity measure. The simplest type of edit distance uses equal costs for all required changes in the strings (i.e. the Damerau-Levenstein (D-L) Metric, Damerau 1964, Levenstein 1966). Apparently, different weighting of different types of changes could be reasonable. For instance, substitution of characters could decrease the comparator value to a larger extent than transposition of characters. For a discussion of several enhancements of the D-L metric, e.g. the Needleman-Wunsch distance or the Smith-Waterman distance, see Hall and Dowling (1980).

Jaro (e.g. Winkler 1985, 1990) introduced a string comparator more straightforward to implement and more closely related to the type of human decisions in comparing strings than the D-L metric. Basically, it accounts for the proportion of common characters in both strings and the number of transpositions that have to be made to create the sequence of common characters of one string from the sequence of common characters of the other string. Several enhancements to the Jaro comparator are available as well. The Winkler enhancement (Winkler and Thibaudeau 1991) gives increased value to agreement on the beginning characters of a string. The McLaughlin enhancement (Porter and Winkler 1997) assigns a constant greater than zero to each disagreeing but similar character, where similar characters might occur because of scanning errors such as '1' versus 'I' or 'l', or keypunch errors such as 'V' versus 'B'. The number of common characters as defined in the original Jaro comparator is increased by the constant for each similar character. The final enhancement due to Lynch and Winkler (Winkler 1994) adjusts the string comparator value if the strings are longer than six characters and more than half the characters beyond the first four agree.

Token-based algorithms measure the similarity of strings via the similarity of tokens (words) contained in the strings. One simple example of a token-based algorithm is the Jaccard similarity (Jaccard 1912). It is defined as the number of common tokens in the two strings divided by the total number of different tokens in both strings. The TF–IDF measure (term frequency multiplied by the inverse document frequency; Salton and McGill 1983) weights agreement on rare terms more heavily than agreement on more common terms.

A very common token-based string comparison method consists in comparing the bigrams that two strings have in common, where a bigram is two consecutive letters of a string. The result of the bigram function is the total number of common bigrams in the two strings divided by the average number of bigrams in the two strings (Porter and Winkler 1997). Other bigram variants use a different denominator: instead of the average number of bigrams, the number of bigrams in the first (or in the second) string is used. Apparently, transposition of characters yields a lower value of the string comparator than single erroneous characters. Yet, it does not make any difference whether errors occur in the beginning or at the end of the string. Bigrams are known to be a very effective, simply programmed means of dealing with minor typographical errors. For instance, Porter and Winkler (1997) and Denk, Hackl, Rainer (2005) have shown empirically that bigrams work well.

3.4. Scoring and classification

In the fourth phase, statistical models are specified and corresponding matching scores that assess the likelihood of a pair belonging to the set of true matches (duplicates) or non-matches (distinct entities), respectively, are estimated from the comparison vectors. Scores are then used to classify record pairs into links (duplicates), non-links (distinct entities), and potentially also possible links (possible duplicates).

3.4.1. Probabilistic record linkage

The SEI framework primarily focuses on probabilistic record linkage as proposed by Fellegi and Sunter (1969) and recent approaches to score estimation. This classical approach is based on the conditional probabilities of observing a particular comparison pattern $\gamma((a_i,b_j))$ given that the considered record pair is a match/duplicate $(a_i,b_j)\in M$ and given the record pair is a non-match $(a_i,b_j)\in U$, i.e. shortly: $m(\gamma)=P(\gamma|M)$ and $u(\gamma)=P(\gamma|U)$. The likelihood ratio $LR(\gamma)=m(\gamma)/u(\gamma)$ or a monotonously increasing transformation thereof, usually a dual or natural logarithm is used as the matching score (originally termed matching weight), i.e. $s(\gamma)=\log(LR(\gamma))=\log(m(\gamma)) - \log(u(\gamma))$. To simplify the estimation of the conditional probabilities a conditional independence assumption is made: the components of the comparison vector are assumed to be mutually statistically independent with respect to each of the conditional distributions, which means that $m(\gamma)$ and $u(\gamma)$ can be calculated as products of the corresponding conditional probabilities of observing a particular comparison outcome in the i-th component (i.e. matching variable) given a match or non-match, respectively. The likelihood ratio is now computed as product of individual likelihood ratios for each component and the composite (or total) matching score (if defined as logarithm of the likelihood ratio) as sum of component (or individual) scores. After estimating the scores, score thresholds are determined for the classification of record pairs into links (duplicates), possible links/duplicates and non-links (distinct entities) based on fixed false match and false non-match error levels. This kind of linkage rule is optimal in the sense that the number of possible links is minimized for fixed error levels. However, this optimality is heavily dependent on the accuracy of the estimates of the conditional probabilities and, thus, on the validity of the conditional independence assumption. If only one threshold is determined separating links from non-links, this classification rule is equivalent to a Bayes test for minimum error.

However, the minimization of error probabilities or of the number of possible matches/duplicates is not always the main target. Subject to the objectives of entity identification, different types of errors affect results more or less seriously and, thus, incur different costs. For instance, when creating or maintaining a register, false matches are hardly acceptable. False non-matches are less serious, as there is a chance that they may be detected and corrected in an updating process later on. Similarly, studies based on comparisons of characteristics of linked pairs require a low false match rate, that is, high confidence in linked pairs being true matches. False non-matches will not affect the findings derived from the linked pairs unless the characteristics under study are distributed differently in the detected matches and the erroneous non-matches. For this reason, often ‘matching to a man’ is sought, i.e. obtaining an authentic data set by deterministic matching and/or manually reviewing all possible links and doubtful linked pairs in order to avoid ‘synthetic’ linked records that do not represent existing real-world entities but rather ‘synthetic’ entities that are merely similar to real-world entities. In coverage evaluation, on the other hand, both types of error affect the results in opposite directions and the desired procedure is one that leads to a balance between both types of error. A cost function can be used to take into account the costs of both error types in the estimation of the matching scores. In the two-class scenario (link, non-link), this is equivalent to a Bayes test for minimum cost. For the three-class scenario (link, non-link, possible link) Verykios, Elmagarmid, and Houstis (1999) introduce a cost-minimizing model based on ideas of Tepping (1968) and Fellegi and Sunter (1969).

In their fundamental paper, Fellegi and Sunter (1969) introduced two ways of estimating the crucial conditional probabilities directly from the datasets being matched without requiring a training data set with known true matching status. The first method uses frequency-based scores. The basic idea is that agreement on rarely occurring values of a variable has more distinguishing power than agreement on commonly occurring values. Usually, agreement on a rare value is also better than the general (non-value-specific) yes/no agreement. Thus, instead of only allowing agreement and disagreement as comparison outcomes, agreement on particular values is taken into account explicitly.

The second method for score estimation is specified for the simple case of only three matching variables with agreement/disagreement configurations. If more than three matching variables are used, it is possible to apply general equation-solving techniques; maximum likelihood based methods such as the Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin 1977) are preferable for reasons of numerical stability (Jaro 1989). The estimation of matching scores via EM algorithm is also possible under less restrictive assumptions, when considering the successive incremental discriminating power of matching variables. Moreover, the combination of frequency-based and simple agreement/disagreement EM-derived matching parameters is feasible (Winkler 2000). Further different extensions of the EM algorithm have been developed, for instance to cope with three classes instead of two, with convex constraints, or with deviations from the conditional independence assumption. For a comparison of the application of different EM-type algorithms for the estimation of matching scores see for example Winkler (1991, 1993, 1995).

After scoring, thresholds must be determined to enable the classification of record pairs. Fellegi and Sunter (1969) provided methods for calculating the thresholds directly from the conditional probabilities. However, experience has shown that these methods are rarely suitable since estimated probabilities usually deviate much too severely from the true underlying probabilities (e.g. Winkler and Thibaudeau 1991) which is primarily due to the failure of the conditional independence assumption. Belin and Rubin (1995) introduced a method for determining thresholds at desired error levels when the distribution of observed scores is viewed as a mixture of scores for matches and non-matches based on a training data set with known true matching status. Strictly speaking, only one threshold dividing the total set of record pairs into links and possible links is computed. The relationship between true matching status and matching score is estimated by discriminant analysis based on the training data set is necessitated. In practice, thresholds are often determined by manually reviewing a set of record pairs that are ordered by decreasing matching score. Based on experience, cut-off scores with rough a priori bounds on the error rates can be determined quite rapidly.

3.4.2. Other statistical approaches

Viewing entity identification as a statistical classification problem, standard unsupervised (no training data required) and supervised (training data necessary) statistical classification methods are a straightforward choice. Clustering techniques, usually k-means clustering, are used to obtain the required number of clusters (typically three) of record pairs based on the comparison vectors. The critical issue is the decision on which cluster represents which link status. The approach proposed by Elfeky, Verykios, and Elmagarmid (2002) is rather plausible and, thus, adopted here. A perfectly matching record pair agreeing with respect to all matching variables is located at the *k*-dimensional 1 point (*k* being the number of matching variables, i.e. the dimension of the comparison vector), a completely disagreeing pair of records is located at the origin (zero point) of the *k*-dimensional comparison space. Hence, the cluster with the nearest centroid to the origin (in terms of a standard distance measure depending on the scale of comparison outcomes, e.g. the Euclidean distance) is regarded as the set of non-matches, whereas the farthest cluster from the origin is assumed to represent the set of matches. The record pairs in the remaining cluster receive matching status undecided. Cao et al. (2004) report on problems of hierarchical algorithms, especially single linkage clustering, in the deduplication scenario and present an approach based on the compact set and the sparse neighbourhood criteria, assuming that distances between duplicates might be larger than distances between non-matching records, although duplicates are usually closer to each other than they are to other distinct records, and that the local neighbourhood of duplicates is usually empty or sparse.

In case of available training data, a classical methodological choice is discriminant analysis. As already stated above, the Belin-Rubin method tries to predict class membership conditional on the matching scores assigned to record pairs. Discriminant analysis based on comparison vectors is also conceivable. Non-parametric methods that are independent of distribution assumptions, such as nearest-neighbour-approaches or classification trees, are preferable (Neiling 1998, Schuermann 1996).

Another probabilistic approach applicable in case of available training data is logistic regression. As independent variables either comparison outcomes or the matching score can be used. Chatterjee and Segev (1992, 1994) and Aizawa and Oyama (2005) suggest two similar approaches to the estimation of matching scores by means of logistic regression models.

3.5. Decision

The decision phase fulfils three different tasks. First, if 1:n or 1:1 assignment of records is the objective of the entity identification process, the m:n assignment resulting from the scoring and classification phase has to be refined to achieve a final classification decision for each record pair. If m:n assignment is sufficient (as usually in duplicate detection), this step is omitted.

In a 1:1 matching situation, the application of appropriate algorithms can dramatically improve matching performance, at least by lowering the number of possible links. For example, non-matches such as husband–wife or brother–sister pairs agreeing on address information (and – probably – on surname) usually receive sufficiently high weights to be designated as possible links. If 1:1 matching is used, these possible links can be automatically identified as non-links in case the true matches are also available in the combination of the two datasets (Winkler 1994).

A greedy algorithm is one that always associates a record with the corresponding available record having the highest matching weight. Subsequent records are only compared to remaining records that have not yet been assigned. There are several variants of greedy algorithms. However, it is known from experience, that greedy algorithms often make erroneous assignments (Jaro 1989, Winkler 1994).

Looking for a one-to-one matching scheme that maximizes the sum of matching scores (or another indicator of compliance) of assigned links, Jaro (1989) introduced a linear sum assignment procedure (LSAP). The original LSAP algorithm was proposed by Burkard and Derigs (1980). In practice, often a mixed approach using a greedy variant combined with experience-based decision rules is applied.

In the second step of the decision phase, undetermined pairs are (usually manually) reviewed to come to a decision on their final estimated link (duplication) status. This step is skipped, if the set of possible links is empty.

In a final step, value conflicts in linked record pairs have to be resolved (Lim et al. 1996). Time stamps, integrity rules, and plausibility checks, as well as additional (meta-) information, such as address registers, or domain-specific ontologies, are typically used.

3.6. Evaluation

Finally, the sixth phase enables the estimation of quality measures to evaluate the entity identification process. Confusion matrices, misclassification rates and other overall quality criteria, including visualizations as presented in section 2.1., are supported as well as phase-specific quality measures as discussed below. Each phase of the entity identification process can be evaluated individually and in comparison to the overall performance of the complete process. These incremental changes can be graphically represented as relative gains, for instance in precision and recall.

Actually, every entity identification process should be concluded by an evaluation phase, as also suggested by Elfeky, Verykios, and Elmagarmid (2002). However, this phase can be rather cost and time intensive, since training data are required to provide sound estimates of quality measures. Training data might come from previous studies, from samples of the current data set for which manual review is carried out, or from other geographical locations. The confusion matrix for the training data set is then used as a basis for the calculation of error estimates.

In their fundamental theory for record linkage, Fellegi and Sunter (1969) provide a simple method of obtaining estimates for error rates from the estimated conditional probabilities. However, Belin's studies of various weighting procedures (1993), amongst others, reached the conclusion that this method tends to be grossly optimistic, due to the often invalid conditional independence assumption. The Belin/Rubin method (1995) presented above for threshold estimation has actually been developed for estimating error rates. In any case, estimated error rates can be used to adjust statistical analyses of the resultant data set for matching error (e.g. Scheuren and Winkler 1993, 1997, Winkler 1999, Winkler and Scheuren 1996).

One opportunity to estimate the variability of matching results is to carry out a sensitivity analysis. For instance, by varying matching weights or weight thresholds in a probabilistic linkage application, the effects on the classification of record pairs may be estimated. Record pairs classified differently when using adjusted parameters should be manually reviewed. Winkler (1985) proposes an approach similar to bootstrapping (Efron 1979) or multiple imputation (Rubin 1987) to estimate the variance of score thresholds and error rates in probabilistic linkage.

Fellegi and Sunter (1969) discuss methods of choosing among alternative blocking procedures by taking into account costs of different errors introduced by blocking. Kelley (1984, 1985) provides further guidance on how to make an objective choice among alternative blocking procedures by weighing the reduced costs of computation against the errors introduced by not looking at all comparison pairs. Referring to Elfeky, Verykios, and Elmagarmid (2002), Baxter, Christen, and Churches (2003) propose three performance metrics for blocking procedures requiring training data. The reduction ratio RR is defined as the relative reduction in the number of record pairs to be compared, i.e., the difference between the number of all possible record pairs and the number of pairs remaining after blocking is divided by the number of all possible pairs. The second quality indicator for blocking is the pairs completeness metric PC which is defined as the ratio of the number of true matching record pairs in the set of record pairs produced for comparison by the blocking procedure and the number of true matches in the entire data. Eventually, they propose a score capturing the tradeoff between pairs completeness and reduction ratio. It is computed as (2x*PC*x*RR*)/(*PC*+*RR*). In the evaluation of the entire process blocking is a confounding factor. If feasible, all quality measurements should be reported without use of blocking or the blocking approach (including procedure and parameters) is published together with the resulting number of removed pairs of records.

The quality of string comparator metrics may be illustrated by visualizing the distributions of different comparators in matches and non-matches. The general measures to evaluate the quality of the overall entity identification process can also be used to evaluate individual phases, such as for instance the string comparison phase. Analogously, if training data are available, different 1:1 assignment algorithms may be compared by looking at the resulting error rates. The quality of the assignment algorithm can be assessed by the amount of reduction of multiple links.

4. Implementation

The implementation of the SEI framework is structured according to the six (or, rather, seven; cf. Denk 2007a) stages of the statistical entity identification process. For each stage there is one component, i.e. one function, that establishes an interface to the lower-level functions which implement the respective methods. The outcome of each stage is a list containing the processed data and protocols of the completed processing stages. Table 2 provides an overview of the functionality of the components and the spectrum of available methods. Methods not yet implemented are italicised. A detailed description of the implementation is provided in Denk (2007a) including a discussion of the functionality of the framework components as well as its illustration by means of a demo sample of a typical CRM dataset.

The preparation component provides an interface to the `phoncode()` function from the StringMatch toolbox (Denk 2007b) as well as the functions `standardise()` and `parse()`. By this means, it phonetically codes, standardizes, or parses the variable(s) in the input data frame according to the specified method(s) and appends the resulting variable(s) with the defined label(s) to the data frame. The default method is American Soundex. At the moment, a selection of popular phonetic coding algorithms and standardization with user-provided dictionaries are implemented, whereas parsing is not yet supported.

The candidate selection component provides an interface to the functions `crossproduct()`, `blocking()`, `sortedneighbour()`, and `stringranking()`. Candidate record pairs from the two input data frames are created and filtered according to the specified method (default is blocking) based on the specified variables.

The comparison component makes use of the `stringsim()` function from the StringMatch toolbox (Denk 2007b) as well as the functions `simplecomp()` for simple (dis-)agreement and `metcomp()` for similarities of metric variables. It computes the similarity profiles for the candidate pairs in the input data frame with respect to the specified matching variable(s) according to the selected method(s) and appends the resulting variable(s) with the defined label(s) to the data frame.

The scoring and classification components estimate matching scores (likelihood ratio is the default) for the candidate pairs in the input data frame from the specified similarity profile (`scoring()`) and determine a classification rule for the candidate pairs according to the selected method (default is empirical Fellegi-Sunter) based on the estimated matching scores and prespecified error levels μ and λ (`classification()`). The score as well as the estimated matching status are appended to the data frame as variables with the defined labels.

Table 2: *Overview of implemented components*

Component	Functionality	Methods
Preparation	parsing	*address and name parsing in different languages*
	standardisation	dictionary provided by the user *integrated dictionaries*
	phonetic coding	American Soundex, Original Russel Soundex NYSIIS, ONCA, Daitch-Mokotoff Koelner Phonetik, Reth-Schek-Phonetik *Metaphone, Double Metaphone* *Phonex, Phonet, Henry*
Candidate Selection	single-pass	cross product / no selection, blocking sorted neighbourhood, string ranking *hybrid*
	multi-pass	*sequence of single-pass*
Comparison	universal	binary frequency-based
	metric variables	tolerance intervals (absolute distance)^p, Canberra
	string variables phonetic coding	see above
	string variables token-based	Jaccard, n -gram, maximal match longest common subsequence, TF-IDF
	string variables edit distances	*Damerau-Levenstein, Hamming* *Needleman-Wunsch, Monge-Elkan* *Smith-Waterman*
	string variables Jaro algorithms	Jaro, Jaro-Winkler Jaro-McLaughlin, Jaro-Lynch
Scoring & Classification	binary outcomes	two-class EM *two-class EM interactions, three-class EM*
	frequency based	*Fellegi-Sunter, two-class EM frequency based*
	similarities	*two-class EM approximate*
	any	*logistic regression*
	no training data	Fellegi-Sunter empiric, Fellegi-Sunter pattern
	training data	*Belin-Rubin*
Decision	assignment	greedy *LSAP*
	review	*possible links, inconsistent values*
Evaluation	confusion matrix	absolute, relative
	quality measures	false match rate Fellegi-Sunter & Belin-Rubin false non-match rate Fellegi-Sunter & Belin-Rubin accuracy, precision, recall, f-measure, specificity unclassified pairs
	plots	*varying classification rules*

The decision component provides an interface to the function `assignment()` that enables 1:1, 1:n/n:1 and particular m:n assignments of the examined records. Eventually, features supporting the review of undetermined record pairs and inconsistent values in linked pairs are intended. `decision()` comes to a final decision concerning the matching status of the record pairs in data frame data based on the preliminary classification, the matching score, and the specified method (default is greedy). A variable representing the final classification is appended to the data frame.

5. Conclusion and outlook

Data quality management is a crucial challenge in scientific and statistical database management, in particular in official statistics, improving the usability and reliability of the data. Entity identification deals with matching records from different data sets or within a single data set that represent the same real-world entity and, thus, enables data integration (at record level) as well as the detection of duplicates, which can both contribute to the enhancement of data quality. Due to the tremendous growth of available data sources, the ongoing trend towards multiple uses and joint usage of data sources, entity identification even has become one of the most crucial issues in data warehousing. It plays a major role in the ETL process and constitutes an essential preparatory step in data mining projects as well.

The statistical entity identification framework presented in this paper emphasizes the importance of a holistic approach taking into account all elementary phases of the process, including preparatory steps, such as standardization of string variables, as well as the evaluation of the quality of the entity identification process. Moreover, it stresses the gain in general applicability and automatisability when making use of statistical models instead of the widespread rule-based approaches. The implementation of the framework poses a considerable step towards statistical entity identification in R. It consists of components corresponding to the stages of the entity identification process, viz. the preparation of matching variables, the selection of candidate record pairs, the creation of similarity patterns, the estimation of matching scores, the (preliminary) classification of record pairs into links, non-links, and possible links, the final decision on the classification and on inconsistent values in linked records, and the evaluation of the results. The projected and current range of functionality of the implementation were presented.

Future work consists in the integration of additional algorithms. The main focus is on further scoring and classification algorithms that significantly contribute to the completion of the framework which should finally be provided as an R package. Moreover, it is intended to test and evaluate the framework with data from the register-based census that has been matched by a conventional equivalence class approach, partly followed by clerical review. Extensions of the conceptual framework with respect to schema-level integration, conventional deterministic entity identification methods, and statistical matching are envisaged as well.

References

Aizawa A., Oyama K. (2005) A Fast Linkage Detection Scheme for Multi-Source Information Integration, in: *Proc. WIRI'05*, 30-39.

Altareva E., Conrad S. (2005) Evaluating and Improving Integration Quality for Heterogeneous Data Sources Using Statistical Analysis, in: *Proc. IDEAS'05*, 406-414.

ANU Data Mining Group (2005) Febrl – Freely extensible biomedical record linkage. Release 0.3, April 7, 2005. http://datamining.anu.edu.au/software/febrl/febrl-03.html, last visited April 10, 2008.

Baxter R., Christen P., Churches T. (2003) A Comparison of Fast Blocking Methods for Record Linkage, in: *Proc. 1st Workshop on Data Cleaning, Record Linkage, and Object Consolidation*, 9th ACM SIGKDD, Washington, DC.

Belin T.R. (1993) Evaluation of Sources of Variation in Record Linkage through a Factorial Experiment, *Survey Methodology*, 19, 13-29.

Belin T.R., Rubin D.B. (1995) A Method for Calibrating False-Match Rates in Record Linkage, *JASA*, 90(430), 694-707.

Bertolazzi P., De Santis L., Scannapieco M. (2003) Automatic Record Matching in Cooperative Information Systems, in: *Proc. ICDT'03*, Siena, Italy.

Bilenko M., Mooney R., Cohen W., Ravikumar P., Fienberg S. (2003) Adaptive Name Matching in Information Integration, *IEEE Intelligent Systems*, 18(5), 16-23.

Bilke A., Naumann F. (2005) Schema Matching using Duplicates, in: *Proc. 21st ICDE'05*, 69-80.

Burkard R.E., Derigs U. (1980) *Assignment and Matching Problems: Solution Methods with Fortran-Programs*, Springer-Verlag, New York.

Calvanese D., DeGiacomo G., Lenzerini M., Nardi D., Rosati R. (1999) A Principled Approach to Data Integration and Reconciliation in Data Warehousing, in: *Proc. DMDW'99*, Heidelberg, Germany.

Cao X., Tung A.K.H., Ooi B.C., Tan K.L., Li S.C. (2004) String Join Using Precedence Count Matrix, in: *Proc. SSDBM'04*, Santorini, Greece.

Chatterjee A., Segev A. (1992) Resolving Data Heterogeneity in Scientific Statistical Databases, in: *Proc. 6th SSDBM*, 145-159.

Chatterjee A., Segev A. (1994) Supporting Statistics in Extensible Databases: A Case Study, in: *Proc. 7th SSDBM*, 54-63.

Chaudhuri S., Ganti V., Motwani R. (2005) Robust Identification of Fuzzy Duplicates, in: *Proc. ICDE'05*, 865-876.

Christen P., Churches T. (2005) A Probabilistic Deduplication, Record Linkage and Geocoding System, in: *Proc. ARC Health Data Mining Workshop*, University of South Australia.

Christen P., Goiser K. (2005) Assessing Deduplication and Data Linkage Quality: What to Measure? In: *Proc. 4th AusDM 2005*, Sydney.

Cochinwala M., Dalal S., Elmagarmid A.K., Verykios V.S. (2001) Record Matching: Past, Present and Future, Technical Report CSD-TR #01-013, Department of Computer Sciences, Purdue University.

Cohen W.W., Ravikumar P., Fienberg S.E. (2003) A Comparison of String Distance Metrics for Name-Matching Tasks, in: *Proc. IJCAI-2003, Workshop on Information Integration on the Web*.

Damerau F.J. (1964) A Technique for Computer Detection and Correction of Spelling Errors, *Comm. ACM*, 7(3), 171-176.

Dasu T., Johnson T. (2003) *Exploratory Data Mining and Data Cleaning*, John Wiley & Sons, Hoboken, NJ.

Dempster A.P., Laird N.M., Rubin D.B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm, *JRSS B*, 39, 1-38.

Denk M. (2002) *Statistical Data Combination: A Metadata Framework for Record Linkage Procedures*, Dissertation Thesis, Dept. of Statistics, University of Vienna.

Denk M. (2006) A Framework for Statistical Entity Identification to Enhance Data Quality, Technical Report wp6dBiz14 br1, EC3, Vienna, Austria.

Denk M. (2007a) A Framework for Statistical Entity Identification in R, in: *Studies in Classification, Data Analysis, and Knowledge Organization Vol. 35 – Proc. 31st Annual Conference of the German Classification Society on Data Analysis, Machine Learning, and Applications (GfKl 07)*, 7.-9.3.2007, Freiburg/Br., Germany, Preisach C., Burkhardt H., Schmidt-Thieme L., Decker R. (Eds.), Springer-Verlag, Heidelberg-Berlin.

Denk M. (2007b) The StringMatch Toolbox: Determining String Compliance in R, in: *Proc. IASC 07 – Statistics for Data Mining, Learning and Knowledge Extraction*, 30.8.–1.9.2007, Aveiro, Portugal.

Denk M., Froeschl K.A. (2000) The IDARESA Data Mediation Architecture for Statistical Aggregates, *Research in Official Statistics*, 3(1), 7-38.

Denk M., Froeschl K.A., Hackl P., Rainer N., (Eds.) (2004) *Special Issue on Data Integration and Record Matching*, Austrian Journal of Statistics, 33, 1-264.

Denk M., Hackl P. (2003) Data Integration and Record Matching: An Austrian Contribution to Research in Official Statistics, *Austrian Journal of Statistics*, 32(4), 305-321.

Denk M., Hackl P., Rainer N. (2005) String Matching Techniques: An Empirical Assessment Based on Statistics Austria's Business Register, *Austrian Journal of Statistics*, 34(3), 235-250.

Denk M., Oropallo F. (2002) *Overview of the Issues in Multi-Source Databases*, DIECOFIS Deliverable 1.1, ISTAT, Rome.

Dey D., Sarkar S., De P. (1998) A Probabilistic Decision Model for Entity Matching in Heterogeneous Databases, *Management Science*, 44(10), 1379-1395.

Dey D., Sarkar S., De P. (2002) A distance-based approach to entity reconciliation in heterogeneous databases, *IEEE Transactions on Knowledge and Data Engineering*, 3, 567-582.

DIECOFIS (2003) The DIECOFIS Project Web Site, http://petra1.istat.it/diecofis/, last visited April 10, 2008.

Dunn H.L. (1946) Record Linkage, *American Journal of Public Health*, 36, 1412-1416.

Efron B. (1979) Bootstrap Methods: Another Look at the Jackknife, *Annals of Statistics*, 7, 1-26.

Elfeky M.G., Verykios V.S., Elmagarmid A.K. (2002) TAILOR: A Record Linkage Toolbox, in: *Proc. ICDE'02*, San Jose, CA.

Fellegi I.P., Holt D. (1976) A Systematic Approach to Automatic Edit and Imputation, *JASA*, 71, 17-35.

Fellegi I.P., Sunter A. (1969) A Theory for Record Linkage, *JASA*, 64, 1183-1210.

Galhardas H., Florescu D., Shasha S., Simon E. (2000) An Extensible Framework for Data Cleaning, in: *Proc. ICDE'00*, San Diego, CA.

Gill L.E. (1997) OX-LINK: The Oxford Medical Record Linkage System, in: *Record Linkage Techniques*, Alvey W. and Jamerson B. (Eds.), Washington, DC: FCSM, 15-33.

Gill L.E. (2001) *Methods for automatic record matching and linking in their use in National Statistics*, GSS Methodology Series, NSMS25, Office for National Statistics UK.

Gravano L., Ipeirotis P.G., Koudas N., Srivastava D. (2003) Text Joins for Data Cleansing and Integration in an RDBMS, in: *Proc. ICDE'03*, 729-731.

Hall P.A.V., Dowling G.R. (1980) Approximate String Matching, *ACM Computing Surveys*, 12(4), 381-402.

Hernandez M.A., Stolfo S.J. (1998) Real-world data is dirty: Data Cleansing and the Merge/Purge problem, *J. Data Mining and Knowledge Discovery*, 2(1), 9-37.

Hernandez M.A., Stolfo S.J. (1995) The Merge/Purge problem for large databases, in: *Proc. ACM SIGMOD*, 127-138.

Jaccard P. (1912) The distribution of the flora of the alpine zone, *New Phytologist*, 11, 37-50.

Jamieson E., Roberts J., Browne G. (1995) The Feasibility and Accuracy of Anonymized Record Linkage to Estimate Shared Clientele among Three Health and Social Service Agencies, *Methods of Information in Medicine*, 34, 371-377.

Jarke M., Lenzerini M., Vassiliou Y., Vassiliadis P. (2000) *Fundamentals of Data Warehouses*, Springer-Verlag, Heidelberg.

Jaro M.A. (1989) Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida, *JASA*, 84, 414-420.

Jin L., Li C., Mehrotra S. (2003) Efficient Record Linkage in Large Data Sets, in: *Proc. DASFAA'03*.

Kelley R. (1984) Blocking Considerations for Record Linkage Under Conditions of Uncertainty, in: *ASA Proc. Social Statistics Section*, 602-605.

Kelley R. (1985) Advances in Record Linkage Methodology: A Method for Determining the Best Blocking Strategy, in: *Record Linkage Techniques*, Kilss B., Alvey W. (Eds.), Washington, DC: FCSM, 199-203.

Kent W. (1979) The Entity Join, in: *Proc. 5th VLDB*, Rio de Janeiro, Brazil, 232-238.

Lee M.L., Hsu W., Kothari V. (2004) Cleaning the Spurious Links in Data, *IEEE Intelligent Systems*, 19(2), 28-33.

Lee M.L., Ling W., Low W.L. (2000) IntelliClean: A knowledge-based intelligent data cleaner, in: *Proc. KDD2000*, Boston, MA, 290-294.

Levenstein V.I. (1966) Binary Codes Capable of Correcting Deletions, Insertions, and Reversals, *Sov. Phys. Dokl.*, 10, 707-710.

Lim E.P., Srivastava J., Prabhakar S., Richardson J. (1996) Entity Identification in Database Integration, *Information Sciences*, 89(1), 1-38.

Lynch B.T., Arends W.L. (1977) *Selection of a Surname Coding Procedure for the SRS Record Linkage System*, Washington, DC: U.S. Department of Agriculture, Statistical Reporting Service.

McCallum A., Nigam K., Ungar L. (2000) Efficient clustering of high-dimensional data sets with application to reference matching, in: *Proc. 6th ACM SIGKDD*, 169-178.

McCallum-Bayliss H. (2000) Identity Resolution in a Global Environment, *IEEE IT Professional*, 6(6), 21-26.

Minton S.N., Nanjo C., Knoblock C.A., Michalowski M., Michelson M. (2005) A Heterogeneous Field Matching Method for Record Linkage, in: *Proc. 5th ICDM'05*, Houston, Texas, 314-321.

Missier P., Lalk G., Verykios V., Grillo F., Lorusso T., Angeletti P. (2003) Improving Data Quality in Practice: A Case Study in the Italian Public Administration, *Distributed and Parallel Databases*, 13(2), 135-160.

Monge A.E. (2000) Matching algorithms within a duplicate detection system, *IEEE Transactions on Data Engineering*, 4, 14-20.

Monge A.E., Elkan C.P. (1996) The field matching problem: Algorithms and applications, in: *Proc. 2nd Int. Conf. Knowledge Discovery and Data Mining*, 267-270.

Neiling M. (1998) Data Fusion with Record Linkage, in: *Proc. 3rd Workshop Foederierte Datenbanken*, Magdeburg.

Newcombe H.B., Kennedy J.M., Axford S.J., James A.P. (1959) Automatic linkage of vital records, *Science*, 3381, 954-959.

Odell M.K., Russel R.C. (1918) US Patent no. 1,262,167.

Odell M.K., Russel R.C. (1922) US Patent no. 1,435,663.

Porter E., Winkler W.E. (1997) Approximate String Comparison and its Effect on an Advanced Record Linkage System, RR97-02, U.S. Bureau of the Census.

Pu C. (1991) Key Equivalence in Heterogeneous Databases, in: *Proc. 1st International Workshop on Interoperability in Multidatabase Systems*, 314-316.

Rahm E., Do H.H. (2000) Data Cleaning: Problems and Current Approaches, *IEEE Technical Bulletin on Data Engineering*, 23(4), 3-13.

Rubin D.B. (1987) Multiple Imputation for Nonresponse in Surveys, John Wiley & Sons, New York.

Salton G., McGill M. (1983) *Introduction to Modern Information Retrieval*, McGraw-Hill, New York.

Scheuren F., Winkler W.E. (1993) Regression Analysis of Data Files that are Computer Matched, *Survey Methodology*, 19, 39-58.

Scheuren F., Winkler W.E. (1997) Regression Analysis of Data Files that are Computer Matched II, *Survey Methodology*, 23, 157-165.

Schuermann J. (1996) *Pattern Classification*, John Wiley & Sons, New York.

Smith T.F., Waterman M.S. (1981) Identification of common molecular subsequences, *J. Molecular Biology*, 147, 195-197.

Tayi G.K., Ballou D.P. (1998) Examining Data Quality, *Comm. ACM*, 41(2), 54-57.

Tepping B.J. (1968) A Model for Optimum Linkage of Records, *JASA*, 63(324), 1321-1332.

Verykios V.S., Elmagarmid A.K., Houstis E.N. (1999) Record Matching to Improve Data Quality, Technical Report CSD-TR #99-005, Department of Computer Sciences, Purdue University.

Wang Y.R., Madnick S.E. (1989) The Inter-Database Instance Identification Problem in Integrating Autonomous Systems, in: *Proc. 5th Int. Conf. on Data Engineering*, Los Angeles, CA, 46-55.

Winkler W.E. (1985) Preprocessing of Lists and String Comparison, in: *Record Linkage Techniques*, Kilss B., Alvey W. (Eds.), Washington, DC: FCSM, 181-187.

Winkler W.E. (1990) String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage, in: *ASA Proc. Section on Survey Research Methods*, 354-359.

Winkler W.E. (1991) Error Model for Analysis of Computer Linked Files, in: *ASA Proc. Section on Survey Research Methods*, 472-477.

Winkler W.E. (1993) Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage, in: *ASA Proc. Section on Survey Research Methods*, 274-279.

Winkler W.E. (1994) Advanced Methods for Record Linkage, in: *ASA Proc. Section on Survey Research Methods*, 467-472.

Winkler W.E. (1995) Matching and Record Linkage, in: *Business Survey Methods*, Cox B.G. et al. (Eds.), J. Wiley, New York, 355-384.

Winkler W.E. (1999) Issues with Linking Files and Performing Analyses on the Resultant Merged Files, in: *ASA Proc. Section on Government Statistics*.

Winkler W.E. (2000) Frequency-Based Matching in the Fellegi-Sunter Model of Record Linkage, RR2000-06, U.S. Bureau of the Census, Statistical Research Division.

Winkler W.E. (2004) Methods for Evaluating and Creating Data Quality, *Information Systems*, 29(7), 531-550.

Winkler W.E., Scheuren F. (1996) Recursive Analysis of Linked Data Files, RR96-08, U.S. Bureau of the Census, Statistical Research Division.

Winkler W.E., Thibaudeau Y. (1991) An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 U.S. Census, RR91-09, U.S. Bureau of the Census, Statistical Research Division.

Yesilcay Y. (1993) The BOYS Algorithm for Determining Optimum Matching Rules, *Computational Statistics & Data Analysis*, 16, 219-234.

The Swiss Census 2010: Moving towards a comprehensive system of household and person statistics

Markus Schwyn, Jean-Paul Kauthen
Swiss Federal Statistical Office, CH-2010 Neuchâtel
e-mail: markus.schwyn@bfs.admin.ch, jean-paul.kauthen@bfs.admin.ch

Abstract: This paper provides a brief overview of the ongoing conception of the Swiss Census 2010 and the setup of a new integrated system of household and person statistics. The Swiss Census that was held (almost) every 10 years from 1850 until 2000 will be modernized. It will be based on an annual register survey, an annual structural survey of a large sample and five topic-based sample surveys alternating every year and a small annual omnibus survey. The process design will be organized along the EDIMBUS process model.

Keywords: census, register, survey, sample, data collection, data integration, integrated system, EDIMBUS, process model

1. The basic principles of the 2010 census

1.1 The 2010 census

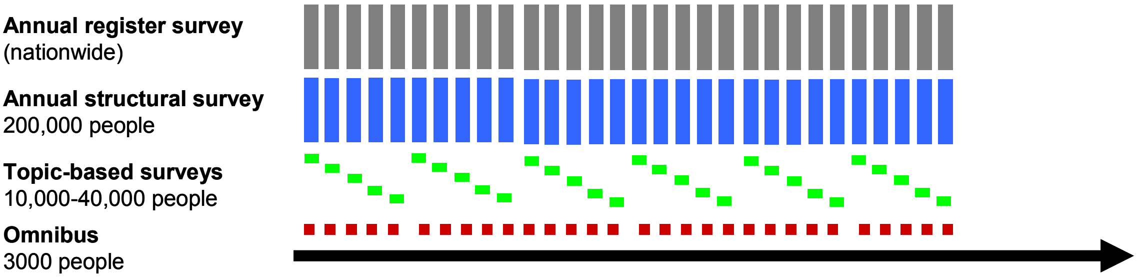
On June 22, 2007 the Swiss parliament passed a completely revised version of the act concerning the federal census. The act came into force on January 1, 2008. The intention of the Swiss Federal Council and parliament in passing this new act was to modernise official statistics. The new census represents a comprehensive change of the system. The traditional census, which was previously carried out every ten years with the entire population, will be replaced by an integrated statistical system. This combines the use of existing person data registers with sample surveys which are carried out and evaluated in an annual cycle.

The new system offers a wide range of benefits. The information will be available more frequently, on a wider range of topics and within a very short period of time. The availability of the latest data on an annual basis will improve the ability to monitor important, politically relevant, sociocultural topics on a regular and systematic basis. The new system can also be constantly updated and developed. Importantly, it also offers an excellent cost/benefit ratio. Improved coordination and the increased use of synergies will result in a significant reduction in costs and administrative work, while at the same time lightening the burden on the interviewees and the municipalities.

1.2 Key elements of the census

The Census Act specifies the requirements for the new system. Data concerning the population structure and societal development in Switzerland are to be collected every year. The relevant topics are described in article 1 of the act. The broad range of topics is covered by four different surveys which will be carried out annually: the register survey, the structural survey, also called Swiss Population Survey, one topic-based survey and the CH omnibus (Figure 1).

Figure 1: *Survey time line*



The annual register survey is based on the population registers of the municipalities and cantons, the most important federal person data registers and the National Register of Buildings and Dwellings. Therefore, the survey provides basic information annually about the population and about buildings and dwellings at the smallest spatial resolution. The act concerning the harmonisation of official person data registers, which came into force on January 1, 2008, has fulfilled one of the central requirements for the easy and efficient use of the register data. The act specifies the identifiers and the attributes which the registers must include, determines the content and form of the registers and controls the exchange of data between them.

The structural survey is an annual sample survey of 200,000 people. As it is a population survey, it includes important attributes which are not currently available in the registers. The survey covers people living in private households who are aged 15 or over. The interviewees provide information about themselves and their households. An annual sample survey of 200,000 people allows statistical analyses to be carried out for all the cantons and for groups of 15,000 people with a sufficient accuracy. After five years it will be possible to make assertions about groups of 3,000 people using data pooling, by combining five consecutive annual structural surveys. Such a method is used for example in the American Community Sample. Within these groups, units of 140 people can be identified after one year and of 28 people after five years. The cantons can supplement the survey within their own area at their own expense to improve the results further.

Since the new Swiss Federal Census Act foresees a duty to give information for the Swiss Population Survey, the response rate is expected to be very high. The anticipated accuracy of the Swiss Population Survey has been analyzed in Eichenberger, Hulliger and Potterat (2007). The concepts of estimation of the size of a group, the resolution of a survey, i.e. the smallest estimable size, and the comparison of proportions are introduced and discussed in that paper.

Topic-based sample surveys will also be carried out annually using a sample size of 10,000 to 40,000 people. The following topics will alternate on a five-yearly cycle: "mobility", "education", "health", "families" and "language, religion and culture". The existing health survey and the transport micro-census will be incorporated into this system. Because of the size of the samples, the topic-based surveys allow conclusive results to be produced for the whole of Switzerland and the seven major regions. The micro-census on mobility and transport will also provide results for the large urban areas. The cantons can also supplement these surveys at their own expense.

The survey referred to as the CH omnibus is a new flexible tool which will provide rapid answers to current questions. This annual sample survey of around 3,000 people offers interested groups the opportunity to join in by asking specific questions. The survey produces results for the whole of Switzerland which can be rapidly processed and published.

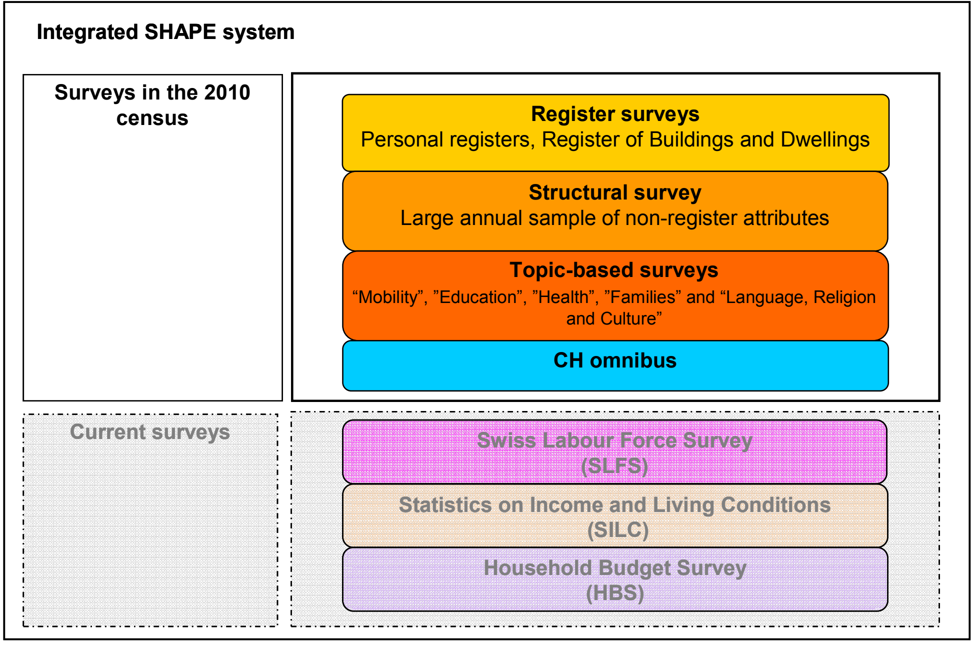
1.3 The census as part of an integrated system

The new focus of the census has transformed it into the backbone of a new integrated system for household and person statistics (SHAPE). In the future it will be part of a system which combines the systematic use of existing administrative registers and sample surveys of people and households. The content, method and organisation of the various surveys will be linked and coordinated with one another. In particular in the long term, the integrated system provides added value which makes it much more than the sum of its parts.

The different areas which make up the census supplement existing surveys in order to create a comprehensive picture of socioeconomic statistics (Figure 2). Three annual surveys are already carried out on the topics of "work" and "income, consumption and living conditions".

- The Swiss Labour Force Survey (SLFS) provides data about the labour market and about working life in general. In order to coordinate with European statistics, the results will in future be produced quarterly.
- The Household Budget Survey (HBS) provides regular information about the consumption habits and income of private households in Switzerland.
- The new survey on income and living conditions introduced in 2007 (SILC: Statistics on Income and Living Conditions) covers a wide area including income, education, work, childcare, the composition of households, the housing situation and health.

Figure 2: *Integrated system for household and person statistics (SHAPE)*


Integrated SHAPE system
Surveys in the 2010 census
Register surveys
Personal registers, Register of Buildings and Dwellings
Structural survey
Large annual sample of non-register attributes
Topic-based surveys
“Mobility”, ”Education”, ”Health”, ”Families” and “Language, Religion and Culture”
CH omnibus
Current surveys
Swiss Labour Force Survey (SLFS)
Statistics on Income and Living Conditions (SILC)
Household Budget Survey (HBS)

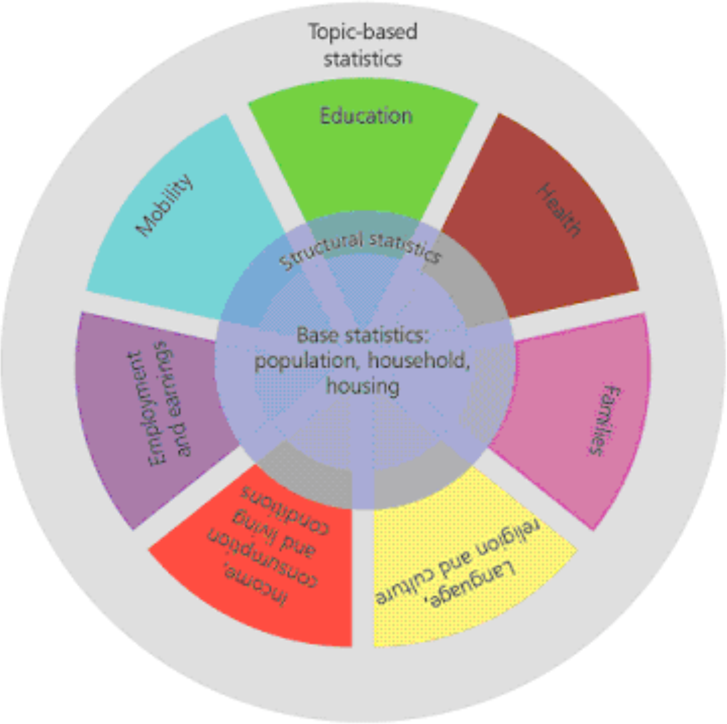
2. The new census: The information provided and the survey programme

2.1 Statistics and surveys

At the heart of the new statistical information system is the resident population of Switzerland, in other words, the people and their households. For every person, a link with the dwellings and residential buildings is formed. The new census system allows four different surveys and their resulting data to be brought together. In future, this integrated system will make four types of statistics available with a different topic-based and spatial depth of focus.

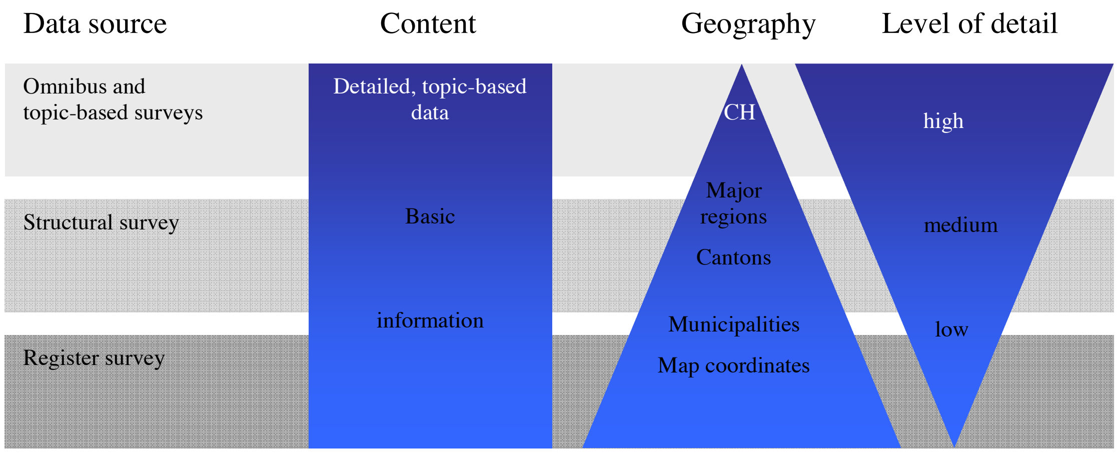
- Basic annual statistics on the population, households and housing on the basis of the nationwide register survey.
- Annual structural statistics based on the structural survey. These also include the traditional topics of the census. They provide additional information to the basic statistics and form the basis for the analysis of the topics at canton level.
- Detailed annual statistics on the topics of "labour" and "income, consumption and living conditions" using the topic-based surveys.
- Detailed annual statistics on one of the other topics.
- Annual statistics on current issues on the basis of the CH omnibus.

Figure 3: *Integrated system: Statistics and topics*


Topic-based statistics
Education
Health
Families
Language, religion and culture
Income, consumption and living conditions
Employment and earnings
Mobility
Structural statistics
Base statistics: population, household, housing

The system includes all the information about persons, households and housing from the basic statistics, structural statistics and detailed topic-based statistics (see Figure 3). These cross-sectional perspectives and the integration of the results from the various surveys allow the seven topic areas to be covered comprehensively. The principles of the data integration in simplified form are as follows:

Figure 4: *Integrating the statistics*



2.1.1 Embedding into the European statistics systems

When the bilateral agreement on statistics with the EU came into force on January 1, 2007, one important institutional and legal condition changed. Switzerland is now responsible for ensuring on a systematic basis that Swiss statistics are comparable with those of the EU and EFTA countries. For this purpose, Switzerland has adopted the corresponding legal instruments of the EU. The United Nations Economic Commission for Europe (UNECE) and the Statistical Office of the European Communities (Eurostat) have issued joint recommendations for population and housing censuses in 2010. These describe the attributes to be included in the survey, the recommended additional attributes and the classifications and definitions which ensure that the data can be subjected to international comparisons. Switzerland will follow the UNECE/Eurostat recommendations concerning the key area of the census results and consider the Regulation on Population and Housing Censuses in the EU Member States.

2.1.2 A new sample survey system supported by register data

From 2010 onwards, samples can be drawn using auxiliary information collected in the population registers and the National Register of Buildings and Dwellings. The population statistics play a central role in providing reference figures for estimations based on the sample surveys. The information on the overall population, on population groups and on residential buildings and dwellings is very important for the purpose of planning, weighting and calibration procedures of the sample surveys.

2.1.3 Basic statistics on population, households and housing based on register data

The basic statistics on population, households and housing form the core of the new statistics system. The main source of these statistics is the register survey. Demographic evaluations of the entire population are carried out annually. These cover the entire resident population living in private and collective households. The basic information on population and households is combined with data on residential buildings and dwellings from the National Register of Buildings and Dwellings. This link provides annual, geocoded information on the population, households, dwellings and living conditions. It also allows annual, small-scale evaluations at the level of the municipalities or below to be carried out. Alongside cross-sectional statistics of this kind, the sources of basic demographic statistics can also be analysed from a chronological perspective.

Indicators for the basic output on population and households

The basic statistics on the population and households provide classic demographic indicators, such as population structure, size and evolution. They cover the entire resident population living in private and collective households and provide annual results on the following indicators:

- Size and structure of the population
- Population balances
- Change of status: Status of the resident foreign population
- Acquisition of Swiss citizenship
- Migration: Internal and international migration
- Private and collective households
- Marriages and civil partnerships
- Divorces and dissolutions of civil partnerships
- Widowhood
- Recognition of paternity
- Adoptions
- Live births and stillbirths
- Deaths
- Population scenarios

Indicators for the basic output on housing

The basic statistics on housing provide information about the building and housing stock and its structure, together with living conditions. They cover all the residential buildings and dwellings in Switzerland and provide annual results on the following indicators:

- Building and housing stock
- Age of buildings and dwellings and the date of the most recent renovation
- Housing supply
- Residential moving behaviour
- Heating and energy sources
- Occupied or unoccupied dwellings

2.1.4 Structural statistics based on the structural survey

The structural statistics combine the results of the structural survey and the register survey and therefore provide information which goes beyond the restricted scope of the basic statistics. They complement the basic annual statistics with a large-scale sample survey of attributes not included in the registers. They also forge a link between the basic statistics generated from register data and the detailed information of the individual topic-based statistics, by creating general overviews of the most important population structures. The fact that the statistics are available annually also allows important changes in the population structure to be monitored regularly at a detailed level. The main source of structural statistics is the link between the newly introduced annual structural survey of 200,000 people and the register survey described above.

Indicators for the basic output of structural statistics

The structural statistics provide additional information on the basic statistics, together with base information for the analysis of the topic areas. The corresponding person statistics cover the resident population, excluding those people who live in collective households. The information on dwellings relates to occupied dwellings. The structural statistics provide annual results on the following topic areas and indicators:

Basic statistics on population, households and housing

- Migration
- Housing rents
- Home ownership ratio and rate
- Housing situation

Introduction to the topic areas:

Work

- Employment and participation in the labour market
- Unemployment

Mobility

- Commuting
- Means of transport
- Traffic volume

Education

- Highest level of education obtained
- Current education
- Education and the labour market
- Original training and current occupation

Language, religion and culture

- Languages
- Religions

Families

- Household structures, family types and living arrangements
- Life/work balance

2.1.5 Detailed statistics on the topic areas

The annual structural statistics cover seven topic areas in a regular cycle. The topics of "work" and "income, consumption and living conditions" are based on the existing SILC, SLFS and HBS surveys and are surveyed and analysed annually. They already allow analyses to be carried out at the level of the major regions and meet the political requirements of the Swiss Confederation.

New surveys on the five topics of "mobility", "education", "health", "families" and "language, religion and culture" will be introduced and integrated into the 2010 census. The intention is to cover these topics in a five year rhythm. This level of frequency is adequate because the annual information from the structural survey gives a general overview of the most important changes, which usually take place more slowly and continuously than those in the labour market or in household incomes. The geographical depth of the analyses will depend on the requirements and on their political relevance for Switzerland. The evaluations cover the permanent resident population, but generally exclude those people living in collective households.

The output is based on the detailed topic-based surveys and modules which supplement the existing surveys in the integrated SHAPE system. The indicators for the topic areas will be defined as part of the design process of the individual surveys in cooperation with the cantons and other interested groups.

2.2 Consolidation into an integrated system

The new census can only exploit its full potential if it takes the form of an integrated system. It is more than the sum of the various individual statistics. In order to create an integrated system, integration components are needed which will bring together surveys based on different data sources. The four integration components are as follows:

- The basic populations shared by all the surveys.
- The new social security number which as a person identifier uniquely identifies a person in different data sources.
- The building and dwelling identifiers which allow the formation of households to be identified and the attribution of people and households to buildings and dwellings.
- The core variables which ensure that the results in different surveys are comparable and that the population groups and basic populations are uniformly defined.

2.2.1 Common basic populations

The different surveys can only be consolidated on a common foundation consisting of the same basic populations which are uniformly defined and harmonised. They determine the framework of the person, household and housing statistics using the statistical units which are being monitored.

The following three basic populations form the common foundation of all the statistics in the SHAPE system:

- The permanent resident population, without people living in collective households, which are defined in the Register Harmonisation Ordinance (e.g. homes for elderly people, prisons, etc.).
- All private households, in other words, all groups of people living under one roof in the same dwelling. Collective households are excluded.
- The third common basic population consists of the occupied residential buildings and dwellings.

2.2.2 Personal identification number

As part of the process of harmonising the official population registers, the new social security number will be included in the registers listed in the Register Harmonisation Act. The number can be used as a personal identification number (PIN) for statistical purposes. The introduction of the thirteen-digit PIN into the registers specifically referred to in the act is a central element of the process of linking data for statistical purposes. Data with the new social security number are regarded as non anonymised data. For this reason, measures will be needed to guarantee the protection and the confidentiality of the data. The use of one or more statistical identification numbers (pseudonymised numbers) which are specific to each area and other technical protection measures will be investigated.

2.2.3 Dwelling and building identification numbers

The process of register harmonisation involves assigning to each person in the population register the federal building identification number (EGID) and the federal dwelling identification number (EWID) of the dwelling in which they live, from the

National Register of Buildings and Dwellings. This allows the composition of households to be determined on the basis of the register. The federal dwelling identification number is a three-digit identifier of the dwellings in the Federal Buildings and Dwellings Register. It is unique within each building and is always assigned in combination with the nine-digit federal building identification number.

2.2.4 Harmonised core variables

The definition of core variables is a further precondition for a harmonised structuring of the statistical information. The attributes should, in future, be used uniformly in all the surveys. They allow population groups to be defined and identified in a standardised way. The core variables also generate a lowest common denominator which guarantees that the results of the different statistics and the statistical monitoring of specific population groups are comparable.

The concept of the harmonised core variables allows assertions to be made about the statistical units in the different topic areas. For example, conclusions can be drawn about the mobility and travel behaviour, health prevention measures, use of cultural activities, language skills or religious practice of groups according to the "highest level of education attained". As a result, definable groups of people (for example, people with university degrees) can be described and analysed in the context of the topic areas.

3. Supplementing the structural survey

The structural survey, also called Swiss Population Survey, is a sample survey of persons. This means that the information obtained can be extrapolated to produce statistical results for the entire population. The results of these projections are estimates which are subject to certain random sampling errors.

The standard programme consists of a survey of 200,000 people aged 15 years or more who are living in private households. As a result, around 2.7% of the entire resident population is surveyed every year. This corresponds to approximately 3.5% of the people aged 15 years or over. The precision of the assertions made on the basis of a sample of this kind can be described in terms of two factors. The depth of focus indicates the smallest geographical unit for which reliable estimates can be made. In contrast, the resolution represents the smallest possible group which can be precisely estimated independently of the size of the geographical unit.

The standard programme allows statistical assertions relating to individual attributes to be made with a sufficient level of accuracy for groups of 15,000 people. These groups can correspond to regional or socioeconomic boundaries, for example, women with a university degree aged between 30 and 40 or single mothers. Using this depth of focus, sound results can be achieved for all the attributes in the structural statistics for all the cantons, larger municipalities and larger districts of large cities.

Estimates for small groups define the mesh size of the monitoring net. In the standard programme the size of a group for which estimates can be made is 140 people, regardless of the attribute being investigated. These people become trapped in the monitoring net, so to speak. Where attributes apply only to a smaller group of people, for example, if only 100 people in a municipality have a tertiary education, they cannot be identified with certainty in the analysis grid.

Pooling or combining the data from structural surveys over several years allows a correspondingly larger sample to be formed. As a result, the precision and significance of the results also increase. However, this information does not refer to a clearly defined survey date, but represents an average over the period being investigated. Data will be pooled over three and five years, resulting in sample sizes of 600,000 and 1,000,000 people. The depth of focus and the resolution change accordingly.

Details about the anticipated accuracy of the Swiss Population Survey can be found in Eichenberger, Hulliger and Potterat (2008). The Swiss cantons have the possibility to increase the sample sizes for the structural survey and the topic-based surveys at their cost and reduce in this way the sample errors.

4. Integration of sample survey and register data

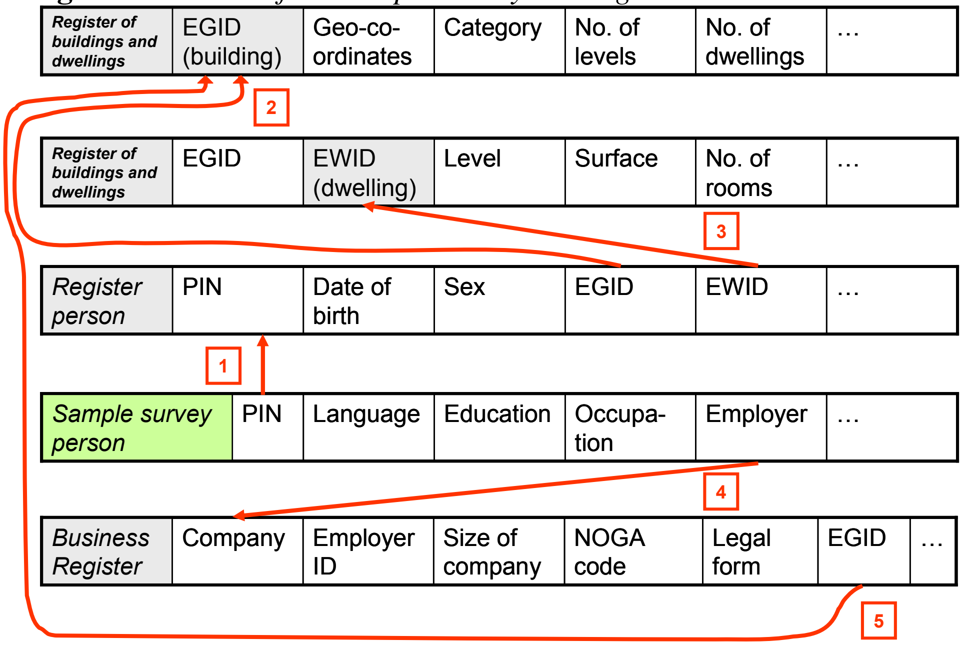
One of the main principles of the new integrated system is that in a sample survey, information that is available in a register will no longer be questioned. For example, information about sex, civil status and nationality will not be questioned in the Swiss Population Survey since this information is already present in the municipality registers. This procedure has the aim to reduce the burden of the survey respondents. This is also a preoccupation of the Swiss Constitution in which register harmonization is put forward to reduce census efforts (Article 65).

In the new integrated system of household and person statistics (SHAPE), several types of data integrations will be carried out:

- data from person registers – data from the National Register of Buildings and Dwellings (links 2 and 3 in Figure 5),
- person data from sample surveys – data from person registers (link 1 in Figure 5),
- person data from surveys – data from the business register (link 4 in Figure 5).

Figure 5 contains a schematic view of the combination of sample survey and register data in the new integrated system.

Figure 5: *Outline of the sample survey and register data combinations*



Register of buildings and dwellings	EGID (building)	Geo-co-ordinates	Category	No. of levels	No. of dwellings	…

Register of buildings and dwellings	EGID	EWID (dwelling)	Level	Surface	No. of rooms	…

Register person	PIN	Date of birth	Sex	EGID	EWID	…

Sample survey person	PIN	Language	Education	Occupation	Employer	…

Business Register	Company	Employer ID	Size of company	NOGA code	Legal form	EGID	…

In what follows, we will briefly describe the expected benefits of this register data and sample survey data integration. Overall it is expected that the quality of the annual Swiss Population Statistic will be improved immensely.

4.1 Register person – building in the Register of Buildings and Dwellings (link 2)

This link provides geo-coordinates (East and North coordinates) for every person. Thus the improved statistical information will be geo-encoded and can be made available for very small geographical areas. In the future it will be possible to produce basic demographic information down to the level of city neighbourhoods.

During the editing and imputation phase, this link may also help to increase – e.g. through automatic imputations – the quality of the building status (in project, in construction, existing, dismantled), category (one family home, several families home, etc.) and number of dwellings.

4.2 Register person – dwelling in the Register of Buildings and Dwellings (link 3)

The combination of EGID and EWID allows linking every person to a dwelling. Thus exhaustive information about housing conditions may be obtained. Since all persons who have been attributed the same EGID-EWID combination form a household, households are also linked to dwellings.

This link also allows defining the set of inhabited or temporarily inhabited dwellings. During the editing and imputation phase, this link may also help to increase – e.g. through automatic imputations – the quality of the dwelling attributes like status (in project, in construction, existing, cancelled), number of rooms and surface.

4.3 Survey person – register person (link 1)

This link allows the enrichment of sample survey data with demographic data from registers thus allowing numerous cross tabulations of sample survey data and register data on the person, building and dwelling level. The formation of population sub-groups based on sample survey and/or register attributes and comparison of results between sub-groups in the same survey or across surveys becomes also possible.

4.4 Survey person – business register (link 4)

Finally this link allows to couple sample survey person data with data of the business register. Thus information on NOGA classification, size and legal form etc. of a possible employer can be added to the sample survey data. The business register is maintained by the Federal Statistical Office.

The employer information in the business register also contains a building identifier (EGID) and thus the possible working place of a sample survey respondent can also be geo-encoded (link 5), thus allowing e.g. the computation of the commuting distance.

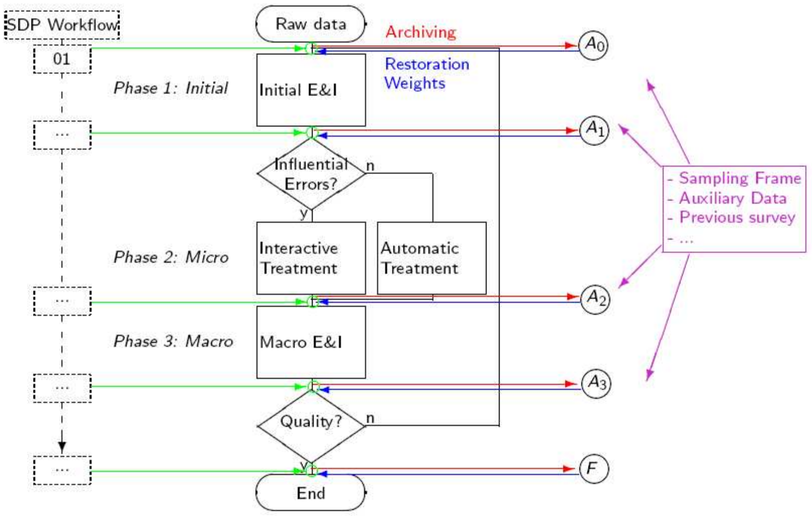
The same holds for a possible school or education site.

5. Process design for the 2010 census

5.1 EDIMBUS process model

In the conception and process design phase of the integrated SHAPE system, the project team at the Federal Statistical Office used the EDIMBUS process model to design the data preparation phase. The applied statistical data preparation (SDP) process has been developed in the "Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys in the European Statistical System" (EDIMBUS) project carried out by the Italian Statistical Institute, Statistics Netherlands and the Swiss Federal Statistical Office. Although primarily designed for business surveys, its principles can be carried over to person, household, building and dwelling surveys.

Figure 6: *EDIMBUS model*



The data preparation process is subdivided into phases as can be seen in Figure 6. The first phase contains an initial data preparation where simple (e.g. systematic) errors are treated first. In a second phase, the data flow is separated in a critical and non-critical data stream with manual and automatic treatments respectively. In phase 3, data quality is checked on a macro level. At the end of each phase, a copy of the data is preserved for further possible iterations. We refer to the EDIMBUS website for more details.

The EDIMBUS process model has been applied to the three major types of surveys of the new integrated census system, namely the register survey, the structural survey and the topic-based surveys. If applicable, different data channels (registers, paper census form, Internet questionnaire and CATI) have been considered.

5.2 Implementation and next steps

The next steps in the setup of the integrated system are very briefly sketched out in the following:

- detailing of the process steps;
- specifications for the IT company in charge of the development of the system;

- decision for a combination of off-the-shelf solutions or a proprietary development;
- development and testing of the system.

6. Further information

Further up-to-date information on the integrated system can be found at:

- www.bfs.admin.ch / Modernisation projects / 2010 census
- www.bfs.admin.ch / Modernisation projects / SHAPE
- www.bfs.admin.ch / Modernisation projects / Register harmonisation

Published information:

- Brochure "SHAPE: The New Statistical System on Households and Persons"
- Message for the fully revised version of the Act on the Federal Census of November 29, 2006 (06.093)
- Message for the harmonisation of official person data registers of November 23, 2005 (05.083)

The most important legal foundations of the 2010 census and SHAPE:

- Statistics Act (BStatG) of October 9, 1992 (SR 431.01)
- Act of June 22, 2007 concerning the Federal Census (Federal Census Act) (SR 431.112)
- Data Protection Act (DSG) of June 19, 1992 (SR 235.1)
- Act on the Harmonisation of Population Registers and Other Official Personal Data Registers (RHG) of June 23, 2006 (SR 431.02)

International requirements for the 2010 census:

- Recommendations for the 2010 round of population and housing censuses within the ECE region: The United Nations Economic Commission for Europe and the Statistical Office of the European Communities, New York and Geneva 2006 (only available in French and English).
- Regulation on Population and Housing Censuses in the EU Member States: Presidency Common Position, Working Group of Demography and Census, Luxemburg, 6-7 March 2008.

EDIMBUS

- EDIMBUS website: http://edimbus.istat.it

References

Eichenberger Ph., Hulliger B., Potterat J. (2008) Anticipated Accuracy of the Swiss population Survey, submitted for publication.

Proposal for a quality framework for the evaluation of administrative and survey data

Piet J.H. Daas, Judit Arends-Tóth, Barry Schouten, Léander Kuijvenhoven
Statistics Netherlands, Kloosterweg 1, 6412 CN Heerlen, The Netherlands,
e-mail: pdas@cbs.nl

Abstract: Statistics Netherlands is increasingly making use of data sources collected and maintained by others, such as administrative data, for the production of statistics. Since the quality of the statistics produced is affected by the quality of the data sources used, it is of vital importance that Statistics Netherlands is able to unambiguously determine the quality, i.e. the statistical usability, of external data sources. For this purpose a quality framework was developed for administrative data. The framework is discussed in this paper. It was found that the framework could also be used for the evaluation of survey data. As such, a single framework could potentially be used for the evaluation of all input data sources used for the production of statistics.

Keywords: Quality framework, Administrative Data, Survey Data, Quality aspects.

1. Introduction

National Statistical Institutes (NSI's) collect data for the production of statistics. Apart from the data obtained through surveys, NSI's are increasingly making use of data that is collected and maintained by others for non-statistical purposes. Administrative data is an example of such a data source (Wallgren and Wallgren 2007). It is produced as a result of administrative processes of organizations but it is -very often- also an interesting data source for NSI's. During the last decade, more and more NSI's have realized this (UNECE 2007). This is especially the case for the NSI's in the Nordic countries. In these countries administrative data is already the main data source for the production of official statistics (Statistics Finland 2004, UNECE 2007, Wallgren and Wallgren 2007).

A major advantage of the use of administrative data for statistics is the fact that it drastically reduces the costs of data collection and the response burden on enterprises and persons. Since administrative data often completely covers whole populations, in various time references, it is also particularly well suited for the creation of detailed and longitudinal statistics on subpopulations and regions (Wallgren and Wallgren 2007). An additional stimulus for its use is the increased use of information and communication technology in public administrations. As a result, more and more administrative data is becoming available in an electronic form that can be easily collected and processed by the NSI (Børke and Bergstrøm 2006).

From a statistical point of view, administrative data has some disadvantages. The most important one is the fact that the collection and maintenance of administrative data are beyond the control of the NSI. It is the administrator of the data source that manages these aspects. The same is true for the units and variables an administrative data source contains. These are defined out of administrative rules and may therefore not be identical to those required by the NSI (Wallgren and Wallgren 2007). It often takes considerable effort to unambiguously determine the statistical usability of administrative data (ESC 2007, Everaers and Van der Laan 2003). Since the production of high quality statistics depends on the quality of the input data, it is of vital importance that NSI's are able to unambiguously determine the quality, i.e. the statistical usability, of administrative data; preferably in a cost efficient way. Although administrative data has been used by statistical offices for quite some time, the determination of the quality of those data sources prior to there use has not received a lot of attention (UNECE 2007, Sæbø et al. 2003). Most of the quality studies performed at NSI's have focused on the quality of data collected by surveys (Biemer and Lyberg 2003, Van den Brakel et al. 2007) and on the quality of the statistics produced (Eurostat 2003a-b, 2005b). Only a relative small number of studies has focused on the quality aspects of administrative data used for statistical purposes (see section 2). For all clarity, the word 'aspect' is used in this paper to describe a measurable part of quality.

In this paper an overview is given of the quality framework developed for the determination of the quality of administrative and survey data at Statistics Netherlands. The framework was originally developed for the evaluation of administrative data but early on in the project it was found that it could also be applied to survey data. The main goal of the work described in this paper is to identify all quality aspects relevant for the statistical use of data sources.

2. Statistical quality

With the adoption of the European Statistics Code of Practice, the NSI's of EU-Member States have committed themselves to an encompassing approach towards high quality statistics (Eurostat 2005a). NSI's of the EU-Member states involved and NSI's of some other European countries, such as Norway, report the quality of their statistical products by using six quality dimensions. The dimensions used are: Relevance, Accuracy, Timeliness and punctuality, Accessibility and clarity, Comparability, and Coherence (Eurostat 2005b). For the determination of the quality of the input data of NSI's, such as administrative data, the six standard quality dimensions are not always applicable. This was also highlighted in a publication of Eurostat (Eurostat 2003c). The study of the quality aspects of administrative data was the starting point for the work described in this paper.

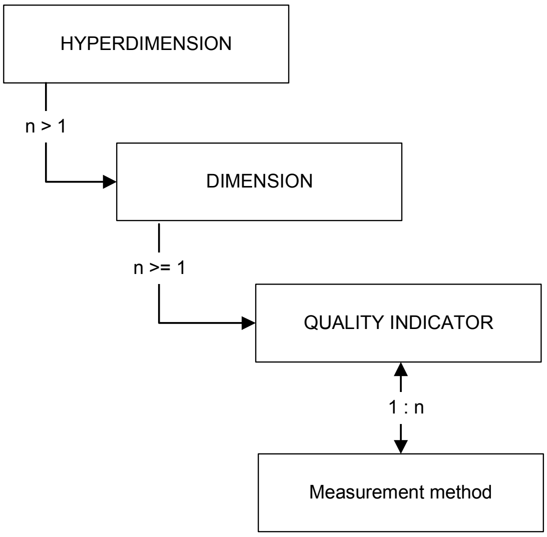
2.1 Quality aspect identification

An extensive literature study revealed that the views on the composition of the quality of administrative data -to be used for statistics- varied greatly. Unfortunately hardly any publications where found that attempted to construct a complete quality framework for administrative data. The most important developments in this area are described in a limited set of papers and books, these are: Wallgren and Wallgren (2007), Daas and Fonville (2007), Eurostat (2003c), Karr et al. (2006), UNECE (2007), Thomas (2005), and ONS (2005). When the results of these studies are compared, a remarkable difference between the number and types of quality groups or dimensions identified for the statistical quality aspects of administrative data is observed. In our opinion this points out the complexity of the problem but also suggests that every researcher or group of researchers had a slightly different view on this topic. The progress in this field would be considerable if these heterogeneous views could somehow be combined into a single framework. This exercise was performed by the authors of this paper. The main objective of this paper is to bring together these different views.

By combining the various quality aspects identified at Statistics Netherlands (Daas and Fonville 2007) and those mentioned in the publications of others (listed above), the authors attempted to get a complete overview of the quality aspects of administrative data relevant for statistical use. Every quality aspect identified in every study was compared with those observed in any of the other studies. During this exercise two important findings emerged. First, there is a general level of mutuality; in a lot of studies many (very) similar quality aspects were identified. Second, the statistical quality of administrative data is more than a simple dimensional concept. Depending on the perspective from which the administrative data source is looked upon, different quality aspects prevail. Such a perspective, a high level view at statistical quality, was described by Karr et al. (2006). In there point of view, statistical quality not only consists of dimensions and indicators but also of a concept they identified as a 'hyperdimension'. A hyperdimension is a way of looking at quality at a level higher than that of a dimension; hence the name 'hyper'dimension.

In a hyperdimension several dimensions of quality are grouped. The quality aspects included are highly influenced by the contextual view on the quality of the data source (Karr et al. 2006). With the above in mind, a quality framework was developed for administrative data that consists of hyperdimensions, dimensions, quality indicators and measurement methods (figure 1). A hyperdimension is composed of two or more dimensions and each dimension contains one or more quality indicators. A quality indicator is measured or estimated by one or more methods. The relation between the various quality aspects included in the framework is shown in figure 1.

Figure 1: *Relation between the different aspects of quality in the framework developed*



2.2 Quality framework

The identification and comparison of all the quality aspects identified for administrative data revealed four discernible contextual ways of looking at the quality of such a data source. The four hyperdimensions identified were called: Source, Metadata, Data, and

Process. Each hyperdimension highlights specific quality aspects of the data source. The quality indicators in the first three hyperdimensions (Source, Metadata, and Data) are all different. These three hyperdimensions are also ordered according to an increase in the level of detail. The quality indicators in the Data hyperdimension, for instance, report on quality aspects in the data source at a much more detailed level than the quality indicators included in the Metadata hyperdimension. This same is true for the Metadata and Source hyperdimensions. The four hyperdimensions identified are briefly discussed below. More detailed information is presented in tables 1, 2, and 3.

2.2.1 Source hyperdimension

In the Source hyperdimension the data source is viewed upon as a file that is delivered by the data source maintainer to the NSI. Only quality aspects related to this view are included in this hyperdimension. In the Source hyperdimension, five dimensions are distinguished: Supplier, Relevance, Privacy and security, Delivery, and Procedures (table 1). Examples of quality indicators in the Source hyperdimension are: contact information of the NSI, effect on response burden, costs of data source use, data source delivery arrangements, communication of planned changes and dependency risks. Mainly qualitative methods are present in this hyperdimension, only a few quantitative methods occur. In table 1 the dimensions, quality indicators, and measurement methods for the Source hyperdimension are listed.

2.2.2 Metadata hyperdimension

The Metadata hyperdimension specifically focuses on the conceptual metadata of the data source. Clarity of the definitions and completeness of the meta information are some of the quality aspects included. The Metadata hyperdimension is composed of four dimensions: Clarity, Comparability, Unique keys, and Data treatment by data source maintainer. Examples of quality indicators in the Metadata hyperdimension are: clarity of the population definition, time differences between the reporting periods of the NSI and data source maintainer, presence of unique identification keys, and data checks performed by the data source maintainer. The Metadata hyperdimension solely contains qualitative methods. In table 2 the dimensions, quality indicators, and measurement methods are listed for the Metadata hyperdimension.

2.2.3 Data hyperdimension

The Data hyperdimension focuses on the quality aspects of the data in the data source. This hyperdimension solely contains accuracy related quality aspects. The Data hyperdimension is composed of nine dimensions: Over coverage, Under coverage, Linkability, Unit non-response, Item non-response, Measurement, Processing, Precision, and Sensitivity. The dimension Precision was added mainly because of its application for surveys (see section 2.3). Examples of quality indicators are: units not belonging to the population, correctly coupled units, missing values, and measurement error. The Data hyperdimension almost exclusively contains quantitative methods. In table 3 the dimensions, quality indicators, and measurement methods for the Data hyperdimension are listed.

A considerable part of the measurement methods in the Data hyperdimension are based on a so-called Representative index (R-index). The R-index is a concept that has been developed by Statistics Netherlands (Schouten and Cobben 2007). R-indexes measure the extent to which the composition of the units in a data source, at a certain point in time, deviate from the population. For surveys this is a familiar concept. Here, representative means that all units in the population have the same probability of responding to the survey request. Representative is, however, also important for administrative data because the composition of the units present in the data source may be time-dependent. In the Netherlands, for instance, the composition of the companies that provide VAT-data to the Dutch Tax Office varies during the monthly collection period. This affects the quality of the data that is provided to Statistics Netherlands.

2.2.4 Process hyperdimension

The Process hyperdimension is quite different in comparison to the other three. In the Process hyperdimension the focus is not on the data source itself but on the processing of the data source. Therefore, it was excluded in our initially studies and is not discussed in this paper. It is, however, a subject that is under investigation in our office.

2.3 Quality framework and survey data

The overall quality framework constructed is shown in tables 1 through 3. For each hyperdimension a separate table is created that shows its dimensions, quality indicators, and methods of determination. Although the quality framework was originally developed for administrative data, it also interested the authors to see if it could be used for the evaluation of survey data as well. The initial reason for doing this was to see if the framework could be used to determine the quality of survey data collected by an organization other than Statistics Netherlands. Evaluation results indicated that this could indeed be the case. To enable the use for survey data it was, however, required to add some survey specific quality indicators; such as those included in the Precision dimension (table 3). The time-dependence of the population composition in administrative data is another, non-survey specific, reason for doing this (section 2.2.3). In the Source and Metadata hyperdimension only the wordings of some of the measurement methods had to be adjusted to enable its use for survey and administrative data sources.

There are additional advantages of extending the administrative data quality framework to that of surveys. The most important one is the fact that this enables the use of a single framework for the determination of the quality of the two most important data sources used for the production of statistics in our office. Currently more detailed evaluation studies are performed to verify this initial finding.

2.4 Use of the quality framework

While evaluating a (potential) data source, the user of the quality framework must first determine the results for the quality indicators in the Source hyperdimension. Subsequently, the quality indicators in the Metadata and Data hyperdimension need to be evaluated. The authors have developed a checklist for the evaluation of the first two hyperdimensions. This approach could not be followed for the Data hyperdimension.

When the results for some of the quality indicators in a hyperdimension reveal problems, these have to be sorted out before the next hyperdimension can be evaluated. If these problems cannot be solved, the evaluation of the data source must be stopped and it has be concluded that the data source cannot be used for the statistics the user had in mind. If the user wants to evaluate the data source again but with another (new) statistical use in mind, the same sequence of events should be repeated. However, if the problems for that data source occurred in the Source hyperdimension it is to be expected that the data source can also not be used.

If the evaluation of the last hyperdimension, Data, is successful, the data source can be used for the production of statistics. It is conceivable, however, that the user would like to perform one or more additional -very specific- checks after the evaluation of all hyperdimensions. The additional checks all occur at the data level; i.e. in the Data hyperdimension. An example of such a specific check is a comparison of the estimated percentage of unemployed persons obtained, after editing and weighting, from an administrative data source (such as the Job-seeker administration of the Centre of Work and Income in the Netherlands) with that of the estimated percentage obtained through the Labour Force survey of Statistics Netherlands. Since the quality framework was created to be used at a general level, it only contains general applicable quality indicators. Very specific check were not included, simply because it is impossible to include all possible conceivable checks. Different users of a data source may have different population parameters in mind that pose different quality constraints. Necessarily, the quality framework has to be restricted to some extent as it is impossible to meet all conceivable uses. The necessity to restriction is especially applicable to the quality indicators of the Data hyperdimension.

2.5 Future work

Future studies will focus on the usability of the quality framework on both administrative and survey data. In these studies the checklists constructed for the Source and Metadata hyperdimensions will be evaluated. For the Data hyperdimension another approach will be followed; one of the options is to include the measurement methods in a specific software program. Various registers and surveys on persons and businesses will be evaluated. The study of the quality aspects of the Process hyperdimension is also a topic for future work.

Table 1: *Quality framework for data sources, Source hyperdimension*

DIMENSIONS	QUALITY INDICATORS	METHODS
1. Supplier		
	1.1 Contact	- Name of the data source - Data source contact information - NSI [a] contact person
	1.2 Purpose	- Reason for use of the data source by NSI
2. Relevance		
	2.1 Usefulness	- Importance data source for NSI
	2.2 Envisaged use	- Potential statistical use of data source
	2.3 Information demand	- Does the data source satisfy information demand?
	2.4 Response burden	- Effect of data source on response burden
3. Privacy and security		
	3.1 Legal provision	- Basis for existence of data source
	3.2 Confidentiality	- Does the Personal Data Protection Act apply? - Has use of the data source been reported by NSI?
	3.3 Security	- Manner in which the data source is send to NSI - Are security measures required? (hard- and software)
4. Delivery		
	4.1 Costs	- Costs of using the data source
	4.2 Arrangements	- Are the terms of delivery documented? - Frequency of deliveries
	4.3 Punctuality	- How punctual can the data source be delivered? - Rate at which exceptions are reported - Rate at which data is stored by data source maintainer
	4.4 Format	- Formats in which the data can be delivered
	4.5 Selection	- What data can be delivered? - Does this comply with the requirements of NSI?
5. Procedures		
	5.1 Data collection	- Familiarity with the way the data is collected
	5.2 Planned changes	- Familiarity with planned changes of data source - Ways in which changes are communicated to NSI
	5.3 Feedback	- May NSI contact data source maintainer is case of trouble? - In which cases and why?
	5.4 Fall-back scenario	- Dependency risk of NSI - Emergency measures when data source is not delivered according to arrangements made

[a)] NSI, National Statistical Institute

Table 2: *Quality framework for data sources, Metadata hyperdimension*

DIMENSIONS	QUALITY INDICATORS	METHODS
1. Clarity		
	1.1 Population definition	- Clarity score of the definition
	1.2 Definition of variables	- Clarity score of the definition (and categories)
	1.3 Time dimensions	- Clarity score of the definition
	1.4 Geographic demarcation	- Clarity score of the definition
	1.5 Definition changes	- Familiarity with occurred changes
2. Comparability		
	2.1 Population definition comparison	- Comparability with NSI definition
	2.2 Variable definition comparison	- Comparability with NSI definition
	2.3 Time differences	- Comparability with NSI reporting periods
	2.4 Geographic differences	- Comparability with NSI reporting area
3. Unique keys		
	3.1 Identification keys present	- Presence of unique keys - Comparability with unique keys used by NSI
	3.2 Unique combinations of variables present	- Presence of useful combinations
4. Data treatment by data source maintainer		
	4.1 Checks	- Population unit checks performed - Variable value checks performed - Combinations of variable values checked - Extreme value checks
	4.2 Modifications	- Familiarity with data modifications - Are modified values marked and how? - Familiarity with default values used

Table 3: *Quality framework for data sources, Data hyperdimension*

DIMENSIONS	QUALITY INDICATORS	METHODS
1. Over coverage	1.1 Non-population units	- Percentage of units not belonging to population
2. Under coverage	2.1 Missing units	- Percentage of missing population units
	2.2 Selectivity	- R-index [a] for population composition
	2.3 Effect on core variables	- Maximum bias of average for core variable - Maximum RMSE [b] of average for core variable
3. Linkability	3.1 Linkable units	- Percentage of units linked
	3.2 Mismatches	- Percentage of units incorrectly linked
	3.3 Selectivity	- R-index for units linked
	3.4 Effect on core variables	- Maximum bias of average for core variable - Maximum RMSE of average for core variable
4. Unit non response	4.1 Units without data	- Percentage of units without data
	4.2 Selectivity	- R-index for unit composition
	4.3 Effect on core variables	- Maximum bias of average for core variable - Maximum RMSE of average for core variable
5. Item non response	5.1 Missing values	- Percentage of cells with missing values
	5.2 Selectivity	- R-index for variable composition
	5.3 Effect on variable	- Maximum bias of average for variable - Maximum RMSE of average for variable
6. Measurement	6.1 External check	- Has an audit or parallel test been performed? - Has the input procedure, e.g. questionnaire, been tested?
	6.2 Incompatible records	- Fraction of fields with violated edit rules
	6.3 Measurement error	- Size of the relative measurement error
7. Processing	7.1 Adjustments	- Fraction of fields adjusted
	7.2 Imputation	- Fraction of fields imputed
	7.3 Outliers	- Fraction of fields with outliers
8. Precision	8.1 Standard error	- MSE [c] of average value for core variable
9. Sensitivity	9.1 Missing values	- Total percentage of empty cells
	9.2 Selectivity	- R-index for composition of totals
	9.3 Effect on totals	- Maximum bias of average for totals - Maximum RMSE of average for totals

[a)] R-index, Representative index (explained in section 2.2.3); [b)] RMSE, Root Mean Square Error;
[c)] MSE, Mean Square Error

References

Biemer P. P., Lyberg L. E. (2003) *Introduction to Survey Quality*, Wiley, New Jersey.

Børke S., Bergstrøm Y. (2006) Electronic Data from Business to Government - Development with Implications on Administrative Registers?, *Seminar on Strategies for Social and Spatial Statistics*, Oslo.

Daas P. J. H., Fonville T. C. (2007) Quality control of Dutch Administrative Registers: An inventory of quality aspects, *Seminar on Registers in Statistics - methodology and quality*, Helsinki.

ESC (2007) Pros and cons for using administrative records in statistical bureaus, *Seminar on increasing the efficiency and productivity of statistical offices*, Economic and Social Council conference of European statisticians, Geneva.

Eurostat (2003a) *Definition of quality in statistics*, Assessment of the quality in statistics, Item 4.2: Methodological documents, Luxembourg.

Eurostat (2003b) *Handbook "How to make a Quality Report"*, Assessment of quality in statistics, Item 4.2d: Methodological documents, Luxembourg.

Eurostat (2003c) *Quality assessments of administrative data for statistical purposes*, Assessment of quality in statistics, Item 6, Luxembourg.

Eurostat (2005a) *European Statistics Code of Practice for the national and community statistical authorities,* Luxembourg.

Eurostat (2005b) *Standard quality indicators*, Quality in statistics, Luxembourg.

Everaers P. C. J., Van der Laan P. (2003) The Dutch System of Social Statistics: Micro-Integration of Different Sources, *Expert Group Meeting on Setting the Scope of Social Statistics*, United Nations, New-York.

Karr A. F., Sanil A. P., Banks D. L. (2006) Data quality: A statistical perspective, *Statistical Methodology*, 3, 137-173.

ONS (2005) *Guidelines for measuring statistical quality*, version 3.0, Office of National Statistics, United Kingdom.

Sæbø H. V., Byfuglien J., Johannesen R. (2003) Quality Issues at Statistics Norway, *Journal of Official Statistics*, 19, 287-303.

Schouten B., Cobben F. (2007) *R-indexes for the comparison of different fieldwork strategies and data collection modes*. Discussion paper 07002, Statistics Netherlands, Voorburg/Heerlen.

Statistics Finland (2004) *Use of Register and Administrative Data Sources for Statistical Purposes*, Handbook 45, Helsinki.

Thomas M. (2005) Assessing Quality of Administrative Data, *Survey Methodological Bulletin*, 56, 74-84.

UNECE (2007) *Register-based statistics in the Nordic countries – Review of best practices with focus on population and social statistics*. United Nations Publication, Geneva.

Van den Brakel J., Smith P., Compton S. (2007) *Quality procedures for survey transitions, experiments and discontinuities*, Discussion paper 07005, Statistics Netherlands, Voorburg/Heerlen.

Wallgren A., Wallgren B. (2007) *Register-based Statistics: Administrative Data for Statistical Purposes*, Wiley, Chichester.

4

Integration of registers and samples 1

Integrated use of register and interview data for EU statistics on income and living conditions (EU-SILC) at Statistics Finland

Veli-Matti Törmälehto
Statistics Finland
e-mail: veli-matti.tormalehto@stat.fi

Abstract: We describe the combined use of register and interview-based data in the Finnish implementation of EU-regulated Statistics on Income and Living Conditions (EU-SILC). The use of different data sources at Statistics Finland and the exploitation of registers in survey sampling, data collection, record linkage, processing, and estimation phases of the EU-SILC survey process are presented.

Keywords: EU-SILC, Registers, Survey

1. Introduction

The present paper deals with the combined use of register and interview data for EU Statistics on Income and Living Conditions (EU-SILC) at Statistics Finland. The EU-SILC is a multidimensional instrument with the objective of timely production of statistics and micro data on income, housing, labour, child care, social exclusion, health, education and demography. The EU-SILC is primarily used for the calculation of cross-nationally comparative indicators of monetary poverty, income inequality and living conditions.

The EU-SILC is implemented as ex ante output harmonised household sample surveys, governed by EU regulations and conducted annually by National Statistical Institutes in nearly all European countries. The EU regulations set up a common framework for the survey implementations. The regulations specifically take into account the possibility of combining survey data with registers. Consequently, the countries taking part in the EU-SILC can be classified into "register" and "survey" countries. The register countries currently include the Nordic countries, the Netherlands and Slovenia.

The paper describes the register-based implementation of the EU-SILC in Finland. It begins with a description of the EU-SILC contents and the available data sources. Issues concerning statistical units, reference periods and statistical variables in the combined use of register and interview data in the Finnish EU-SILC are then discussed. The fourth part examines the phases of the survey process: sampling and record linkage, survey data collection and processing, estimation and quality control. The paper concludes with a consideration of the advantages and disadvantages of the combined method used in Finland.

2. EU-SILC and the register infrastructure at Statistics Finland

2.1. Requirements of the EU-SILC

The EU-SILC regulation defines the *core areas* and within the core areas common definitions of *target variables*, i.e. a specification of the required output and ex ante measures to harmonise the output. For each target variable, the *statistical units*, *modes of data collection*, and the *reference periods* are specified. For example, it is defined that in the core area 'income', variable PY090G 'Unemployment benefits' which were received during the income reference period must be recorded for all adult (16+) household members in the sample either from personal interviews or from registers.

The regulation sets also the definitions and procedures in relation to the reference populations, sampling, the survey units, the tracing rules, weighting, quality control and so on. As long as these are followed, the countries have flexibility in collecting the primary data with different modes of collection: from survey respondents, administrative and statistical registers, as a combination of survey and registers, or constructed by modelling or imputation.

In addition to being multi-national and multi-dimensional, the EU-SILC is a complex household sample survey because both cross-sectional and longitudinal data must be collected at individual and household level. The longitudinal component is much more limited in content than the cross-sectional part. Both the cross-sectional and the longitudinal component require information which cannot be taken solely from registers[1]. Direct data collection from households is needed.

To meet both the cross-sectional and the longitudinal requirements, Eurostat has recommended a rotational design with four panels. Initially selected sample persons are followed for at least four years, and each year one panel is dropped out and a new one is substituted. Finland has a slightly non-standard solution since it has integrated the cross-sectional EU-SILC into its own long-running Income Distribution Survey (IDS). This survey has been conducted since 1977 and has a two-year rotating panel design. It has always combined registers with survey data and the subject areas are much the same as in the EU-SILC[2]. For a sub-sample, the panel length is extended to four years in order to collect the longitudinal EU-SILC information.

The targeted outcome of the EU-SILC is a set of output harmonised micro data sets. Statistics Finland collects and processes the EU-SILC data in Finland, and delivers edited, imputed, and weighted micro data to Eurostat. Eurostat calculates indicators from the micro data and publishes them on their web site for Finland and for all participating countries, and combines and disseminates the user micro data files (the EU-SILC user database) which in 2006 contain data from 27 countries, to researchers.

[1] When the EU-SILC was planned in Finland, a register-based survey, i.e. a survey where all data are derived from registers without any own data collection was considered for the longitudinal part. It was deemed to be unfeasible due to the data requirements and conceptual problems with the statistical variables available in the registers.

[2] In this paper, the abbreviation EU-SILC is used for the integrated IDS/EU-SILC, and both the EU-regulated and the national contents of the survey are considered.

2.2. The EU-SILC core areas and data sources at Statistics Finland

Figure 1 below gives an overview of the EU-SILC core areas and data sources in the Finnish implementation. In Finland, both the cross-sectional and the longitudinal EU-SILC is now compiled from data coming from the *Survey on Income and Living Conditions*, by collecting and processing data from several *income registers*, and by record linkage of register data from the *entirely register-based statistical systems* of Statistics Finland. The EU-SILC is a separate statistical system operated in an MS SQL Server database, and the unique *personal identification number* (PIN) is the primary link variable. The informed consent principle is followed in the survey part, i.e. every respondent is informed that register data on them will be merged with the information they give in the survey. The survey data collection is conducted mostly with computer assisted telephone interviews (CATI) by the professional interviewers of Statistics Finland.

Statistics Finland has a well-functioning, co-ordinated system of statistical registers and Finland has drawn its population censuses entirely from registers since 1990. The register system consists of the Population Statistics Database, education-related registers, the Business Register, and income registers. The Population Statistics Database includes sub-systems on families, dwelling units, buildings and dwellings, and employment. Income is a focus area in the EU-SILC, and most of the income data in Finland is retrieved from personal income registers (tax register and several others; more than 10 registers overall) which are part of the register-based statistical system at Statistics Finland.

Figure 1: *Overview of the EU-SILC core areas and the data sources at Statistics Finland*

EU-SILC CORE AREAS							
HEALTH	SOCIAL EXCLUSION	HOUSING	CHILD CARE	LABOUR	INCOME	EDU CATI ON	BASIC DATA

DATA SOURCES AT STATISTICS FINLAND				
Survey on Income and Living Conditions Blaise questionnaire CATI/CAPI	Business Register (Sybase)	Income registers (SAS)	Population Database (MS SQL Server): population, families, household-dwelling units employment	Register of Completed Education and Degrees; register of educational institutions

ADMINISTRATIVE DATA		
Tax Administration	Population Information System of the Population Register Centre: Persons Buildings and dwellings	Employment registers
Pension and social security administration (pensions and social security, employment history)		Job-seeker register
		Other registers

The units at Statistics Finland responsible for the entirely register-based statistical systems have combined and processed the data from several administrative sources, and created statistical variables which may be linked to other statistical systems with the personal identification number (or a surrogate number), the domicile code, and the business code. The Population Database is the backbone of all social statistics. The coverage of individuals in the Population Database is essential for the representativeness of the EU-SILC sample while the link between persons and dwellings greatly facilitates the record linkage to survey data.

The primary data of the Population Database are based on the population information system on persons, buildings, and dwellings. Every person resident in Finland has a unique standardised 11-digit *personal identification number* (PIN). Each person must be registered in the municipality where he/she has a permanent place of residence. The link between a person and his/her permanent dwelling/address is the unique standardised *domicile code*. Even the homeless are registered in municipal registers but without information on an address[3]. (Statistics Finland, 2007a).

The unit responsible for the EU-SILC is responsible for the income register files. The quality control of the income registers is conducted on the EU-SILC sample data. The quality controls of the registers are easier to do with sample data than with complete registers. This method of using a sample for the quality control of registers is also endorsed in Wallgren & Wallgren (2007).[4]

The use of the Business Register and the *business identification code* is less extensive in the EU-SILC. Tax data are used in the EU-SILC to connect a person to the unit that pays his/hers wage or salary (identified with the business code) and further to the statistical variables (industry and sector) in the Business Register.

3. Units, reference periods and statistical concepts

The legislation in Finland allows the combining of registers with interview data and record linkage variables exist in the registers and for the sample. In addition, combined use of survey and register data needs to take into account differences in statistical units (objects), reference periods, as well as the processing of primary data into variables which satisfy the statistical concepts. Furthermore, the stability and timeliness of the registers are important issues for punctual production of the EU-SILC.

3.1 Units

The units of analysis in the EU-SILC are individuals and households[5]. While the units of analysis at the population level in the survey and register countries are the same, the

[3] The quality of the domicile code is studied regularly in connection with the Labor Force Survey.

[4] With regard to income data, the register data may often be edited in the administrative process, e.g. to assess final taxes for an individual. A distinction relevant for income data is whether the data come from administrative sources or whether they are self-declared to the register authority. The data on wages and salaries, social transfers, and taxes are very accurate as they are based on administrative data which is transmitted electronically to tax authorities. The data on e.g. self-employment income and rental income are largely self-reported by the recipient to tax authority, and obviously may contain higher degree of measurement error due to tax avoidance or hidden economy than other types of income.

[5] In the labour domain, the analysis is restricted to persons aged 16+, while in the child care domain, the analysis concerns children aged 0-12 years.

collection units and the units of analysis at the sample level need not be the same. This is elaborated on in Table 1 which is taken from Eurostat's EU-SILC documentation and presents the survey units for sampling, analysis and data collection in the EU-SILC.

Table 1. *Survey units for sampling, analysis, and data collection in EU-SILC (Source: Eurostat, EU-SILC User Database description)*

Sampling unit		**Analysis units**	**Collection unit/source**	
Selected	constructed		'survey country'	'register country'
Address Or Household Or *Person* (aged 16+)	*Household*	Set (a): household	Household respondent (HR)	Registers +HR
		Set (b): all household members	Household respondent*	Registers +HR
		Set (c): household and personal income and basic variables	Personal interview (all members 16+)	Registers (all members 16+)
		Set (d): detailed variables		
		All members 16+	Personal interview**	
		Selected respondent		Personal interview

* Combined with set (a) household interview ** Combined with set (c) personal interview

The main difference between the survey and register countries lies in *the collection unit of the individual level variables*. When personal income data and basic variables can be taken from registers, as is the case with Finland, the regulation states that only one person per household needs to be interviewed for "detailed" personal variables. These persons must represent the population of individuals, which is the case with the "register countries" when they sample persons ("*selected respondents*") from the frame and construct households around the selected persons. The selected respondents represent the population of individuals 16+ and special selected respondent weights are applied in the analysis of data collected at selected respondent level only. In the "survey countries", a sample of addresses (households) is drawn and all current members aged 16 years and older must be interviewed.

The household level variables and child care variables are collected either from a household respondent by interviews[6] or from administrative registers. Some register variables on income are household-level variables although they appear to be individual-level variables, i.e. they are registered to only one PIN and thus to only one household member. Housing subsidies, income support and child allowances are typically these kinds of variables.

In the registers, individuals are identified with PINs, and all persons sharing the same domicile code in the register constitute *dwelling-units*, i.e. households defined solely according to the co-residence criterion. The register definition of a household is not sufficient for the EU-SILC where sharing of resources among household members should be the decisive criterion and comparability of the units in sub-groups of the population is important (e.g. economic situation of students in Finland compared to students in Italy).

The linkage between identification PINs and domicile codes enables pre-entry into the EU-SILC questionnaire of all persons permanently registered at the same address

[6] Mail survey is used in Germany.

(dwelling-unit) before the interviewer contacts the household. These pre-entries are then corrected in the interview, i.e. the register-based dwelling unit is the basis for establishing the surveyed *economic household* members.

3.2 Reference periods

The EU-SILC regulation has several reference periods depending on the target variable. In the questionnaire of the Survey on Income and Living Conditions, the reference periods have been set taking into account the reference periods in the registers. This is important for the construction of variables which utilise both register and survey data, for quality assessments and benchmarking, and for consistency editing of register and survey-based variables.

Obviously and most importantly, the analysis of data coming from many sources requires coherence of reference periods. For example, when the employment status of population below the low income threshold is analysed, the variable on employment status and income data should refer to the same time period as far as possible (see also Debels & Vandecasteele, 2008).

The data in registers usually refer either to a fixed point in time or to a certain period of time. In the register sources used for the EU-SILC, the reference times are almost always either the end of the year (e.g. age, citizenship, marital status, debt amounts, current activity status) or the calendar year (e.g. income, main activity status)[7].

In the EU-SILC regulation, the reference periods *current* and *income reference period* are relevant to the combined use of register and survey data in Finland. The survey data collection for year N is conducted from January to May in N+1. Year N is always the income reference period because that is the income reference period in registers. In the questionnaire, the reference period "current" is interpreted as "at the time of the interview" in current education, health, and social exclusion domains whereas it is interpreted as "the end of last year" in domains where register variables may be used.

For example, the current activity status asked in the questionnaire refers to the status last December rather than at the time of interview to increase the consistency of labour variables with the register data, including income. The same applies to household composition, main residence, housing costs as well as dwelling characteristics. For national use, labour variables are usually defined in relation to main activity during the whole income reference period, e.g. person is unemployed if he/she was unemployed for six months or more during the last calendar year, rather than being unemployed in December, or as being unemployed at the time of the interview.

The harmonising of reference periods in the survey and in the registers leads to retrospective questions and quite probably to more recall errors. To dampen this effect, the fieldwork period is set early in the year. Furthermore, consistency checks with registers and subsequent editing can be used to improve data quality.

[7] Other reference periods do exist (e.g. being registered at an educational institution in September) or it may be possible to construct variables with different reference periods from the underlying administrative data (within year periods).

3.3 Statistical variables

Income and basic demographic data are the EU-SILC core areas where most of the target variables are based direct on registers. The quality of these data, especially those on income, is regarded as being far better than interview data in Finland, and the level of detail allows for a wide variety of income concepts to be constructed (e.g. different kinds of detailed classifications of social transfers received). Some income variables are constructed from both register and interview data, since supplementary income data still need to be collected in the interviews. In addition, register data may be used to construct the required sub-categories for interview-based variables more reliably.

In the social exclusion, health, child care and labour domains data cannot usually be derived from registers because they do not exist, or exist only partially. For example, in the social exclusion domain, data on enforced lack of durables must be collected. Data on the ownership of certain durables, such as cars could be linked from registers. For the EU-SILC, one needs to find the reason for not having a car, and therefore two questions on owning a car and a follow-up question of reason for not owning a car need to be asked. This is both psychologically wise and cost-effective as there is no need for further consistency editing.

In some domains, data are available from registers but it is either not used or used only as auxiliary information in the questionnaire, editing, imputations, or calibration. There are instances where a register variable cannot be transformed to the required statistical target variable because of conceptual differences or coverage problems, lack of detail in registers, or for the sake of internal consistency or cost-effectiveness considerations, and finally for timeliness considerations.

Let us consider, for example, one of the key variables in the EU-SILC, the variable on self-defined current economic status (PL030). For this variable, the regulation stipulates the reference period as "current", the units as "all current household members aged 16 and over", and the mode of collection as "personal interview, proxy or registers". The variable has 9 categories. It is based on four survey questions in order to divide the population into employed, unemployed, pensioners etc., with supplementary information on the category "permanently disabled" based on register data on disability pensions.

In the register-based employment statistics, a variable on activity status at the end of year is constructed from several registers (about 20 altogether). It is one of the key variables of the register-based statistical system and has eight categories. Since the reference period "current" is interpreted as end of year in the Finnish EU-SILC, and the register data are annually available for the whole population, it could be linked to the EU-SILC sample.

Even after disregarding the term "self-defined" of the definition of the EU-SILC variable, the register variable on activity status could not be used for the EU-SILC because the required categories are more detailed than the register variable categories including, for example, part-time work. Besides that, a question on current activity is necessary in any case for filtering the questions on e.g. number of hours worked, reason for part-time work and other employment variables. This is, in fact, cost-effective: it reduces the respondent burden of several labour variables and decreases the amount of consistency editing at the processing phase. Finally, the timeliness of register-based employment statistics would cause problems. For example, the results from the 2005 register-based employment statistics were published in spring 2008 while the 2006 EU-

SILC data (referring to year 2005) were finalised almost one year earlier and delivered to Eurostat in June 2007.

In a similar fashion, extensive data in the housing domain are collected by the survey although data on a number of dwelling characteristics (floor area, tenure status) exist in registers and can be linked to the survey data. However, the required target variable of housing costs is not available and it makes sense to ask about housing characteristics on the questionnaire and condition housing cost questions according to tenure status (owner/tenant) and dwelling type (detached, block of flats, etc.). The data on dwelling characteristics in registers are linked with PINs to the EU-SILC database and used in edits, imputations and logical checks.

4. The phases of the survey process

In this section, we go through the use of registers in the different phases of the EU-SILC process at Statistics Finland; sampling, record linkage, data collection, processing, estimation, and quality control.

4.1 Sampling frame and sampling

The sampling frame for the EU-SILC is created by record linkage of the population database data of individuals to their dwelling units (indicated by domicile code), and further linkage to the previous year's tax file using PINs. The contribution of frame imperfections to total survey error is negligible because of the high quality of the population information system on the population and dwellings: the coverage is excellent and the system is continuously updated. The sampling frame is used to feed information onto the electronic survey questionnaire and is occasionally also used for methodological studies (Monte Carlo simulations of planned changes to the sampling design).

The frame allows precise definition of the target population which is defined as private households residing in the territory of Finland at year-end. Individuals in non-private households (e.g. permanently in hospital, prison, collective housing, abroad, etc.) at year-end are excluded from the frame using the domicile code (last three digits begin with 9, the so-called '900'-group).

The frame is sorted geographically by domicile code. Tax data on the previous year's income source and income class are used to create the stratification variable. The frame is updated with end-of-year population database information to correct for the slight frame over-coverage of the initially selected sample. The final sample of register-based household dwelling-units, i.e. the selected persons and the members of their dwelling units, is then selected and a file is created for the electronic Blaise questionnaire. This file contains basic information on the sample units from the frame, information from the previous wave for waves 2-4, and even some information derived from statistical registers (degree from the register of completed education).

The frame information is fed onto the Blaise questionnaire in the first wave of interviews, while for the consequent waves (2-4) interview data from the previous wave is fed forward. The register data used from the frame are location, personal demographic data, education, and dwelling unit composition. Register data on personal demographics (age, sex etc.) are used as target variables but also for filtering to bypass questions on the Blaise questionnaire. For example, register data on completed education are also used to bypass questions on education.

4.2 Record linkage

Since all units in the frame have a valid PIN, every sample unit selected from the frame must have a valid PIN. The register on dwelling-units, or individuals registered at the same address, is important for the record linkage because the link variable is readily available for a large majority of the sample, i.e. for all selected persons and all members of the selected persons' dwelling-units. A valid PIN may only be missing for new members added to the household in the interviews.

Inglic (2007) describes the experience in the Slovenian SILC which is a register-based implementation but without a register on households and dwellings, i.e. without the link between PIN and address/domicile code. Consequently, much more work on the PINs must be conducted because when a sample of persons is selected their PINs are known, but PINs need to be constructed for all household members. This is done by collecting personal data on each member in the interviews (name, surname, birthday and gender), and by then applying automatic (85% of cases) and manual (15% of cases) searching from the population register for the PINs.

In Finland, it is up to the survey team to make sure that *new* members added to the household roster in the interviews have a valid PIN. This reduces the dimension from thousands to hundreds[8]. For example, in the sample of the first wave of the EU-SILC 2007 there were 14,433 individuals of whom 213 had been added to the household roster in the interviews. Roughly 2/3 of these had not reported a complete PIN and these had to be retrieved manually from the population register by name or date of birth. Without the register on dwelling-units it would have been necessary to search PINs for approximately 6,000 individuals instead of 213[9]. Furthermore, there would not have been any PINs and consequently no register data on the individuals in the non-responding part of the sample which now can be used in the non-response analysis.

When a valid PIN cannot be found, a person reported as belonging to the household in the interview is treated as an incorrect entry by the interviewers and outside the scope of the target population. In the 2006 survey, there were 21 persons out of 27,000 for whom no valid PIN could be found. At the end of the day, a valid PIN is the necessary precondition for a respondent being selected into the sample and for any member to be included in the sample.

4.3 The questionnaire data collection

The use of registers leaves its mark on the EU-SILC questionnaire and consequently on the data collection of the survey. A major part of the EU-SILC concerns detailed income

[8] New members are added to the household roster in the interviews if the register-based dwelling-unit does not match the statistical concept of economic household, if the registered dwelling-unit is incorrect or not up-to-date, or if the interviewer or the respondent misunderstands the instructions and the household becomes incorrect. A valid PIN is asked in the interviews for any added member, but is not always given by the respondent. The PIN given for an added member in the interview may also be incorrect, although there are logical checks and signals in the electronic Blaise questionnaire to prevent this.

[9] Estimated as two-thirds of the 8,894 individuals in the first wave (14,443 individuals in successfully interviewed households minus 5,549 selected respondents).

data. When these data are derived from registers, the length of the questionnaire is significantly reduced. It is notable that this reduction may affect the mode of collection: telephone interviews have replaced personal interviews in most of the register countries. There are two variants of this: either the first wave uses personal interviews and the subsequent waves telephone interviews, or all waves are conducted by telephone. The latter is the case in Finland. In the survey countries personal interviewing is the dominant method.

The respondent burden is reduced with the use of registers because less information needs to be collected. This follows from two reasons: first, the number of variables/questions is reduced and, second, for some variables data need to be collected only for the selected person, not for all household members. As already noted, the survey countries need to collect detailed information for each 16+ member of a household, while in the register countries only one person per household needs to be interviewed, e.g. in the cases of the health status, access to health care, and some labour variables such as managerial status, type of contract, or industry. Unfortunately, it is evident that some variables that are required for all adult members can never be derived from registers, e.g. whether a person is actively looking for job (PL020) or is available for work within the next two weeks (PL025) must be asked from either every member or from some other household member (proxy answer).

While the use of registers reduces the respondent burden, the questionnaire itself may become more fragmented if "only the holes are filled in". With regard to the outcome or target variables, there will be more derived variables than with a traditional survey (where questions can just be asked according to a classification), and hence more derivation and workload in the processing phase.

There are valid research questions on *the effects of different modes of collection* on the international comparability of the data, both in terms of comparing register data to interview data, and interview data collected with personal or telephone interviews, or even by mail. For example, there is some evidence that income data based on registers may yield lower inequality and monetary poverty estimates than income data collected with interviews (Table 3). There may be negative consequences to comparability across countries if total survey errors are reduced within a country by allowing countries flexibility with regard to the data sources. However, these questions are beyond the scope of this paper.

Table 3. *Register and survey based estimates of inequality and monetary poverty in Finland 1995 and 2000.*

	1995		**2000**	
	Interviews	Registers	Interviews	Registers
Gini-index, %	23.8	22.6	26.5	25.1
At risk of poverty rate, % (50% of median income)	7.1	4.5	8.4	5.9

Source: Nordberg (2003).

4.4 Imputations

The EU-SILC framework regulation states that the files transmitted to Eurostat must be "*fully checked, edited and imputed in relation to income*". Because Statistics Finland uses registers for most income variables there is very little item non-response and hence little need for income imputations. The interview-based variables in other domains do suffer from item non-response to a varying degree[10], while the register-based variables may suffer from undercoverage.

The standard methods of item non-response imputation in the Finnish EU-SILC are hot deck imputation, regression and mean imputation, and in some core areas last value carried forward (for the longitudinal sample) or register imputation (e.g. missing dwelling characteristics may be imputed from previous wave or from registers). In specific cases, item non-response is treated by reweighting. The questions on subjective health, which are asked of selected respondents only and no proxies are allowed, suffer from large item non-response. For these variables, item non-response is compensated for by reweighting the personal weights of selected respondents (representing population aged 16 and over).

To cope with coverage problems in register data, information on missing data may be fed from a register onto the questionnaire and additional information can then be asked in the survey. For example, the education register contains missing data in the year when the highest level of education was attained (variable PE030) whereas the coverage of the highest ISCED level (PE040) in the same register is good and the data are considered highly reliable. Information on missing year is fed onto the Blaise questionnaire and the year is asked in the interview and used to correct for register undercoverage. The problems with register coverage are more eminent in the labour and education domains. Obviously, if the coverage of a variable is suspected to be poor in a register it is not used at all.

4.5 Micro editing

Incorrect and/or outlying values of interview variables are checked variable by variable and edited in most cases with deductive or mean imputation. Register variables are similarly checked variable by variable but the original register variables are never edited. Instead, a new correction variable is created if errors are found in the register data.

The main problem in editing that arises with the use of register data (income or other domain) concerns increased need for *consistency editing*, i.e. after integrating data from different sources and micro-editing of data variable by variable for incorrect/outlying values they need to be checked for consistency and possibly further edited or imputed to improve to the situation. The survey and register variables should be *consistent at unit level* within a domain regardless of the source of data. Consistency editing in practise means replacing incorrect interview data with register data or deductive imputation, or correcting flag variables to indicate "non-visible" item non-response.

For example, detailed information is needed on number of months in different activities (employed, unemployed, retired) during the last calendar year. A battery of questions is asked on the questionnaire to get this information. For several, but not all, activity types the months may be register-estimated using income data or by getting the actual time periods of receiving the income type concerned (e.g. unemployment benefit). New variables on activity months are created by editing the months reported in the interview with register-based months.

[10] The most problematic components are monetary variables, such as electricity bills and other housing costs.

Another example from the EU-SILC data on arrears and on debts serves as an example of the need for consistency editing arising from the item non-response that becomes visible due to the use of registers. Household debt amounts, which are needed for national purposes, are obtained by summing individual level tax register data over household members. For the EU-SILC, data on arrears with loan repayments must be collected at the household level; i.e. the household respondent answers questions on whether the household has had problems once, more often, or never. To restrict the respondent burden the latter question is conditioned in the Blaise questionnaire to only those who are in debt. Being "in debt" has to be asked first for filtering, but the amounts are not asked.

There are cases when there is positive debt amount in the tax register but the household respondent has not reported being in debt in the interview and has not answered the question on being in arrears. This situation is interpreted as measurement error with interview data, probably due to recall errors or lack of information (household respondent may not know whether other household members are in debt or in arrears). A new variable with additional category to indicate this kind of "non-visible item non-response" is created to the database. The estimated number of households who were not even asked the question was three per cent of all households with housing loan (Table 4) in the EU-SILC 2007. The standard item non-response appeared to be almost negligible.

Table 4. *Arrears with loan repayments 2006, households with housing loan. Interview variable and register-edited variable. Estimated number of households.*

Category	Interview response	Edited variable
In arrears, once	15,680	15,680
In arrears, more than once	17,453	17,453
Not in arrears	736,934	736,934
Non-response in interview	244	244
Debt in registers, question was not asked	-	30,772
All households with housing loan	770,311	798,807

4.6 Macro editing

Outlying values in register incomes may cause problems which need to be dealt with by *macro-editing*. The need for macro-editing (impact of individual cases on aggregates) is assessed mainly on the target variables on income. A complexity arises because a register-based measurement of income is "too accurate" and leads to non-representative results in a multi-dimensional sampling space. The accuracy of the measurement of very high incomes, for example dividends or incentive stock options, leads regularly to very influential observations, i.e. individual values affect significantly estimates of certain sub-groups or indicators even at the aggregate level (e.g. Gini-coefficient in Finland). The values, if they are found to be correct as they usually are because taken from registers, are retained. The data may have to be reweighted so that the sampling weight of the problematic observation is reduced.

All edits, imputations as well as weighting are now done with SAS programs. As of spring 2008, the SAS-based BANFF system for edit and imputation has been experimented with in the Finnish EU-SILC. BANFF is a collection of specialised SAS procedures developed at Statistics Canada which can be used to satisfy the edit and imputation requirements of a survey.

4.7 Weighting and calibration

The EU-SILC gives instructions on weighting and recommends integrated calibration as the method to reweight the design weights. This serves two purposes: it improves the coherency of the EU-SILC with other sources and may improve the accuracy of estimates. The weighting method is calibration of design weights to known marginal distributions. SAS macros CALMAR or CLAN are used depending on the circumstances.

The available register information enables particularly efficient calibration models. As noted by Verma (2007), it is essential that survey variables are strictly comparable to the auxiliary information used in calibration. This is the case when exactly matched register variables are used in the calibration.

In the Finnish EU-SILC, a weighting frame is first created by limiting the population in income registers to the target population of the sample (population in private households) the same way the sampling frame was created, with the end-of-year population data and exclusion of non-private households with the domicile code. Marginal distributions are then calculated and fed into the CALMAR or CLAN calibration programs.

The current calibration model includes a demographic part and an income part. In the demographic part, the design weights are calibrated to the distribution of population in age-sex categories, to the distribution of dwelling units, and to the geographical distribution at NUTS 3-level with the capital region separated. In the income part, the total amounts of main income components (wages and salaries, pensions, etc.) as well as the number of income recipients (unemployment benefits, pensions) are fixed.

The availability of registers improves the accuracy of the survey itself as well as its coherence with other statistics in the field (National Accounts, totally register-based income statistics). For example, a standard quality assessment is comparison of the total sums of income surveys to those of National Accounts. Using income totals in calibration in Finland ensures that any discrepancy is not due to sampling error but must be due to either conceptual differences or measurement problems with the variable in itself, and this only concerns a very small amount of non-register based incomes such as inter-household transfers. Consequently, the coherence of the Finnish survey with National Accounts is very good (Kavonius & Törmälehto, 2003).

All countries do not have the possibility to use registers in the calibration to the same extent. A study on the effects of different calibration strategies with the Finnish data was conducted by Ollila (2008). Simple demographic calibration models were compared to elaborated models which use heavily register-based income data as auxiliary variables. The efficiency of the estimation was clearly improved when more income information was used. Unfortunately, these models can only be used in the "register" countries and some variation exists between them as well. The most elaborated models seem to be in use in Denmark and Finland.

4.8 Quality control

The traditional use of register data for quality control is to use them for unit non-response analysis. Register data are available for both respondents and non-respondents and the selectivity of the non-response can therefore be assessed and adjustments made either to the survey process, e.g. sampling, fieldwork practises, or to the weight calibration model. As described in the previous section, the use of auxiliary information in calibration improves the accuracy of the estimates, i.e. the mean square error is lower because calibration increases efficiency and reduces bias.

The tremendous advantage of register data is the possibility to compare and validate the results of a sample with parameter estimates from the register source. The biases due to sampling, data collection and processing can be evaluated with register data because at least some population parameters are known for the sample estimates. The validity of register-based concepts can be examined as well because the concepts in the survey are more comprehensive than in the registers: for example, register-based income statistics rely on income available in registers while the EU-SILC survey fills the (few) missing components by asking the information in the interview.

Table 5 illustrates the comparison with the evaluation of bias in the sample estimates of monetary poverty rate and income inequality. The magnitude of sampling error and conceptual differences can be evaluated by record linking data from the Total Statistics on Income Distribution (TSID) to the SILC sample, and controlling for differences in survey and register definitions of income and households/dwelling-units (Epland & Törmälehto, 2007).

Table 5: *Assessment of the effect of the income and household definitions on low income and inequality indicators in 2005: SILC/IDS sample and Total Statistics on Income Distribution (TSID) in 2005.*

Definition	**(1)**	**(2)**	**(3)**	**(4)**
Data source	SILC/IDS sample	SILC/IDS sample	SILC/IDS sample	TSID (total)
Income concept	SILC/IDS (survey)	TSID (registers	TSID (registers)	TSID (registers)
Income receiving unit	Household	Household	Dwelling-unit	Dwelling-unit
n (persons)	28,039	28,039	28,039	5,178,562
N (persons)	5,175,503	5,175,503	5,175,503	5,178,562
Indicator				
At risk of poverty, %	12.8	13.6	14.0	14.1
Gini-coefficient	0.270	0.275	0.274	0.282

The difference between the sample estimate of poverty rate (12.8%) in the first column based on 28,039 observations, and the TSID "parameter value" (14.1%) in the fourth column based on nearly 5.2 million observations, is not explained only by sampling error. There is undercoverage of income in the TSID compared to the SILC because alimonies and other inter-household transfers are not in registers. This increases the poverty rate from 12.8% to 13.6%. (columns 1 and 2). The effect from different household definitions (economic household in survey/dwelling-unit in registers) is the change of poverty rate from 13.6% to 14.0%. In the Gini-coefficient, the effect of sampling error seems to be more important than in the monetary poverty rate (columns 3 and 4).

5. Summary and conclusions

The EU-SILC is a multi-dimensional survey where the combined use of survey and register-based data is explicitly taken into account in the EU regulation. The survey implementations in the member countries vary depending on the possibility to exploit register data, especially on income, and the countries taking part are divided into "register" and "survey" countries.

The register-based SILC implementation in Finland benefits from a shorter questionnaire, less need for asking sensitive questions on income and, consequently, from lower respondent burden compared to survey-based implementations. The data quality and the accuracy of estimates based on income data are markedly improved because of low or non-existent item non-response and effective calibration models which use extensively auxiliary information from the registers. There is less measurement error related to interviewers, respondents, and their interaction when data are taken from registers, given that the register variables are of sufficient quality in terms of validity and coverage. The EU-SILC sample is used for quality controls of income registers at Statistics Finland, and in quality assessments of entirely register-based income statistics.

In addition to estimation, register data on household composition and personal demographics and location are fed from the sampling frame onto the electronic survey questionnaire to improve data quality, and to reduce the respondent burden and processing after the fieldwork period. Non-response analysis and quality assessment are greatly improved because much information is available for the non-respondents and because sample estimates may be compared to population parameters with common register variables.

There are certain negative aspects as well because the survey questionnaire needs to be adjusted to the data used from registers. The questionnaire becomes more fragmented and more complicated to program, it also may be less comprehensible to the respondent and the interviewer. Variables collected from different sources need to be consistent at the unit level and this increases the need for consistency editing of the data. Typically, there will be more derived variables compared to a pure survey and this further complicates the production process. The costs in the processing phase are reduced by less need for imputations and increased by more need for register estimation and consistency editing.

There is some indication that the "living conditions" variables of the EU-SILC have higher item non-response rates in the register countries. The use of registers seems to dictate the mode of collection which is usually CATI in the register countries while personal interviews dominate in the survey countries. In addition, the selected respondent model of the register countries may lead to more proxy answers in certain variables which need to be obtained by interviewing other household members besides the selected respondent.

Nevertheless, it is fairly easy to conclude that in the Finnish EU-SILC the total survey error is greatly reduced due to use of registers on income, population, buildings and dwellings, and businesses. However, there are core areas where the survey design is affected by the availability of registers and may result in less than optimal outcomes. The advantages of using registers clearly outweigh the disadvantages, even without consideration of data collection costs.

When considering the joint use of interview and register data, the correct strategy for the EU-SILC is perhaps to decide about the use of data source at a core area level rather than at a variable level, i.e. attempt to maximise the use of one source within a given core area (say, housing). This improves the survey questionnaire and reduces the workload in the processing phase which may cause unexpected costs.

The combining of survey and register data in the Finnish EU-SILC brings along better quality data and allows sufficient flexibility with regard to changes in register contents or coverage problems. The national part of the survey has run successfully since 1977 with most of the core areas of the current EU regulation (income, labour, housing) and without major breaks in the essential time series on income inequality and monetary poverty - in spite of several changes in the course of this time in the tax registers, the vast improvements in the population and register-based employment statistics, and changes in the other register sources.

References

Epland J., Törmälehto V-M (2007) From sample surveys to totally register-based household income statistics: Experiences from Finland and Norway. Paper prepared for the conference of the European Survey Research Association, Prague, 25-29 June 2007.

Debels A., Vandecasteele L. (2008) The time lag in annual household-based income measures, *Review of Income and Wealth,* 54 Issue 1, 71-88.

Inglic R. T. (2007) Administrative data and registers in EU-SILC. Paper presented at the Seminar on Registers in Statistics, Helsinki 21-23 May 2007.

Nordberg L. (2003) An Analysis of the Effects of Using Interview versus Register Data in Income Distribution Analysis Based on the Finnish ECHP-surveys in 1996 and 2000, Chintex Working Paper #15, Work Package 5, December 22 2003.

Kavonius I, Törmälehto V-M. (2003) Household income aggregates in micro and macro statistics, *Statistical Journal of the United Nations Economic Commission for Europe,* 20, 9-26.

Ollila P. (2008) EU-SILC: Impact Study on Comparability of National Implementations. Unpublished methodological report. Statistics Finland.

Statistics Finland (2007) Final quality report of the EU-SILC 2005 operation. Unpublished methodological report. Statistics Finland.

Wallgren A., Wallgren B. (2007) *Register-based statistics. Administrative data for statistical purposes,* Wiley, New York.

Verma V. (2007) Issues in data comparability in EU-SILC. Comparative EU Statistics on Income and Living Conditions: Issues and Challenges. Proceedings of the EU-SILC Conference (Helsinki 6-8 November 2006). Eurostat Methodologies and working papers.

Combining the tax and the survey data for the purposes of the short-term statistics production

Rudi Seljak
Statistical Office of the Republic of Slovenia
e-mail: rudi.seljak@gov.si

Abstract: At the Statistical Office of the Republic of Slovenia the first examinations concerning the possibilities of the broad use of the tax data in short-term statistics production began in 2005. At that time we carried out the feasibility study, which compared the monthly turnover indices regularly produced by the »classical« survey with the indices calculated from the tax data. On the basis of the results of the feasibility study we set up the new methodology where the turnover for most of the units is estimated from the tax data, and only for a small number of the largest units the data are still collected with the post questionnaire survey.

In the paper we present the new methodology with the special emphasis on the system for the combining of the survey and administrative data. We will explain the conceptual difference between both types of the data, point out what problems these differences could cause and explain what were the procedures used to solve these problems.

Key words: short-term statistics, administrative data, turnover indices

1. Introduction

The pronounced needs of the contemporary users of the statistical results, especially the needs of the political decision-makers, for quick and quality data on different aspects of the modern society reflect in the constantly rising demands for more and more efficient methods of data collection and statistical results production. The results should hence be produced quickly, with low costs and with high quality. Since these demands are most of the time contradictory, the statisticians are faced with the real problem of finding the balance between all these demands of different users.

The demand for quick production of results is especially outstanding in the case of the short-term business surveys which are by definition designed to provide results in a short period of time. Therefore, in this area we are constantly confronted with the challenge how to improve the timeliness of the statistical results with no significant influence on other quality components. One of the most popular and the most widely used ways, which would at least partly result in the realization of the above described goals, is the efficient use of the administrative sources. Many statistical offices have in recent years been exploring the possibilities of incorporation of these data in the different phases of the regular statistical processes. Since in the past the usage of these data was mostly limited to the process of the sampling frame construction, sample selection and sometimes the usage of the administrative data as the auxiliary variables in the estimation process, lately more and more surveys use these data also as a direct data source.

In the paper we present the new methodology for the estimation of the monthly turnover indices, which was at the Statistical Office of the Republic of Slovenia (SORS) introduced in the last three years and by which the data for most of the observed units are obtained from the administrative data source. These data are monthly provided by the Tax Authorities and were originally used for the monthly settlement of the value added tax. In 2005 we began to examine the possibilities of using these data for the purposes of the regular production of the short-term statistics. We firstly carried out the feasibility study for the wholesale activity group and on the basis of the results of this study the new methodology was set up. In the beginning of 2006 we started to regularly produce turnover indices for the wholesale, obtained by the new methodology. At the same time we started the feasibility study also for the field of other business services and in the beginning of 2007 the "new production" of the turnover indices started also for this field. In the beginning of 2008 the new methodology was then widened also to the retail trade activity as the final step of the "three step" introduction of the new methodology.

In the first part of the paper we will present the new methodology of selecting the set of the observational units. The main change of this part of the process is that we moved from the random sample to the cut-off sample selection. In the second part of the paper we will briefly describe the statistical processing of the data. This part could be easily denoted as the crucial part of the whole process, since we are here faced with a special challenge of merging and processing the data from the administrative source and from the "classical" statistical survey. In the last part of the paper we will point out the main advantages and also some deficiencies of the new methodology.

2. Selection of the observational units

Besides the introduction of the new data source, another very important change that came with the new methodology was the movement from random sampling to the cut-off sampling selection process. The main reason for that change was the fact that with the introduction of the exhaustive administrative data source the data for many units are now available with no additional costs and it's quite obvious that the use of the cut-off procedure should result in much more precise results than random sampling. On the other hand, also the tax data do not cover the whole population of interest. This is due to the fact that the units whose annual turnover is under a certain threshold are not obliged to report their data. In addition, some enterprises that are obliged to report are not obliged to report monthly but quarterly. Due to all these facts, one of the main tasks in our preparation phase was to set up the selection system carefully in order to obtain the target population which would cover a large part of the population of interest and would, according to the available data, lead to a sufficient response rate.

According to the results of the feasibility study, we decided that the target population will be updated twice a year, i.e. semi-annually. The whole procedure is carried out in two steps. In the first step the units of the target population are determined and then in the second step the units for which the data will still be obtained by the "classical survey" are selected. In the first step the units which will be included in the target population are determined by setting the fixed threshold. The threshold is determined by using the combination of the semi-annual turnover and the current number of employees.

The new methodology is mostly based on the usage of the administrative data, but for the smaller part of the largest units the data are still obtained by the post survey. The main reason for this decision was the intention to overcome the methodological differences in the definition of the turnover. We also wanted to keep the direct contact with the largest enterprises in order to more easily control the most important data and to control the demographic changes in the most important part of the population.

The units for which the data will be obtained by the "classical" survey are obtained by using the flexible threshold inside each of the activity groups. The goal is that the turnover of these units would exceed the certain share of the total turnover in each of the activity groups. The target share slightly differs between the activity groups but it is generally between 50% and 60%. In the concrete procedure the target population is firstly sorted by the descending turnover in each of the activity groups. Then so many of the largest units are selected that the share of the turnover of these units exceeds the target share of the turnover. The number of the selected units which are surveyed by the post questionnaire is approximately 3% of the whole target population.

It is of great importance for the efficiency of the whole system that we have an effective and transparent system for the management of the set of the observational units. Therefore we decided to set up and maintain the special database of the units which have ever been included in the set of the observational units. Hence, each unit which is included in the target population for the first time is also inserted into the database and remains there although it is at some point in time excluded from the target population. To enable easier management of the units, especially the management of the demography changes, to each of the units in the register a special 6-digit identification number is assigned. This identification can remain the same even if the business register's identification changes. The business register identification is in fact just one of the attributes in this register.

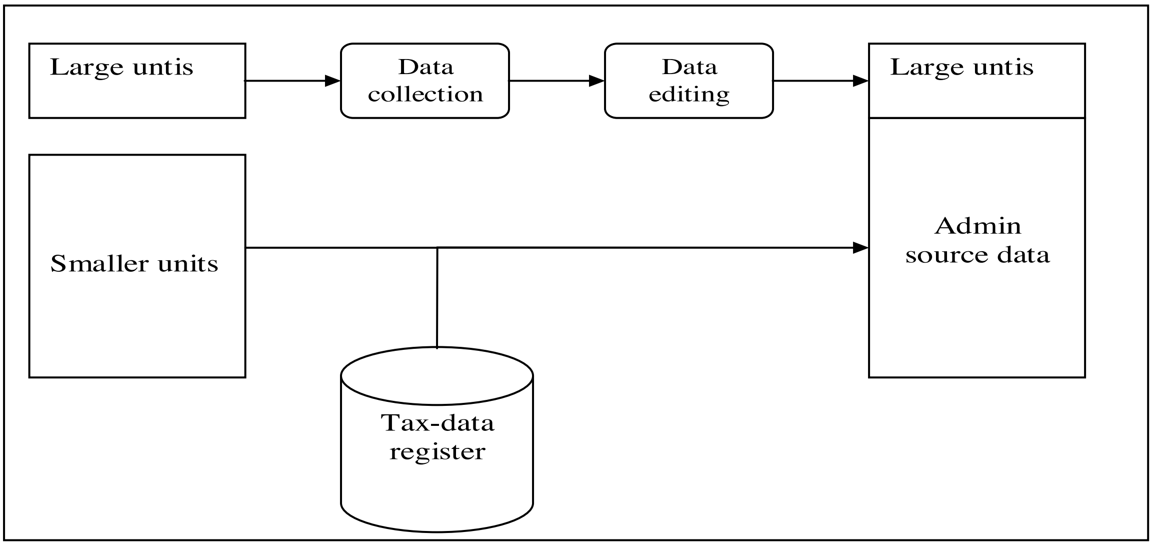
The main role of the database is to centrally store the principal data on the observational units. By using the graphical interfaces the survey manager can access and observe these data easily. The survey manager can also use the graphical interfaces to insert the changes of the data either manually or by using the special automated procedures which updates the data in the database with the reported data. By the manual procedures the survey manager can change the data on NACE code, address and the administrative identification.

3. Data collection and statistical processing

The use of the two different data sources is the main reason that the setting up of the whole statistical process was quite a challenging task. The set of the observational units is thus at the very beginning of the process split into two parts. For the small number of the largest units the data are still collected by using the "classical" post questionnaire which the reporting units should answer and send back by a certain date. These units represent 3% of the whole population in terms of the number of the units, but they cover more then 50% of the total turnover. For the remaining majority of the selected units the sum of the appropriate items from the tax authority's data is used as the estimation of the monthly turnover. In fact these estimates are not completely in line with the methodological definition of the turnover and one of the main goals of the feasibility study was to find out if they can serve well enough for our purposes. In the feasibility study we therefore simulated the new methodology for all the months of 2003-2005 and then compared the monthly turnover obtained by the new and the old methodology. It turned out that the level of the turnover from both sources can sometimes differ significantly but the movements, expressed in the form of the indices, were sufficiently coherent.

For the data that are collected by the "classical" survey, the data editing is also performed classically, meaning that each record is checked by using a set of logical controls. If the record is detected as "suspicious", the reporting unit is re-contacted by the telephone and if necessary the data are corrected. After the cleaning phase these data are then merged together with the data from the administrative database. To enable easier data processing and analyzing, each record in the merged dataset is assigned with the special status, whose values shows weather the data for the particular unit are coming from the survey or from the administrative source. The values of the status are assigned according to the standard 4-digit classification used at SORS. For instance, if the status for the particular unit has the code "21.17", it means that the data on turnover for this unit were obtained from the tax data. The first part of the process is presented in the following figure.

Figure 1: *First part of the statistical process*



4. Statistical data editing

As described earlier, the data from the "field survey" are edited classically using the predefined set of logical controls for the detection of the errors and the telephone re-contacts for the eventual correction of the data. For the units whose values are derived from the tax-data register, we are not allowed to contact the units and verify the correctness of their tax data. Therefore, in this case we are forced to employ some kind of the automatic data editing procedure.

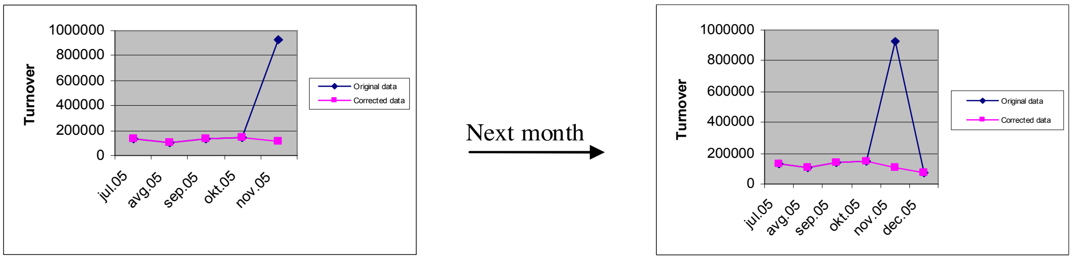
When we explored the historical administrative data for the purposes of the feasibility study, we detected some cases of sudden leaps in the time series of the turnover calculated from the tax data. As we found out through the study, most frequently the reason for such occurrence was the fact that the unit sold the real property. This purchase money was reported to the tax authorities but it shouldn't be included in the turnover. Hence, the main goal of the automated editing procedure was to detect and correct such cases of the overestimation of the monthly indices which could seriously distort the image of the observed phenomena.

The procedure is based on the well-known Hidiroglu-Berthelot method, designed for the cases of the periodical business surveys. With this method the distribution of the month-to-month turnover change is explored. In the first step the distribution is transformed in the way that the transformed distribution is symmetrical. In the second step the extreme values from the tails of the transformed distribution are detected as the outliers. These values are later in the process re-estimated by the imputation procedure. The procedure should be suitably parameterized and the tuning of the parameters was done during the feasibility study.

Since just one variable is the subject of the editing procedure, the procedure might at first sight seem quite straightforward. But since we are dealing with the monthly data, which additionally could be frequently revised, we had to set-up the system carefully to enable the correct and consistent procedure. Just for the illustration we present the two real-data cases.

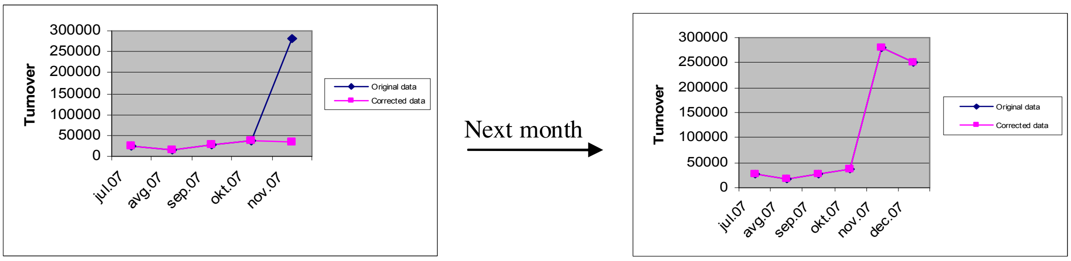
In the first case the turnover for the one particular unit in the second half of 2005 is presented. In the first chart the original and edited turnover from jul.07 to nov.07 is presented. As it is clearly seen from the line chart, the original November turnover significantly deviates from the level of previous months. In the second chart on the right, we added the dec.07 data. Since the December data returned to the "normal" level, all the previous data remained unchanged.

Figure 2: *Data editing – first case*



In the second case the data for the particular unit are presented for the second half of 2007. The image in the first chart is very similar as in the first case, where the original and the edited data fully correspond for the first four months and then we are confronted with the significant increase in the turnover and that's why the data were corrected. The difference comes with the December data. Since the original data still remain on the November's level, the system corrects also the November data to the original value.

Figure 3: *Data editing – second case*



According to the above indicated characteristics of the data editing in the case of the short-term statistics, the "classical" H-B procedure had to be adopted in this case. The basic rules of the implemented procedure are:

- Each month (M) the monthly turnover (T_M), calculated from the administrative data, is checked for all the previous twelve months.
- For the months (M-11)-(M-1) the distribution of T_M/ T_{M-1} (backward ratio) as well as the distribution of T_M/ T_{M+1} (forward ratio) is explored and the outliers according to the H-B method are detected. Only the values which are detected as outliers in both directions are designated to be corrected.
- For the current month only the distribution of the backward ratio is explored and the outliers in the distribution are designated to be corrected.
- All the data which were designated to be corrected are imputed together with the missing values in the next step of the process. Three different imputation methods are used. The first method is only used in the last month of each of the quarters for the units for which only the quarterly (and not monthly) tax data are available. The quarterly data are broken down to the monthly data by using the nearest neighbor principle. For the units for which the data from the previous month exist, the Historical Trend Method is used. Finally, for the units for which none of the above mentioned methods could be used, the Mean Value Method is used.

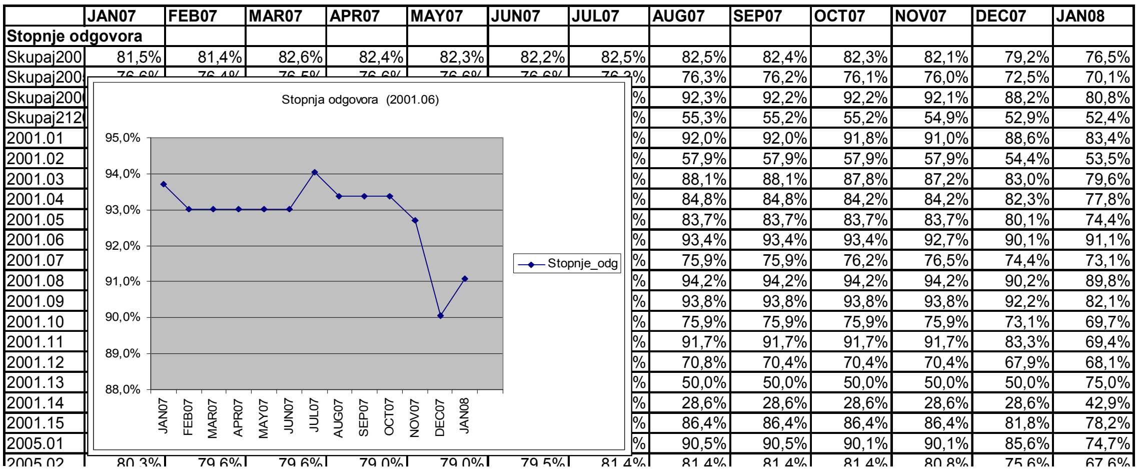
5. Quality indicators

In the last step of the process the data are aggregated into the form of indices and at the same time the set of the process quality indicators is calculated. All the quality indicators are calculated on the basis of the values of the metadata-variable called the status of the variable, which was already briefly introduced before. Besides the information on the data collection method, the status also contains information whether the data were corrected through the editing process or not and in the case they were corrected also the information on the imputation method used. For instance, if the status for the particular unit has the code "52.12", it means that the data on turnover for this unit were corrected through the editing system and the Historical Trend Method was used at the imputation stage.

Two types of quality indicators are calculated. The first group of indicators is called the micro and the second group the macro quality indicators. The micro indicators are the indicators which are calculated solely from the non-aggregated data. An example of such an indicator is the well-known response rate. The values of the indicators are always calculated and graphically presented for the previous 13 months. In the following figure the response rates as they could be seen by the survey manager are presented.

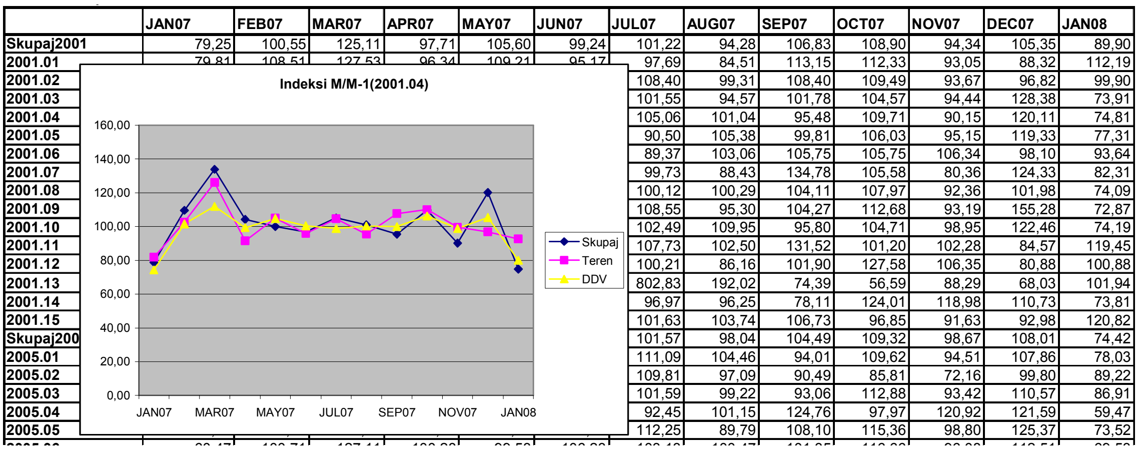
The macro indicators are the indicators calculated from the aggregated data (indices). An example of such an indicator is the relative difference between the index calculated from all the data and the index calculated from the non-imputed data. The data are presented in the tables as well as in the graphical form. The following figure presents the data for January 2008.

Figure 4: *Presentation of the response rates*



	JAN07	FEB07	MAR07	APR07	MAY07	JUN07	JUL07	AUG07	SEP07	OCT07	NOV07	DEC07	JAN08
Stopnje odgovora													
Skupaj200	81,5%	81,4%	82,6%	82,4%	82,3%	82,2%	82,5%	82,5%	82,4%	82,3%	82,1%	79,2%	76,5%
Skupaj200	76,6%	76,4%	76,5%	76,6%	76,6%	76,6%	76,3%	76,3%	76,2%	76,1%	76,0%	72,5%	70,1%
Skupaj200							%	92,3%	92,2%	92,2%	92,1%	88,2%	80,8%
Skupaj212							%	55,3%	55,2%	55,2%	54,9%	52,9%	52,4%
2001.01							%	92,0%	92,0%	91,8%	91,0%	88,6%	83,4%
2001.02							%	57,9%	57,9%	57,9%	57,9%	54,4%	53,5%
2001.03							%	88,1%	88,1%	87,8%	87,2%	83,0%	79,6%
2001.04							%	84,8%	84,8%	84,2%	84,2%	82,3%	77,8%
2001.05							%	83,7%	83,7%	83,7%	83,7%	80,1%	74,4%
2001.06							%	93,4%	93,4%	93,4%	92,7%	90,1%	91,1%
2001.07							%	75,9%	75,9%	76,2%	76,5%	74,4%	73,1%
2001.08							%	94,2%	94,2%	94,2%	94,2%	90,2%	89,8%
2001.09							%	93,8%	93,8%	93,8%	93,8%	92,2%	82,1%
2001.10							%	75,9%	75,9%	75,9%	75,9%	73,1%	69,7%
2001.11							%	91,7%	91,7%	91,7%	91,7%	83,3%	69,4%
2001.12							%	70,8%	70,4%	70,4%	70,4%	67,9%	68,1%
2001.13							%	50,0%	50,0%	50,0%	50,0%	50,0%	75,0%
2001.14							%	28,6%	28,6%	28,6%	28,6%	28,6%	42,9%
2001.15							%	86,4%	86,4%	86,4%	86,4%	81,8%	78,2%
2005.01							%	90,5%	90,5%	90,1%	90,1%	85,6%	74,7%
2005.02	80,3%	79,6%	79,6%	79,0%	79,0%	79,5%	81,4%	81,4%	81,4%	81,4%	80,8%	75,6%	67,6%

Figure 5: *Presentation of the macro indicators*



	JAN07	FEB07	MAR07	APR07	MAY07	JUN07	JUL07	AUG07	SEP07	OCT07	NOV07	DEC07	JAN08
Skupaj2001	79,25	100,55	125,11	97,71	105,60	99,24	101,22	94,28	106,83	108,90	94,34	105,35	89,90
2001.01	79,81	108,51	127,53	96,34	109,21	95,17	97,69	84,51	113,15	112,33	93,05	88,32	112,19
2001.02							108,40	99,31	108,40	109,49	93,67	96,82	99,90
2001.03							101,55	94,57	101,78	104,57	94,44	128,38	73,91
2001.04							105,06	101,04	95,48	109,71	90,15	120,11	74,81
2001.05							90,50	105,38	99,81	106,03	95,15	119,33	77,31
2001.06							89,37	103,06	105,75	105,75	106,34	98,10	93,64
2001.07							99,73	88,43	134,78	105,58	80,36	124,33	82,31
2001.08							100,12	100,29	104,11	107,97	92,36	101,98	74,09
2001.09							108,55	95,30	104,27	112,68	93,19	155,28	72,87
2001.10							102,49	109,95	95,80	104,71	98,95	122,46	74,19
2001.11							107,73	102,50	131,52	101,20	102,28	84,57	119,45
2001.12							100,21	86,16	101,90	127,58	106,35	80,88	100,88
2001.13							802,83	192,02	74,39	56,59	88,29	68,03	101,94
2001.14							96,97	96,25	78,11	124,01	118,98	110,73	73,81
2001.15							101,63	103,74	106,73	96,85	91,63	92,98	120,82
Skupaj200							101,57	98,04	104,49	109,32	98,67	108,01	74,42
2005.01							111,09	104,46	94,01	109,62	94,51	107,86	78,03
2005.02							109,81	97,09	90,49	85,81	72,16	99,80	89,22
2005.03							101,59	99,22	93,06	112,88	93,42	110,57	86,91
2005.04							92,45	101,15	124,76	97,97	120,92	121,59	59,47
2005.05							112,25	89,79	108,10	115,36	98,80	125,37	73,52

6 Conclusions

The introduction of the tax data as the main source in the production of the short-term statistics at the Statistical Office of the Republic of Slovenia started at the beginning of 2006. At that time the new methodology was introduced for the wholesale activity and then in the next two years the methodology was widened to the area of retail trade and other business services. Before the introduction of the new methodology the feasibility study was carried out. The results of this study showed that there are differences in the methodological definitions of the turnover obtained from the administrative data and the one obtained directly from the field survey, but the administrative data could be well used for the purposes of the "change estimation" in the short-term statistics.

By the new methodology only a small proportion of the largest units in the target population is still surveyed classically by using the post questionnaire. The main reason that these units are still surveyed classically is that we wanted to overcome the methodological differences in the definition of the turnover and wanted to remain in direct contact with the largest units.

The use of the administrative data as the main data source demanded considerable changes in the methodological and technical realization of the survey process. The main methodological change is the movement from random sampling to cut-off sampling. Technically the process is completely renewed. Most of the steps in the process are fully automated and could be executed and controlled by the survey manager. The special attention was focused on the editing system for the data originating from the administrative source. Since for these data we are not allowed to check their validity with the reporting unit, we set-up the automated data editing procedure, based on the well-known Hidiroglu-Berthelot method.

The greatest benefit of the new methodology is the response burden reduction: by the old methodology approximately 4000 units were surveyed, while now only approximately 400 units are requested to report their data every month. There is also considerable cost reduction from SORS's side since the material as well as human resources costs have been significantly reduced.

References

Hidiroglou M.A., Berthelot J. M. (1986) Statistical Editing and Imputation for Periodic Business Surveys, *Survey Methodology*, 12, 73-83.

Lyberg L. E., Biemer P., Collins M., De Leeuw E. D., Dippo C., Schwarz N., Trewin D. (1997) *Survey measurement and Survey Quality*, Wiley, New York.

European Commission (2006) *Methodology of short-term business statistics, Interpretation and guidelines*, Luxembourg.

Seljak R., Zaletel M. (2007) Tax Data as a Means for the Essential Reduction of the Short-term Surveys Response Burden, *International Conference on Establishment Surveys*, Montreal 2007.

SORS, Feasibility study on the use of administrative data in the wholesale activity, internal document.

Council Regulation (EC) No 1165/98 concerning short-term statistics amended by the Regulation (EC) No 1158/2005 of the European Parliament and of the Council.

Combining survey data and administrative sources for production of official Labour Statistics in Lithuania

Milda Šličkutė-Šeštokienė
Statistics Lithuania, Lithuania
e-mail: milda.slickute-sestokiene@stat.gov.lt

Abstract: Statistics Lithuania is now in the process of gradually moving to the extensive use of administrative sources. The usage of administrative sources for production of official Labour Statistics is described. The advantages and disadvantages are presented. The methodology of combining register and survey data to obtain statistical information of good quality is overviewed.

Keywords: administrative sources, register based statistics, earnings statistics, labour statistics.

1. Introduction

Statistics Lithuania has the full range of labour statistics that meet the demands of Eurostat and national needs. The challenge is to keep the quality and timeliness and to publish even more detailed information and at the same time to spare the costs and to diminish the burden for respondents. Those, contradictory to each other, challenges encouraged Statistics Lithuania to extensive use of register based information.

While the process leading to extensive use of administrative sources has already been set in motion, Statistics Lithuania is still at an early stage and a lot of work remains to be done. Some administrative sources are already successfully used while other administrative sources are still analyzing, trying to find out most suitable application.

Usage of administrative sources in order to diminish burden for respondents is one of the central goals of Statistics Lithuania. This goal is extremely supervised by authorities.

2. Administrative sources used for Labour Statistics

Administrative sources for production of official statistics in Lithuania were started to analyze in 2003. After the first agreements between register-holders and Statistics Lithuania were signed available administrative sources were started to analyze trying to create methodology based on the extensive use of auxiliary information. It took some time to understand of what register-based statistics are and to reorganize the usual way of working.

At this moment five administrative sources are used for production of official Labour Statistics:

- Register of Statistical Entities;
- Social Insurance Database;

- Tax Inspectorate Database;
- Register of Public Servants;
- Population Register.

The Register of Statistical entities and Social Insurance Database are the most widely used.

Administrative sources are used for following purposes:

- as auxiliary information at estimation stage;
- checking of primary data at micro and macro level;
- construction of frame;
- other purposes.

Usage of administrative sources at estimation stage (as auxiliary information) is most challenging and extremely time consuming task. At the moment two tasks related to the usage of administrative sources at the estimation stage are implemented:

- is being used for estimation of Annual Data of Earnings - Generalized Regression Estimator;
- is adopted to estimate the data of Individual enterprises - Linear Regression Model.

It is also expected to follow up analysis of administrative sources and to implement more tasks related to this field.

2.1. Annual Data of Earnings (Generalized Regression Estimator)

Until 2003 Annual Survey of Earnings (ASE) used to be carried out completely enumerating all enterprises. Since 2004 ASE was rejected and annual data are estimated on the basis of Quarterly Survey of Earnings (QSE) and data of Social Insurance (SI).

The main reason for the use of complete enumeration to get annual data is very detailed breakdown of the results. Such a breakdown does not allow getting reliable results using only the data of the sample. The breakdowns required for Annual Data of Earnings:

- NACE (two digits or sometimes even more detailed) & economic sectors (49 economic activities and 2 economic sectors);
- NACE (section level) & size of enterprise & economic sector (15 economic activities, 6 sizes of enterprises and 2 economic sectors);
- NACE (section level) & county (15 economic activities and 10 counties);
- Municipality (60 municipalities).

Total data for 49 x 2+15 x 6 x 2+15 x 10+60 = 488 partly overlapping domains are required. The sample size in 2007 is 6864 enterprises (15.2% of total number of enterprises). It is evident that it is impossible to get reliable data for such detailed breakdown using direct estimators. As the data of SI became available for the statistical purposes it was decided to use this data as auxiliary information in order to estimate parameters for domains.

Coefficients of correlation between key statistical variables and administrative variables are more than 0.9 almost in each economic activity and in each region. Such a strong correlation ensures that usage of administrative sources at estimation stage guarantees high quality of the results.

For estimation of the variables of interest Generalized Regression Estimator (GREG) is used. Data of SI is used as auxiliary information.

$$\hat{t}_w = \sum_{k\in\mathbf{i}} w_k y_k = \sum_{k\in\mathbf{i}} d_k g_k y_k$$

where

$\hat{t}_w$ - Generalized Regression Estimator (GREG) of variable of interest;

i - sample set;

w_k - calibrated weight of *k*-th sample unit;

y_k - variable of interest of *k*-th sample unit;

d_k - design weight of *k*-th sample unit;

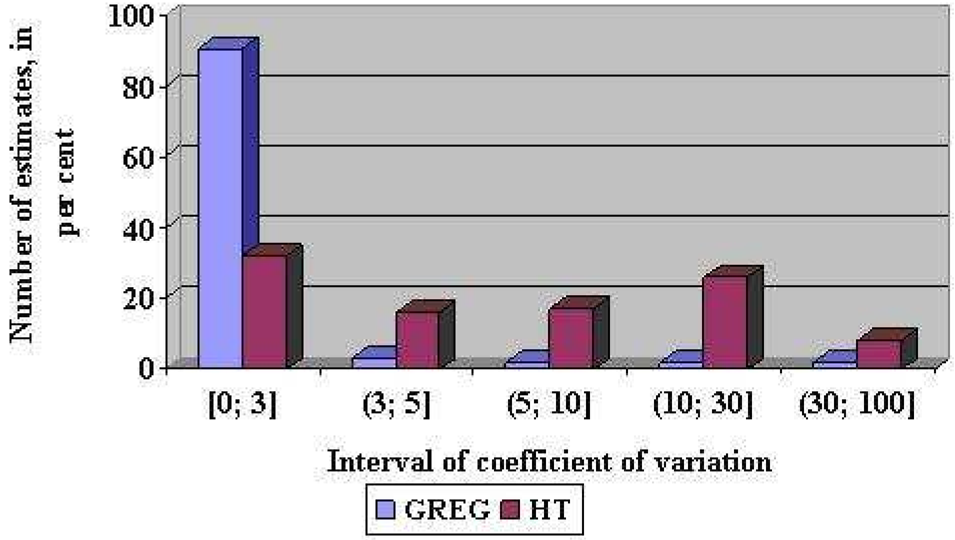
$g_k = 1 + \left(t'_x - \sum_{k\in\mathbf{i}} d_k \mathbf{x}_k^t\right)\left(\sum_{i\in\mathbf{i}} d_i \mathbf{x}_i \mathbf{x}_i^t\right)^{-1} \mathbf{x}_k$;

$\mathbf{x_k} = (x_1,...x_J)$ - vector of auxiliary variables,

J – number of auxiliary variables.

Introduction of GREG estimator significantly improved the quality of the results. The median value of coefficient of variation for GREG is 7 time smaller compare to Horvitz-Thompson (HT) estimator (0.7 for GREG and 5.2 for HT). The distribution of statistical estimates of variable “Number of Full-time Units” by size of coefficient of variation separately for HT and GREG estimators is presented in Figure 1. 91% of GREG estimates fall in interval [0; 3], but for HT estimator only 32% of estimates fall into the same interval. Also the share of non reliable estimates (cv >10) are significantly higher for HT estimator (35%) compare to GREG estimator (4%).

Figure 1: *Distribution of statistical estimates for number of full-time units by size of the coefficient of variation (CV) 2004*



The variable "Number of Full-time Units" is the one for which precision after introduction of GREG estimator was improved most compare to HT estimator. Other variables are not so considerably improved but anyway there is some improve for each variable.

Introduction of GREG estimator considerably diminished burden for respondents. About 45 000 enterprises do not need to fill in questionnaires each year. Also users are able to get results three month earlier than they used to.

2.2 Individual enterprises (linear regression model)

Individual enterprises are not included in surveys of Labour Statistics. Taking into account that individual enterprises cover significant part of employees it was decided to accomplish analysis on inclusion of individual enterprises into the scope of surveys of Labour Statistics.

Individual enterprises cover about 7.7% of total number of employees and about 29.5% of total number of enterprises. In some economic activities the number of employees in individual enterprises is significantly high. In the table below it is presented the number of individual enterprises compare to total number of enterprises and the number of employees in individual enterprises compared to the total number of employees.

Table 1: *The number of individual enterprises and number of employees in individual enterprises compare to total number of enterprises or employees, 2006*

NACE	Number of enterprises	Number of employees
Total	29.5	7.7
C	32.9	8.1
D	42.5	6.8
E	7.8	0.6
F	28.1	8.1
G	9.3	0.3
H	24.5	7.0
I	40.3	16.8
J	47.4	21.8
K	34.8	11.4
L	22.4	1.9
M	17.9	8.1
N	0.1	0.1
O	4.9	0.4

Implementing one of the goals of Statistics Lithuania to diminish burden for respondents it was decided do not include individual enterprises in the sample of survey but to estimate the required data on the basis of administrative sources and data of non individual enterprises.

From the viewpoint of employment individual enterprises are the same as non individual, also individual enterprises are controlled by the same laws as non individual. The only difference is that most of individual enterprises are small (usually have only few employees). The correlation between statistical variables and administrative variables for individual enterprises are the same as for non-individual enterprises.

The test Labour Cost Survey of individual enterprises for reference year 2004 was carried out in order to examine behaviour of individual enterprises. The results of the survey confirmed that in the terms of employment, behaviour of non-individual enterprises are the same as behaviour of individual enterprises. It was decided to prepare methodology for estimation of data of labour statistics for individual enterprises on the basis of non-individual enterprises.

Preliminary methodology for estimation of data of individual enterprises is prepared, but this methodology is still at the stage of improvement and it is supposed to be finished till the beginning of 2009.

At this moment it's most likely that for estimation of variables of individual enterprises linear regression will be applied:

$$y_k = A + B \times x_k$$

where:

y_k - variable of interest;

x_k - explanatory variable (derived from administrative sources);

A and B - regression coefficients.

Regression coefficients A and B are estimated applying least square method using the data of non-individual enterprises (y_k and x_k are known for non-individual enterprises).
The quarterly data of individual enterprises is supposed to be published since 2009.

3. Pros and cons

Usage of administrative sources for the production of official statistics has both: advantages and disadvantages.

Register based statistics undeniable is superior to traditional statistical methods. The biggest disadvantage is transitional period when moving from traditional statistics to register based statistics. The human resources required during this period are significantly higher and also the type of qualification differs. That means that beside day-to-day task, the staffs of Statistical Office have to be retrained. Also a lot of newly aroused problems have to be solved. To tackle the big amount of problems which were not met before is complicated.

Fortunately transitional period for Statistics Lithuania is almost finished and required staff is retrained and can handle the problems.

Statistics Lithuania had to initiate to change the laws in order to get administrative data, to solve technical problems with keeping huge databases and to retrain the staff. Each of those tasks is very time consuming but when the transitional period be passed the statistical data be more detailed, with better quality and lower costs. The burden for respondents will be diminished as well as burden for staff of Statistics Lithuania.

4. Conclusions

During the last year Statistics Lithuania made significant changeover of methods applied for estimation of official statistics. Nevertheless the final results remain the same.

At this moment Labour Statistics significantly diminished the burden for respondents, improved quality of the results and spared the costs by using administrative sources. However a lot of attention still has to be paid trying to extend the usage of such sources. Analysis of newly received administrative sources is ongoing as well as seeking for new possible sources. The methods for estimation based on administrative sources are still under consideration.

It costs a lot of time to know different registers and to use them in the day to day practise of official statistics but after long run when new methods will settle down it is expected that survey costs will be reduced significantly as well as quality of the results will be improved.

References

Wallgren A., Wallgren B. (2006) *Register-Based Statistics – Administrative Data for Statistical Purposes*, Wiley, New York.

United Nations Economic Commission for Europe (2007) *Register-based Statistics in the Nordic Countries – Review of Best Practices with Focus on Population and Social Statistics*, United Nations.

Sarndal C. E., Lundstrom S. (2005) *Estimation in Surveys with Nonresponse*, Wiley, New York.

Sarndal C. E., Swensson B., Wretman J. (1992) *Model Assisted Survey Sampling,* Springer-Verlag, New York

Combination of administrative and survey data for structural business statistics in Spain: Annual Structure of Earnings Survey

Ignacio Moral-Arce, Elisa Martín
Subdirectorate of Labour Market Statistics, INE (Spain)
e-mail: emartin@ine.es

Abstract: This work presents the Annual Structure of Earnings Survey, carried out by INE (Spain), which has as its main feature the use of administrative and tax records for obtaining results. For this purpose this survey combines information from Social Security files, data from the Quarterly Labour Cost Survey, a small survey conducted by the INE, and the information on income from form 190: Annual Summary of the Tax Agency Personal Income Tax (IRPF) Withholdings and Advance Payments on Account. The different cross-checks run with the administrative files, the features of the statistical matching and the identifier keys used for this purpose are all shown.

Keywords: Survey, administrative record, statistical matching, identifier key

1. Introduction

The origin of the Annual Structure of Earnings Survey (ASES) lies in the Structure of Earnings Survey (SES), which was first conducted in Spain in 1995. The main novelty presented by this survey compared with other earnings surveys consists of the capture of wages and salaries on an individual basis in the questionnaire and, alongside these, a large amount of variables related to the worker. Thanks to this, it is possible to establish links between salary and some variables that may help to determine its amount, as is the case of the educational level reached, length of service, type of contract or occupation, amongst others.

After the good results obtained, Eurostat considered the need for all member states to conduct regular surveys of this kind by promoting the creation of regulations governing the conduct of the SES on a four-year basis, the first year being 2002.

In Spain there was a certain gap in the statistical information available on earnings due to the fact that, apart from the survey conducted in 1995, practically the only data available was on salaries aggregated by company, establishment or, at the most, by more or less homogeneous groups within an establishment, but never individual information.

Another of the most innovative contributions of the SES is that not only are average earnings values supplied, but also the distribution of wages/salaries and, therefore, a measure of their inequality. The aims of the survey can basically summarize in twofold: knowledge of earnings levels, not only of average levels, but also of their distribution and the determination of the salary structure, both from the standpoint of composition and from that of the variables that affect it and the extent to which they do so.

This statistical operation, however, also has two major drawbacks. First, its four-yearly basis means that in the years between surveys no official information is available on salaries classified by sex and occupation. Second, but equally important, the excessive length and complexity of the questionnaire, which makes it unfeasible to increase the frequency with which this survey is conducted.

Bearing in mind the above, another important fact that we should remember is the amount of information that can be obtained from administrative files. In Spain there are two government agencies: the Social Security and the Tax Agency, with databases that include a large part of the information requested in the earnings surveys, such as the employee's gender, date of birth, nationality, gross salary, etc.

Taking into account, on the one hand, the limits of the SES and, on the other, the existence of administrative files, the aim of the Annual Structure of Earnings Survey (ASES) is to overcome the drawbacks of the SES, such that, by using administrative and tax files, results may be obtained on a yearly basis on annual earnings per employee classified by gender, age, occupation, nationality and type of contract., without thereby increasing the informative burden on enterprises. Lastly, since the methodologies of the SES and the ASES are very similar, an uninterrupted time series can be obtained, enabling us to study trends and changes in time.

The rest of the work is organized as follows: The next section shows the scope of the survey as well as the sample generation process. Section 3 analyses the different databases available, while section 4 presents the linkage process. Finally, the conclusions are set out in section 5.

2. Scope of Survey and Sample Selection Procedure

2.1. Scope of Survey

The population scope is formed by all workers employed in the local units who have been registered in the Social Security for more than two months during the year. Chairmen, board members and, in general, any personnel whose remuneration is not mainly in the form of wages/salary, but based on commission or profits are excluded.

The geographic scope encompasses the whole of Spain, with results disaggregated by regions.

As for sectoral coverage, we investigate the local units with business activity across the three main sectors: Industry, Construction and Services, specifically those centres with economic activities comprised in sections C to K and M to O of the Nomenclature of Economic Activities NACE Rev.1. Agriculture and fishery activities, the Public Administration, Defence and compulsory Social Security, domestic staff and extraterritorial bodies are all excluded from the survey.

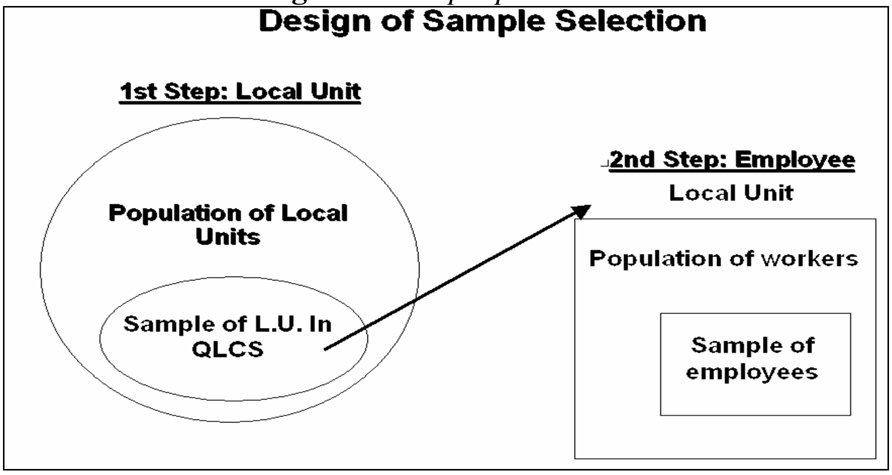
Finally, with regard to the time scope, the reference period is the calendar year (2004, 2005, 2007, etc.).

2.2 Sample procedure

The random unit selection procedure corresponds to two-stage stratified sampling where the first stage units are local units (contribution accounts), while the second stage units are the workers. The first stage stratification criteria are the region, the economic activity and the size of the unit. From the sample of local units obtained in this way we select randomly a nominal and representative list of wage and salary earners, which depends on the size of that unit. Figure 1 shows a diagram of the sample selection process.

To simplify the processes and take advantage of the surveys already being carried out, the sample selected at the first stage is the same as that of the Quarterly Labour Cost Survey (QLCS), so that information will be required from almost 19000 contribution accounts making up the quarterly sample of this survey, so the sample design of the survey is similar to that of the QLCS. In terms of the workers, the sample of employees obtained is around 170,000.

Figure 1: *Sample procedure*



The following section describes the different information sources used to carry out this statistical operation.

3. Sources of information

One of the basic aims of the ASES is to obtain results, but without these entailing a informative burden of the respondent. For this reason, it is necessary to use a range of information sources. Three different sources are used in this survey.

3.1. INE - Quarterly Labour Cost Survey

The Quarterly Labour Costs Survey (QLCS) is a continuous short-term statistic, elaborated quarterly by the INE. The population scope is the Social Security contribution accounts whose economic activity is related to industry, construction or services for the whole country. A contribution account can be defined as a local unit. For each account all employees associated with the account are investigated. The sample size is 19500 establishments, divided into three monthly subsamples of 6500 establishments. The QLCS provides levels and indicators on the average cost of labour per worker and month, the average cost of labour per hour actually worked and the time worked and not worked. This sample therefore offers us all the information associated with the local unit (region, branch of activity, company size, company tax identity number, etc.)

3.2. Social Security Agency - General Register of Affiliations

The General Register of Affiliations of the Social Security General Treasury contains the list of workers registered at the different contribution accounts in the year of reference. The identification of the workers is done by means of their Social Security number, which is unique for every employee. This Social Security number acts as an identifier key. From these lists of workers and after a series of checks and edition criteria, we obtain the framework from which the sample of employees who will form part of this investigation will be selected.

In addition, the information offered by the General Register of Affiliations is of great importance, as for each worker it supplies the tax identity number, date of birth, nationality, gender, date of registration and removal, type of employment contract, as well as the contribution group, and so forth.

3.3. Tax Agency - Personal Tax Annual Register – Form 190

Form 190 is specified in the Personal Income Tax Regulations, which stipulate in section 2 of article 101 that every withholder or person obliged to make payments on account should submit an annual summary of withholdings and interim payments made, where, besides stating his/her identification details, a list of recipients will be included showing for each of them the data that have been taken into account to determine the withholding rate or payments on account.

The data that are included in form 190 are the identification of the taxpayer and of the income receiver, who are identified by their tax identity number. In addition, information on the income receiver's full name and province of residence are also obtained. Regarding remunerations, file 190 also has information of the key identifying the kind of remuneration (we are only interested in key A: employees in general), the whole remuneration (total annual amount of the pecuniary emoluments actually paid), withholding applied (total annual amount withheld), valuation of payments in kind (sum of the valuations corresponding to payments in kind actually made in the year), interim payments made (amount actually deposited by the payer), and the interim payments charged (the interim payments on account actually charged to the income receiver).

After reviewing the different records that are used, along with the most significant variables offered by each of them, in the next section we describe how the process of cross-checking the different records and obtaining the different variables is carried out.

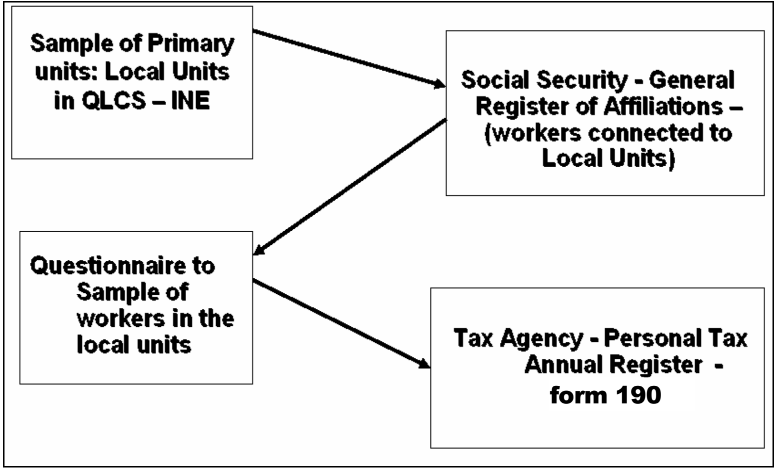
4. Record linkage procedure

The ASES is constructed in four stages, as shown in Figure 2. It shows that the initial starting basis is the information supplied by the Quarterly Labour Costs Survey, which is a representative sample of contribution accounts in Spain in the industry, construction and services sectors.

Information is requested from Social Security on the workers associated with those centres during the year in question. The General Register of Affiliations provides the directory of workers at that centre and systematic sampling is carried out for each individual centre. A questionnaire is sent to the local unit together with the list of individuals selected to answer questions on their occupation and functions.

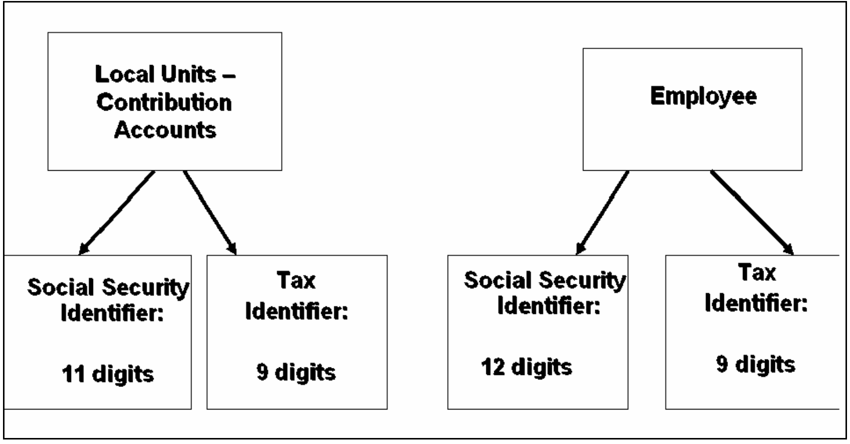
Finally, with those employees for whom all the above information is available, information is requested from the Tax Agency on the remunerations that the employee selected has received at the selected contribution account during the year under investigation.

Figure 2: *Steps of sample procedure*


Sample of Primary units: Local Units in QLCS – INE
Social Security - General Register of Affiliations – (workers connected to Local Units)
Questionnaire to Sample of workers in the local units
Tax Agency - Personal Tax Annual Register - form 190

The basic elements for carrying out this whole procedure are the identifier keys, which will enable us to carry out statistical matching, appropriately at every stage. As already mentioned in the previous section, in Spain we have two relevant statistical units – local units and employees. As may be observed in Figure 3, the units of the Administrative Register do not exactly agree with the units of the Tax Register, and each register has its own system of identification keys for its units. The problem therefore arises from the non-existence of a unique identification key for a unit, as the first of the identifiers (Social Security number) is exclusive to the Social Security file, while the tax identity number is what is used for identification purposes in the Tax Agency file.

Figure 3: *Key identifier*

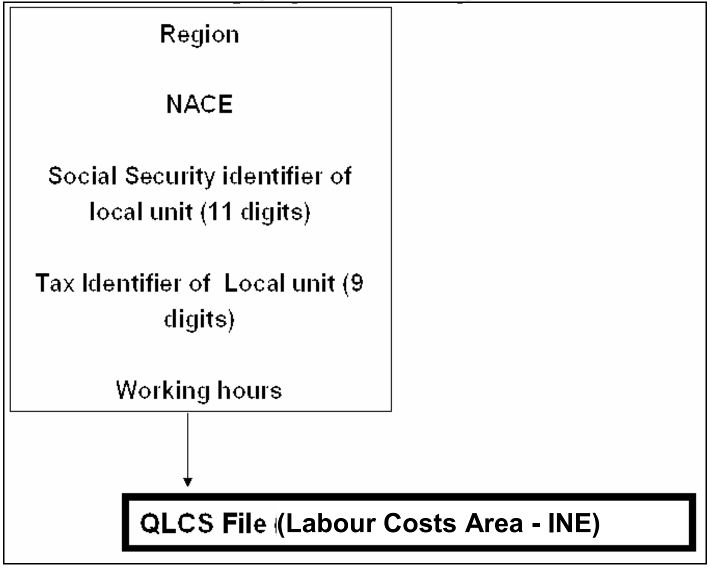

Local Units – Contribution Accounts
Employee
Social Security Identifier: 11 digits
Tax Identifier: 9 digits
Social Security Identifier: 12 digits
Tax Identifier: 9 digits

Bearing in mind the two existing identifier types, we are now going to take a closer look at each individual stage and what variables are obtained at each of them.

4.1. First Step: Sample of Local Units (Contribution Accounts)

At this first stage we obtain all the significant information of the Contribution Account (primary unit in our sampling selection process). To this end, we use one of the surveys conducted at INE: The Quarterly Labour Cost Survey. This statistical operation offers information on the branch of activity, the size and the region where the contribution account is located. Figure 4 offers a summary of the information that we are going to use from this source of information.

Figure 4: *First Step*


Region
NACE
Social Security identifier of local unit (11 digits)
Tax Identifier of Local unit (9 digits)
Working hours
QLCS File (Labour Costs Area - INE)

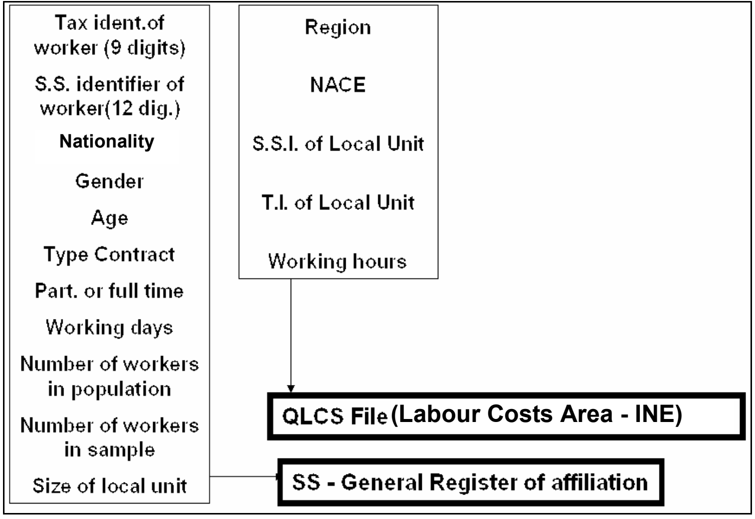
4.2. Second Step: Worker Universe associated with the Contribution Accounts

Using the information from the previous stage, the Social Security is asked for information on workers associated with those Contribution Accounts during the year in question. For this purpose, we use the 11 digits of the Social Security number as the local unit identifier. By way of an exact record linkage, the Social Security feeds back information contained by the following variables:

- Local Unit - SS identifier: 11 digits
- Employee – SS identifier: 12 digits
- Information of type of contract, period, number of contracts, etc.

Figure 5 shows the information that is obtained at this stage. As may be observed, the information obtained makes reference to the employee (secondary unit), while at the previous stage the bulk of the information was on the primary unit (contribution account). With the information received from the Social Security we now have the directory of employees at each individual Contribution Account. Initially, a worker debugging process is carried out, deleting those that do not meet certain conditions. Once the directory has been debugged, we obtain the final sample of workers using systematic sampling (by contribution group and sex). With this sample of secondary units we can now move on to the third stage.

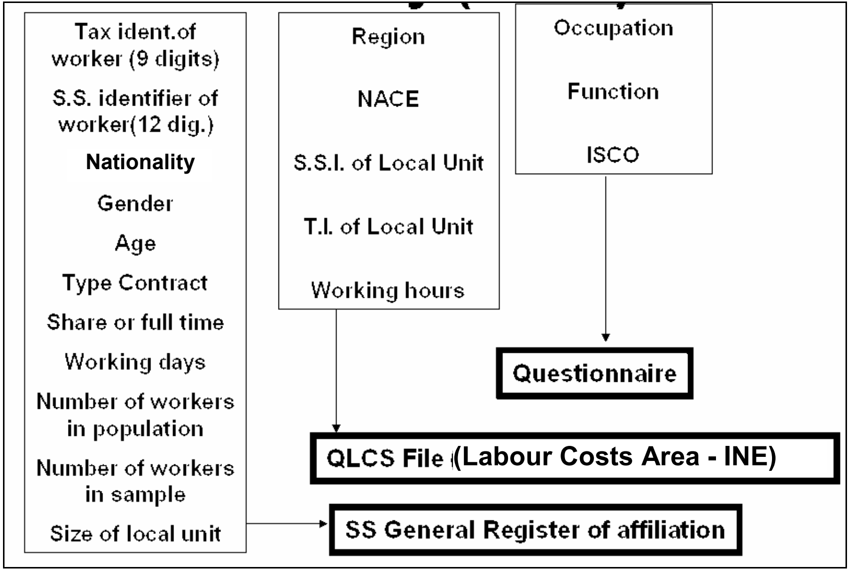
Figure 5: *Second Step*


Tax ident.of worker (9 digits)
S.S. identifier of worker(12 dig.)
Nationality
Gender
Age
Type Contract
Part. or full time
Working days
Number of workers in population
Number of workers in sample
Size of local unit
Region
NACE
S.S.I. of Local Unit
T.I. of Local Unit
Working hours
QLCS File (Labour Costs Area - INE)
SS - General Register of affiliation

4.3. Third Step: Questionnaire to selected employee

One of the most important features presented by salary surveys is to see the relationship existing between the worker's salary and occupation[1]. Unfortunately, the administrative files do not offer reliable information on this variable, so the respondent needs to be asked for it directly. For this purpose, a questionnaire is remitted to the local units asking for information on the occupation and functions performed by the worker selected in the survey. The answers obtained on these two questions are used to produce the International Standard Classification of Occupations (ISCO) 2-digit code. Figure 6 shows the list of variables available after this stage is completed.

Figure 6: *Third Step*


Tax ident.of worker (9 digits)
S.S. identifier of worker(12 dig.)
Nationality
Gender
Age
Type Contract
Share or full time
Working days
Number of workers in population
Number of workers in sample
Size of local unit
Region
NACE
S.S.I. of Local Unit
T.I. of Local Unit
Working hours
Occupation
Function
ISCO
Questionnaire
QLCS File (Labour Costs Area - INE)
SS General Register of affiliation

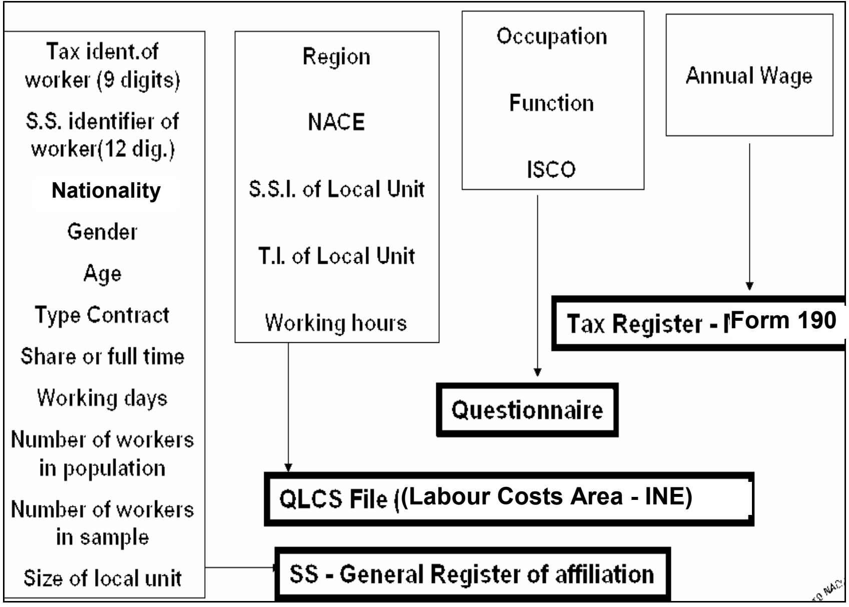
[1] Another extremely important variable is education, which is not satisfactorily obtained from administrative files either. It is not requested in the questionnaire either as we feel it is much more important that the respondent burden is not increased than obtaining data from this variable.

4.4. Fourth Step: Employee Earnings

After receiving the response from the reporting units on occupation and carrying out the ISCO encoding, at the last stage the Tax Agency is asked for information on the earnings that the employees have received at that contribution account. To carry out the cross-check between employee and company we use the 9-digit tax identity number as the identifier. By means of an exact record linkage the Tax Agency returns information to us containing the following variables:

- Local Unit – Tax Agency identifier: 9 digits
- Employee – Tax Agency identifier: 9 digits
- Annual wage

Figure 7: *Fourth Step*



When the whole variable selection process is completed, Figure 7 shows how our final data matrix appears and what the source for each one is.

Bearing in mind the multiple database generation stages and the fact that this statistical operation employs different sources of data (administrative, tax) and uses different identifier keys, it is necessary to analyse the evolution of the coverage at different stages. Figure 8 shows the coverage of the survey for our first reference year (2004).

Figure 8: *Evolution of coverage*

	percentage of coverage (19920 local units)	percentage of coverage (166248 employees)
QLCS	100.00	
SS file of Affiliations		
Debug and sample	96.82	100.00
Questionnaire	87.87	89.44
Tax Agency	85.22	87.55

5. Conclusions

This paper analyses the use of administrative and tax files for conducting a structural business survey. In particular, a new statistical operation is studied - the Annual Structure of Earnings Survey - which uses several sources of information.

The most important conclusion reached from this study is the great wealth of information available in the administrative Social Security and Tax Agency files, which should be used before designing new surveys.

The bulk of the information used in the ASES is obtained from Social Security and Tax Agency files. There is a need, however, to supplement these data with partial information (normally individual) on occupation and hours worked. In spite of this, the informative burden of this statistical operation is extraordinarily low. If, furthermore, the sample of CCC's of the ETCL is used, only the worker's occupation by means of an additional survey is required.

Finally, since the Structure of Earnings Survey, conducted every four years since 2002, is supplemented in the years between surveys by the ASES. By using the files described in this document, it is possible to obtain a time series, as the two surveys are carried out with highly similar methodologies.

References

Denk, M. and Hackl, P. (2003) Data integration and record matching and Austrian contribution to research in official statistics. *Austrian Journal of Statistics*, 32, 305-321.

Haslinger, A. (1997) Data matching for the maintenance of the business register of statistics Austria. 2, *Methods and Techniques*, 199-209.

Personal Data Protection Act 15/1999, of 13 December 1999.

Act 5/1992 of 29 October 1992, governing the automated processing of personal data

Moral-Arce, I. and Martín, E. (2006) Utilización de ficheros administrativos en las encuestas: La Encuesta Annual de Estructura Salarial. I*V Congress on Survey Methodology*. Pamplona. Spain.

5

Integration of registers and samples 2

Investigating road traffic accident statistics - Matching hospital and police data.

Kevin McGrath[1], Matthew Tranter[2]
[1]Office for National Statistics, UK
[2]Department for Transport, UK
e-mail: Kevin.McGrath@ons.gov.uk, Matthew.Tranter@dft.gsi.gov.uk

Abstract: Data sourced from two different administrative systems can sometimes be used to measure the same variable of interest. Often the two estimates differ and this can call into question the credibility of an official statistic. Such a case occurs when comparing road traffic accidents recorded by the police with those admitted to hospital. This paper describes matching the common units in the two datasets to enable researchers to attempt to explain the difference. Initial matches are made on core variables, geography, sex, age and date. The correctness of these matches are determined by constructing rules based on deriving as much information as possible from the values of the other common variables (notably Postcode) on the set of initially matched records. The rule-based matching process is described.

Keywords: Road traffic accidents, rule-based matching, HES, STATS19

1. Introduction and background

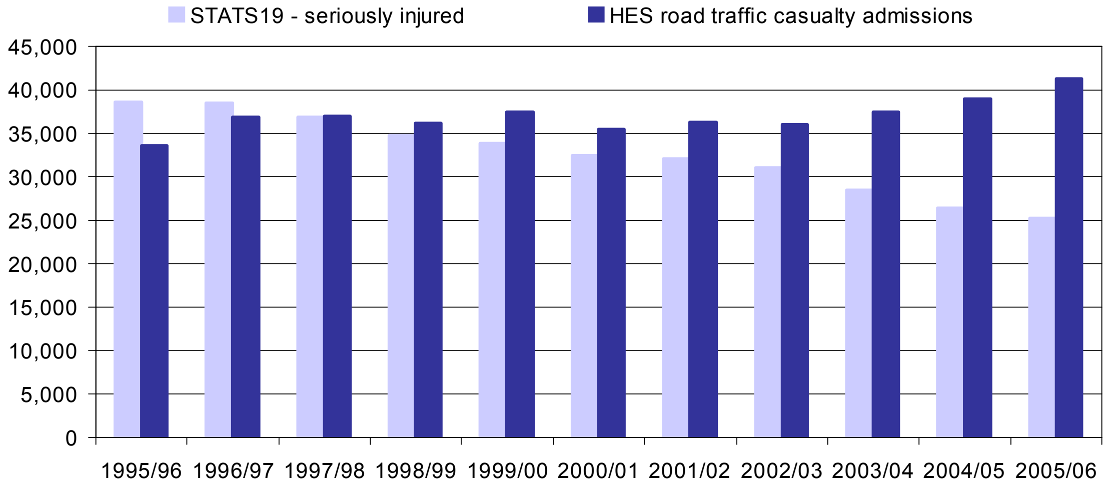
The emphasis in this paper is in the description of the practical implementation of matching two datasets sharing a large number of identical units but which lack a common identifying matching key variable. The matching approach is a mix of exact and rules-based matching rather than probabilistic. The lack of a common identifying matching key variable forces the use of the common characteristic variables, sex, age, and date (of admission to hospital and of accident) to find candidate matches. Derived variables can be generated to refine the matching, based on geographic region, and road user type and casualty class, common in both datasets.

Information on casualties in road traffic accidents in England is available from both a long established database of data collected by the police, (known as STATS19[1]) and more recently from data on hospital admissions (HES - Hospital Episode Statistics). These sources provide an alternative, though not equivalent measure of the number of seriously injured casualties on the roads of England. The use of hospital records can complement and inform the use of police data to monitor Government casualty reduction targets (DfT, 2006a).

[1] Named after the number of the first questionnaire issued when the system was introduced in 1949

The two sources of data have shown (see figure 1) differing trends in recent years (Gill *et al*. 2006). The police records show a larger fall in seriously injured casualties than are admitted to hospital (as recorded by HES). These may be due to definitional differences, changes in levels of reporting to the police or changes in police recording practice. Studies (DfT, 2006a) suggest a police tendency to underestimate severity rather than overestimate because of the difficulty of classifying severity at the scene of accident. In STATS19 the definition of serious injury includes all casualties admitted to hospital and certain injuries, such as fractures, whether or not the casualty was admitted to hospital. The number of the most severely injured casualties may have declined less than those casualties recorded as 'serious injured' but not serious enough to go to hospital. Hospital data systems may have changed resulting in improved accuracy in recent years. It is also possible there has been an increase in the proportion of road casualties admitted to hospital.

Figure 1: *Traffic injury levels for England measured from police statistics (STATS19) and from admissions to hospital (HES) by year.*



For a discussion of government targets for reducing road traffic casualties and a comparison of measures produced from the two datasets, see the Department for Transport (DfT) paper (DfT, 2006a).

In order to investigate these trends further, for the years 1995-2005, matching has been undertaken of police records of people injured in road accidents in England with records of patients admitted to hospital who were injured in a road accident (collated by the Department of Health and supplied from all hospitals in England).

Such a matched database can also have potential uses to link the circumstances of a road accident with its medical consequences. STATS19 data provides information on the circumstances of an accident, with only an overall assessment of the severity of resulting injuries. In contrast HES data provides information on injuries received with little information regarding the circumstances of the accident.

There has been previous matching of STATS19 data with local hospital records for particular geographical areas, and with hospital accident and emergency data (Ward *et* *al.*, 2007). Also matching has been carried out with the HES equivalent (Scottish hospital in-patient system - SHIPS) in Scotland, latest results are to be published later in 2008 (Broughton *et al,* 2007). The matching process is complicated by the fact that the name and address of the accident casualty is neither recorded by the police nor is it released by the Department of Health on the HES dataset. The closest to a unique matching key variable common to both datasets is home postcode of casualty and of patient, although this is subject to varying levels of recording accuracy by police forces. This is the first time that a matched dataset for the whole of England has been available to researchers.

A discussion of the sources of the data follows in section 2. Section 3 contains definitions of the matching variables and the linking methodology. Section 4 describes the matching strategy and matching rules, which are tabulated in Annex 1.The results of the matching is discussed in Section 5 and presented in Annex 2.

2. Data sources and definitions

2.1 Personal injury road traffic accidents (STATS19)

The Department for Transport's (DfT) national database of personal injury road traffic accidents in Great Britain is based on information collected by the police in a system known as STATS19. One record is created for each person injured (casualty) in a road accident on public roads in which at least one vehicle is involved and which becomes known to the police within 30 days.

The scope and detail of STATS19 allows the identification of different accident circumstances, enabling road safety policies to target appropriate interventions to reduce the number of accidents and their resulting casualties.

Some 50 data items are collected for each accident, recording information on the accident, the vehicles involved and the casualties (age, sex, severity of injury, and whether driver, pedestrian or passenger). Casualties are classified as fatal (death within 30 days), seriously injured or slightly injured; the severity of casualty is recorded by the reporting police officer on the basis of information available within a short time of the accident. In STATS19, the definition of serious injury includes all casualties admitted to hospital and certain injuries, such as fractures, regardless of whether or not the casualty was admitted to hospital.

A complex devolved reporting system such as STATS19 will never produce perfect results; the standards that are achieved reflect the efforts of local authorities[2] and police forces to report to the standard national requirement. However while very few, if any, fatal accidents do not become known to the police, research has shown that a significant proportion of non-fatal injury accidents are not reported. Note that it is not a legal requirement that accidents should be reported to the police provided that personal details are exchanged. It is important to get a good estimate of the level of reporting. If there has

[2] Local Authority Districts are the main tier of local government in England. There are 354 in England with an average population of about 140,000.

been a systematic change in the levels of reporting, this would cause a problem in monitoring trends.

The DfT data analysed consists of approximately a quarter of a million records each year between 1996 and 2005 relating to all accidents in England, so this includes accidents classified as slight as well as serious.

Table 1: *STATS19 variables used in matching*

Casualty Data	
Sex	
Age	
Casualty Type (Road User Class)	Bus / car / motorcycle / pedal cycle / pedestrian …etc.
Casualty Class	Driver or rider /passenger/pedestrian
Casualty home post code	
Accident Data	
Date of Accident	
Police force code	
Local Authority code	
Severity of Accident	
Ordnance survey grid reference of accident	

2.2 Hospital Episode Statistics (HES) data

Information on casualties admitted to hospital as in-patients in England is contained on the Hospital Episodes Statistics (HES) database held by the Information Centre of the National Health Service (NHS). It is compiled by the Information Centre from over 300 NHS Trusts in England. The records relate to individual consultant episodes, including information on admission to and discharge from hospital. They exclude casualties treated in Accident and Emergency departments who are not subsequently admitted to a hospital.

Hospital admission figures are based on periods of care ('episodes') under a particular consultant, so patients can be counted more than once (e.g. if they transfer to another consultant). The extract used is selected on the external cause of injury for all admissions, recorded using the International Classification of Diseases (ICD-10)[3]. The criteria is to select those patients who on admission to hospital have been involved in road traffic accidents and specifically injured in a road traffic accident, to differentiate them from those who were involved in accidents that occurred 'off-road' (consistent withSTATS19). All traffic accident casualties admitted to a bed in a hospital in England should be recorded as an admission episode in HES data. HES records also contain the age of the patient and where they lived. There are further codes to identify the hospital,

[3] ICD-10 Reference http://www.who.int/classifications/apps/icd/icd10online/

the length of time the patient stayed in hospital, and the speciality of the consultant who treated them.

Data for the years between 1996 and 2005[4] are supplied by the Information Centre for Health and social care belonging to the NHS and consists of between approximately 55 to 60 thousand records annually.

Table 2: ***HES*** *variables (HES data dictionary www.hesonline.nhs.uk)*

Patient Data	
Sex	
Age at start of episode	Years (or months if under 1)
HES ID	Unique patient Identifier. A combination of; NHS number, local patient identifier, postcode, sex and date of birth.
Postcode	Postcode of home address
Postcode district	
Census Output area	
LAD	LAD of patient's home
Admission and discharge data	
Date of admission	
Method of admission	
Date of discharge	
Method of discharge	
Episode data	
Date episode started	
Episode key	
Episode order	
Days of intensive care	
Diagnosis data	
Primary diagnosis code	
Secondary diagnosis codes 1-14	
Cause code	External cause of injury (ICD-10)
Organisation code	
Provider code	
Primary care code	
Health authority of treatment	

[4] Data was provided on a financial year basis and therefore the 2005 data covers January-March only

3. Data linking methodology and matching variables

This paper describes a practical implementation of a matching process where the aim is to link records from two different datasets where many units are believed to belong to both datasets. There is only partial identifying information, and there is no common identifying key which could link the records exactly. This might have been an example of probabilistic matching – where a numerical measure can be estimated of how well two particular records match, however as this is dependent on creating some sort of benchmark file possibly involving clerical matching this was not feasible within the constraints of the project. What follows is a description of matching two datasets without a common identifying key and of the rules used to assign a match as 'correct'.

Matching was performed using (Proc) SQL commands within the SAS programming language. The data management features of SAS enable conversion of the supplied ascii delimited input files into labelled datasets, with all the various data manipulation procedures necessary - checking the data , formatting and recoding of variables, variable and value labelling, as well as the ability to run various statistical procedures.

The matching process assigns HES as the master database, on to which STATS19 records are matched. For 2004 data, 249553 STATS19 records are matched against 58747 HES records. The SQL option is chosen so that the output matched database contains all the matched records as well as all the non-matched HES records.

After the SQL statement performs the initial match the output records will have several part-duplicate record matches. These arise when several records in one of the datasets have identical values for all the initial matching variables. All of these records will attach themselves to any one record in the other data set which has identical values on the same initial variables. This causes an expansion of the size of the matched database and the part of the matched record contributed from the dataset with the duplicated values will thus be duplicated for all of these records (hence called part-duplicates) but only one will possibly be correct. Formally there is a many-to-many relationship between the two data-sets. As only one of the merged records can be correct it is necessary to use information from other variables to attempt to find which of the merged part-duplicate records is most likely to be the correct match.

Police records in STATS19 do not record the name and address of the casualty and the released HES medical records contain only the home postcode of the patient due to confidentiality concerns. If name and address are supplied then the matching process can be a straightforward process based on these identifiers as an exact or nearly exact match. In the absence of a common identifying variable, other common variables must be used, and in such a way which captures the maximum number of likely correct matches. In general the variables can be classified either as analysis or matching variables.

Standardisation

The fields from two datasets need to be standardised if they are to allow comparison between the different data sources. Standardisation may involve nothing more elaborate than removing inconsistencies. A simple example refers to age; in HES for age under one year old the variable has specific codes for monthly age groups. STATS19 does not record such precise detail, so it is important to recode the age variable in HES from months less than one year old to nought years old.

Blocking

To save comparison of all possible combinations of record matches blocking variables are often used in matching. The blocking variable blocks both datasets into mutually exclusive subsets. It is chosen such that records for the same person are very unlikely to appear in different blocks. Only matches within the same blocks are compared. This reduces the number of comparisons needed by only comparing record pairs where links are more likely to be found. Both files are divided into the same blocks and records within a particular block in one file are compared only with records within the same block on the other file. Blocking is most effective when they break up the population into small groups of similar size.

The matching variables are those which are common to both datasets, with the same definition, suitable variables are;

- sex
- age
- date
- home postcode (of hospital patient and accident casualty)
- Local Authority District (of patient's home and of accident location)

Others can be derived using lookups;

- Strategic Health Authority (of hospital and of accident location)
- casualty type
- casualty class

The analysis variables are those which are to be used to analyse relationships within the problem being considered and includes severity of the accident. It is important that this should not be used to determine the matching as this may lead to a biased analysis.

The matching variables and how they can be used are discussed below.

3.1 Strategic Health Authority: geographic blocking

As previous discussed a blocking variable may considerably reduce the number of record comparisons necessary and so speed up the matching process.

A suitable geographic variable common between the two datasets can be obtained. The HES data has a variable called HATREAT which gives the Health Authority for Treatment at that time. Over the ten years studied these have changed with successive re-organisations of the NHS[5]. A lookup can be obtained which links each health authority of treatment to the 28 Strategic Health Authorities (SHA) extant up to June 2006 and also to the 10 SHAs created in a re-organisation in July 2006. Annex 3 shows a map of the 28 SHAs. In practice the 28 areas serve as conveniently sized geographic units for blocking purposes whilst the 10 areas are too broad.

[5] http://www.statistics.gov.uk/geography/england_health.asp

In the STATS19 data the Local Authority District where the accident occurred is supplied. A straightforward lookup table links this to the 28 Strategic Health Authorities.

The end result is a derived variable for 28 Strategic Health Authorities in both files. This can then be used as a geographic blocking variable. In order to capture more matches after initially matching on exact SHAs, subsequently matching is attempted between adjacent SHAs - where the SHAs of the hospital of treatment are adjacent to the SHAs of the accident location.

3.2 Age of casualty/ age of patient

STATS19 has the age of the casualty as recorded by the police. HES has a STARTAGE variable, age of patient at the time of admission, derived from the date of birth and start of episode. This is thought to be more reliably recorded than the corresponding variable in STATS19.

Figure 2 shows the histogram of the STATS19 age distribution for 2004, spikes occur at ages ending in 0 and, to a lesser extent 5 suggesting rounding within police records to those ages in cases where the exact age of a casualty is not known. Figure 3 shows the corresponding age distribution for the HES data.

Figure 2: *Histogram of age for 2004, STATS19 data*

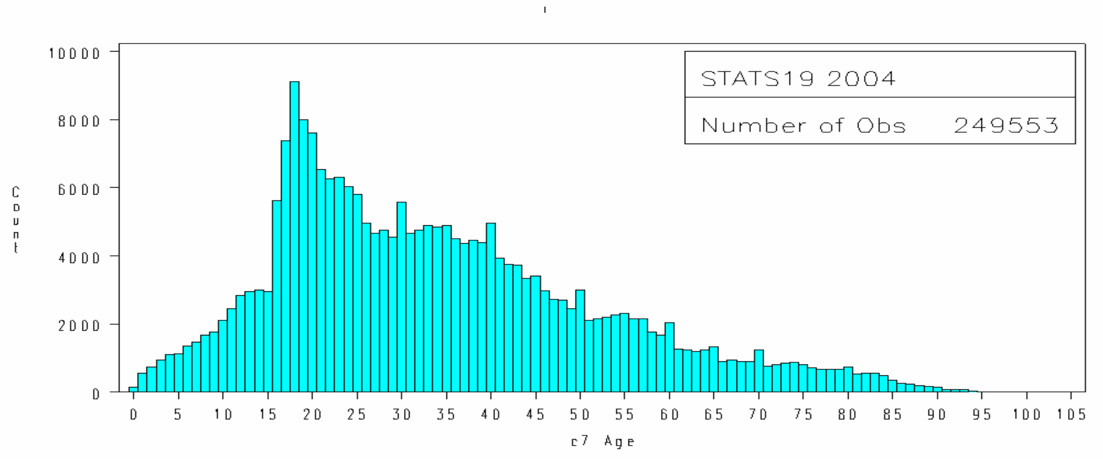
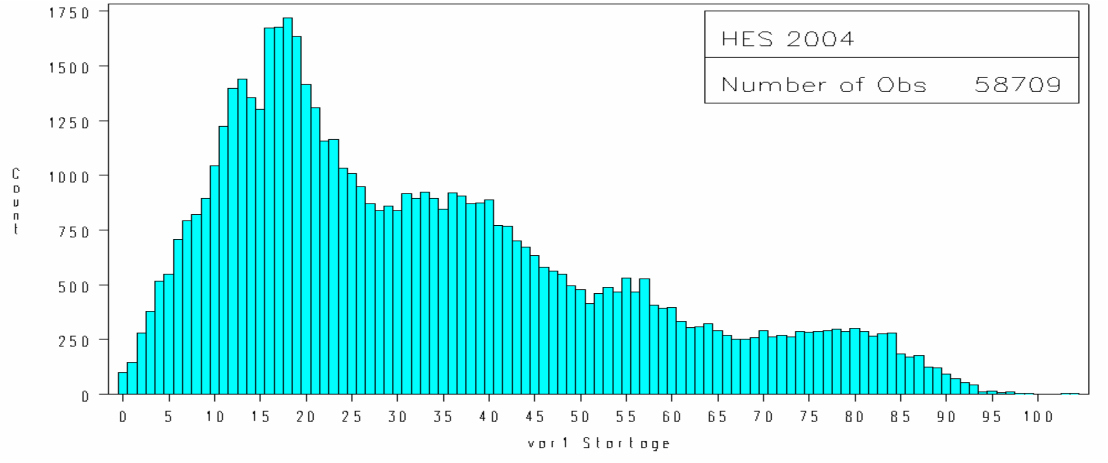


Figure 3: *Histogram of age for 2004, HES (road traffic accidents) data*



The HES age distributions do not show systematic spikes at certain ages and are thought to be accurately recorded within the Heath information system. Comparing histograms shows there to be a bulge in HES relative to STATS19 in the sub-17 age group (as shown by figure 3 below). This could be because the ICD codes used to select the HES accident records include a category of accident which may be disproportionate in this younger age category. For example a large number of young cyclists are injured by falling off their bicycles, often so severely injured that they are taken to hospital. With no other vehicle involved however usually they will not be reported to the police. These codes are included in the HES extract because it is known there is a likely chance of mistakes when coding these categories.

To account for the rounded STATS19 figures, it is reasonable to accept a wider range of HES ages as potential matches. For casualty ages 20 and above, those that end in a 0 or 5 on STATS19 should be allowed to vary by up to 3 years (for example a casualty recorded as 30 on STATS19 can be matched to casualties aged 27 – 33 on HES). All other STATS19 ages can vary by one year (for example a casualty recorded as 31 on STATS19 can match to casualties aged 30 – 32 on HES).

3.3 Date of accident

It is assumed an exact date match will match records by the date of accident with the date of the start of hospital episode. Logistically an exact match includes the day after the accident as casualties from accidents during the night may often arrive at hospital the following day. To introduce some tolerance a match is included with a HES date (i.e. date of hospital admission) up to 2 days after the accident as recorded on STATS19. This allows for less seriously injured casualties who may not immediately go to hospital.

3.4 Postcode

Postal codes in the UK are known as postcodes and were introduced for the purpose of automating mail sorting. As convenient units of geography other uses have been found for them.

They are usually made up of 7 or 6 alphanumeric characters, mostly AA9 9AA or AA99 9AA where 9 denotes number and A any letter. The first part (before the space) denotes the postcode area district or out code and is used to direct mail to the destination sorting office. The inward code (after the space) is used to sort the mail into individual delivery rounds.

Each postcode identifies an address to within 100 properties (with an average of 15 properties per postcode).

It is known that postcodes recorded in STATS19 are incomplete with missing postcodes being common. Some police force areas report no postcode at all. If the exact postcode is used in the initial stages of a matching process there will be many missed matches. Tolerance can be used to increase the matching by allowing matching on incomplete postcodes, for example if 2 out of the first 3 characters match.

The postcode variable is standardised in both datasets following the recommendation made by Leicester Gill (Gill, 2001) in his report on record matching for the ONS. This is to standardise postcode by removing the space character and left justifying the remaining characters, creating a 7 character wide variable.

3.5 Local Authority District (LAD)

The STATS19 database records Local Authority District where the accident occurred. The HES data records the patient's home Local Authority District. As this only matches when the accident occurs in the patients home LAD it is not a powerful matching variable.

3.6 Derived variables; casualty class and casualty type

The external cause code of injury is given by the International Classification of Disease codes – (ICD-10) A table on the DfT [6] website provides a look-up to the ICD code in order to derive the HES values for Casualty (road user) type and Casualty class. These are matched with the corresponding variables of casualty type and class which are directly recorded and so are already available on STATS19. They are used to provide supporting evidence or otherwise on the likelihood of a correct match found using the main matching variables.

4. Matching strategy

The development of the matching strategy had to take into account the main features of the datasets, considering that a common identifying key variable is not available. Therefore, the strategy is fully based on expert judgment to define rules to accept whether or not a match is correct.

The first step of the matching procedure is to search for exact matches on the key variables: Strategic Health Authority (SHA), sex, age and date. However, several STATS19 records have identical values of combinations of SHA, sex, age and date and so they could be attached to any HES record with the same values (although only one of the matched records could possibly be correct). Also, even when there is no duplication in the exact match of the key variables, this does not necessarily imply a correct match.

Therefore, further evidence from other variables is required to determine which match is most likely to be correct. Postcode and additional variables such as casualty class, casualty type, and LAD (of accident and of patient's home) are compared to increase the confidence in whether the record is a true match or otherwise.

It is important to note that only allowing for exact postcode match will automatically eliminate many potentially correct matches (e.g. some police forces do not record postcode) so matching rules are defined in order to permit matching of those records with similar postcode sectors.

[6] http://www.dft.gov.uk/172974/173025/221412/221549/227755/285672/Article6HESandroada1.xls#'ICD 10 Codes'!A1

As a result, a set of rules are derived to determine the criteria for a match. The rules comprise not only levels of tolerance for differences in the values of the key variables (or in the characters of the postcode) but also define how the remaining variables are taken into account to allow a match. After consideration of the above the matching strategy is codified into a set of rules which are described in the next section and presented in Annex 1.

4.1 Definition of the matching rules

After matching on EXACT SHA, sex, age and date, of all the remaining variables the postcode variable is considered to have the most discriminatory power. If the postcodes are exactly the same then a match is accepted as correct, otherwise the first 3 (or 4) characters are analysed depending on the postcode format.

If there is a valid STATS19 postcode (that is, the field contained at least one character and it started with a letter), a match is accepted if:

- the first two characters are identical;
- any 2 of the first 3 characters are identical or any 3 of first 4 characters are identical (if the postcode had 3 or 4 characters on its first section respectively).

On the other hand, if no valid postcode is available at STATS19 then a match is accepted if at least one of the remaining variables (casualty type, casualty class and LAD) has the same value. Greater priority is given to casualty class than casualty type, with least priority to LAD. For example if one record has only matched casualty type and another matched LAD then the first record is selected.

In addition, **rules are also defined to permit matching based on adjacent SHA, sex, fuzzy age, fuzzy date** (details of tolerance levels are defined in Annex 1. In this situation, postcode information is handled in the same form as described for the cases of exact matching in the key variables. However if no valid postcode is available at STATS19 then a match is accepted if any two of the remaining variables (casualty type, casualty class and LAD) have the same value.

Finally, for records remaining unmatched after following the steps above, a final sweep are run, without taking into account SHA, and accepting a match only if age, sex, date and postcode have exactly the same values in both datasets.

Figure 4.1 *Top-level view of the matching Steps*

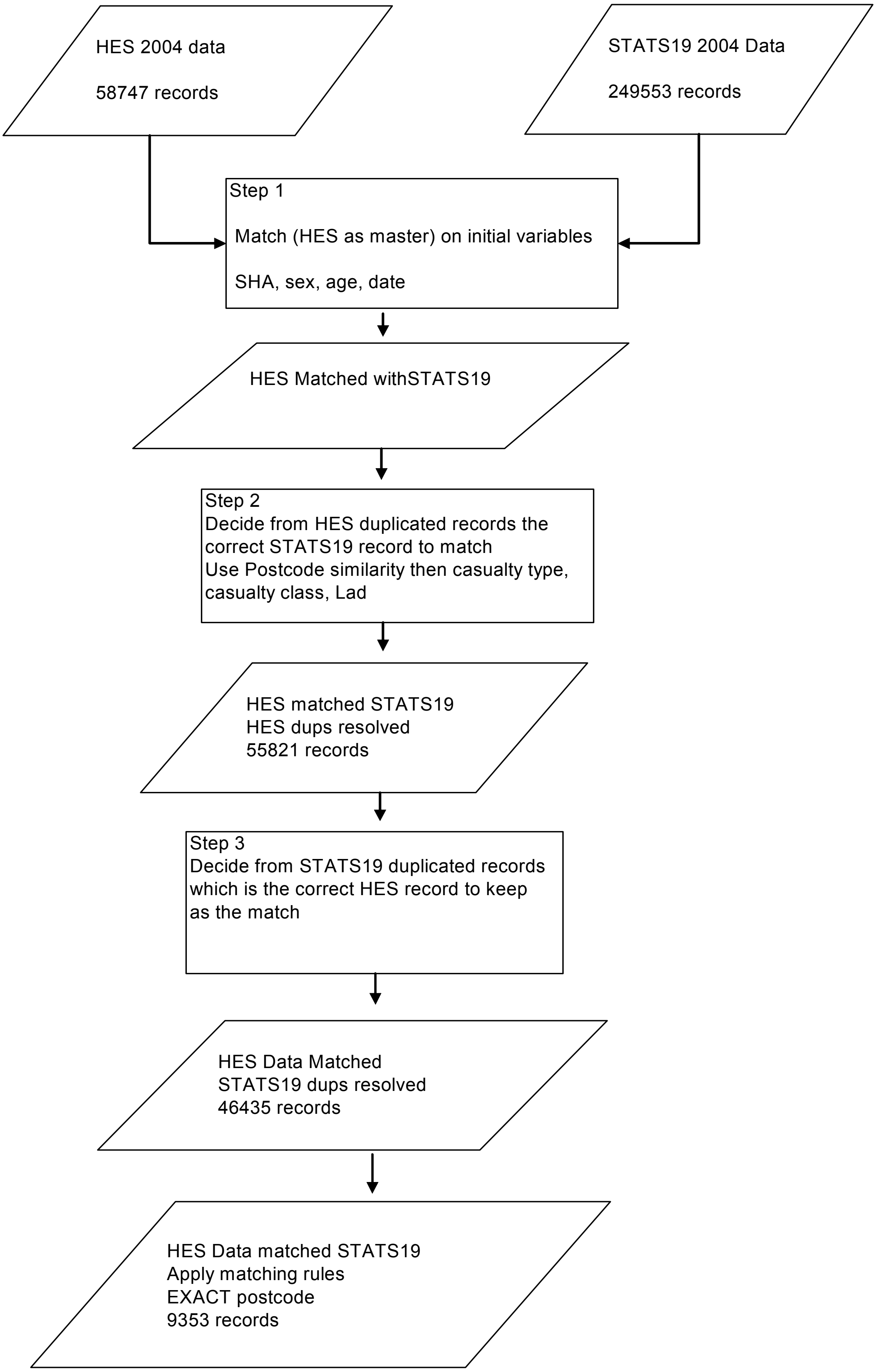


Figure 4.2 *Summary of matching rules, for matched outcome codes 1-9*

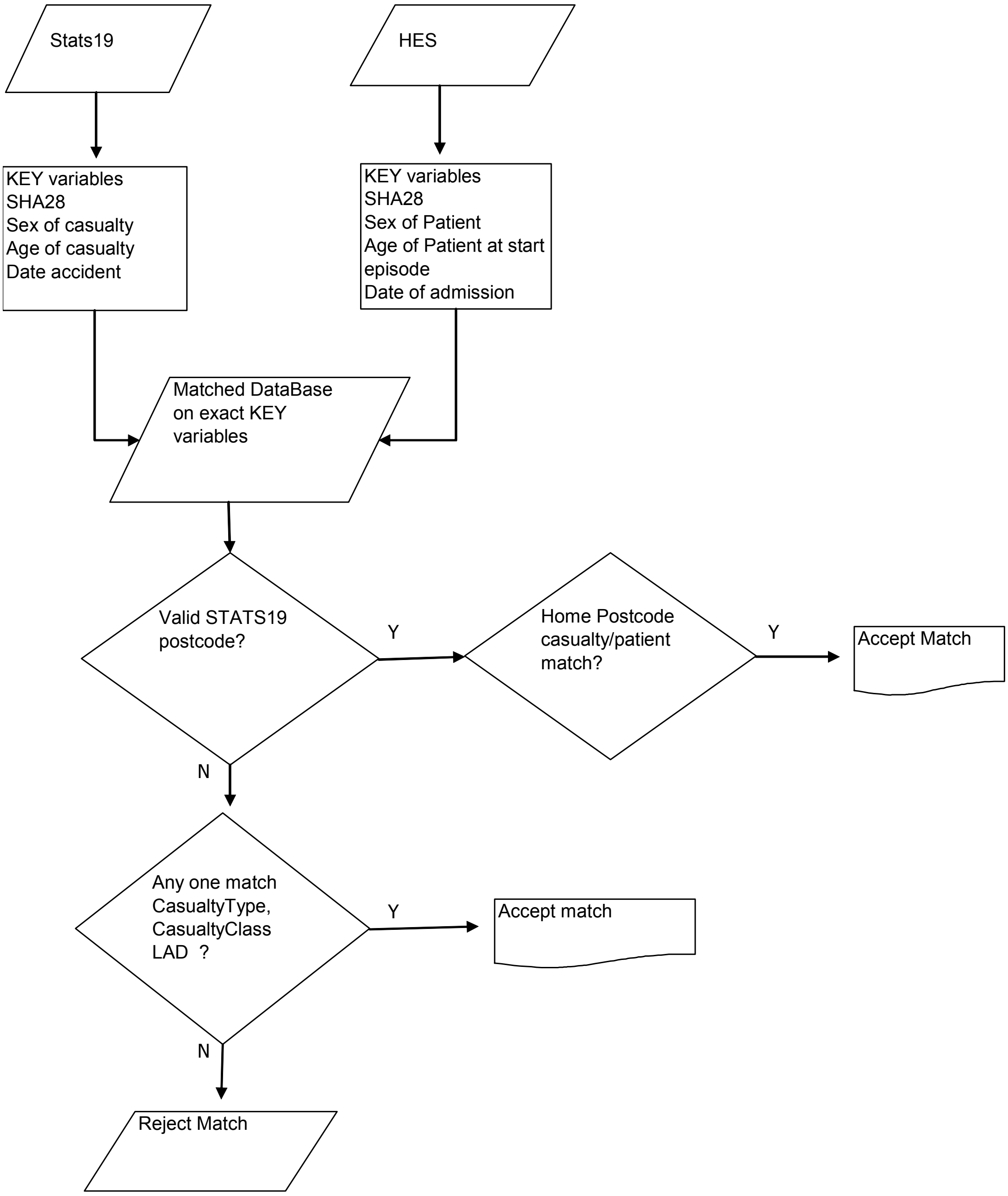


Figure 4.3 *Summary of matching rules, matched outcome codes 11-19*

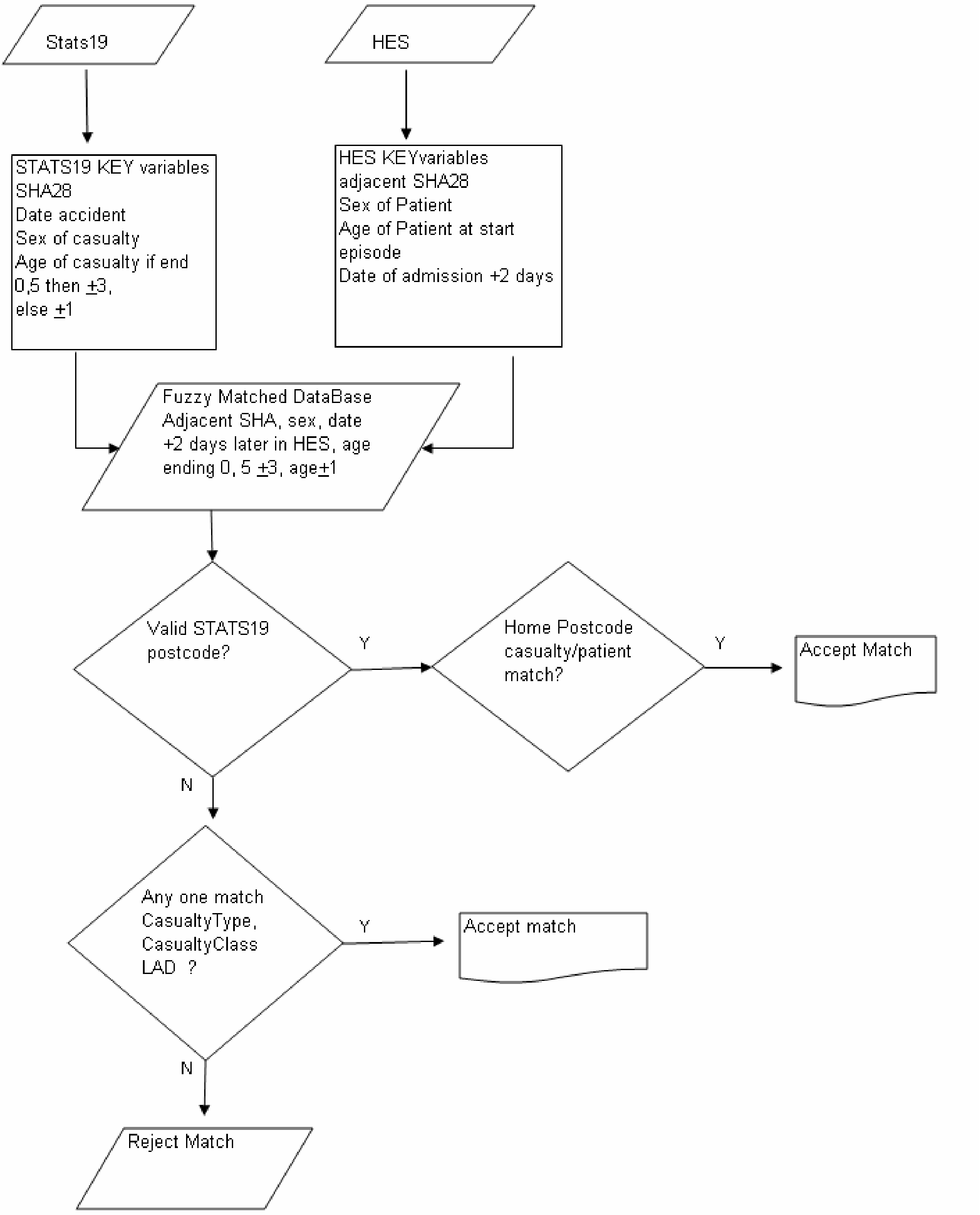
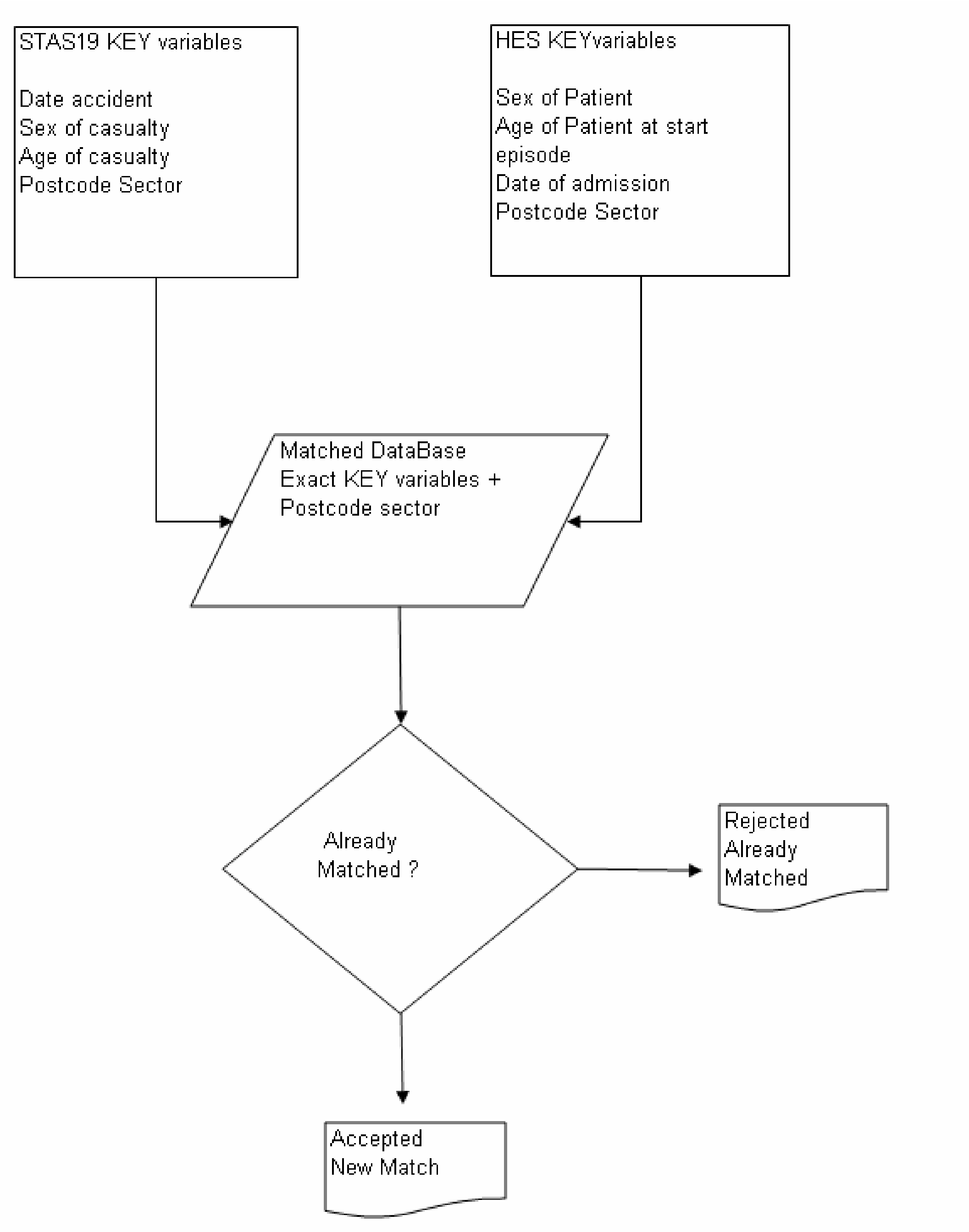


Figure 4.4 *Summary of matching rules, for matched outcome codes 20*



5. Results

The numbers of matches achieved by applying the various matching rules across the years 1996 – 2004 are given in Annex 2.

Taking 2004 as an example – 249,553 STATS19 records are initially matched against 58,747 HES records. At first the matching variables are the SHAs, age, sex and date, allowing tolerances in all variables so as to catch the widest number of possible matches. At this stage there are still many part-duplicates arising (as described in the previous section) due to many records within each of the datasets having identical values on these initial variables. As there should only be one STATS19 record matched to a corresponding HES record many of these initial matches are false. This is to be expected because the initial matching criteria are purposely set so as to catch too many rather than too few possible matches. They provide a pool of candidate matches which are then checked by comparing the matched values for the additional variables of postcode, and subsequently casualty class, casualty type and LAD. The priority and order of comparison of these variables is determined by following the matching rules drawn up with expert consultation.

The initial stage of matching produces 46,484 records which are possible matches. After applying the information supplied from the extra variables by using the matching rules 9353 records are considered to be matches with very high confidence. This is because they have an exact match on postcode, the most powerful of the matching variables. A further 4,551 records have a matching confidence of high as they fail to match on full postcode but have similarities in the first part of the postcode. A further 956 records are considered high given that they match on adjacent SHAs but are exact on all the other variables.

A further 1940 records are rejected as a match as they have no similarities in the postcode despite having a valid STATS19 postcode. For instances where there is no valid STATS19 postcode for matching then it is unfortunate to reject all such cases - some forces do not report postcode. Therefore the other variables are considered and if there is a match on casualty class, casualty type or, last in priority order LAD of accident with LAD of patient's home then the match is accepted otherwise if none of these variables match then the matches are rejected, (456 false positives). It is concluded that they are different as they have exact matching values on SHA, sex, age and date but do not match on their other values of postcode and the other variables as listed above This explains the matching outcome codes 1-9 (see Annex 2). The codes from 11 to 19 are derived from using fuzzy values for the initial matching variables, SHA, age and date then the same logic as above using casualty class, casualty type and LAD.

The percentage of HES records matching to a STATS19 record with very high confidence equates to a 16% matching rate. When all the matches classed as high are included this gives a rate close to 24%. These figures are considered by practitioners in the field as being an acceptable matching rate, similar to previous exploratory DfT matching exercises (DfT, 2006a). In principle all casualties admitted to hospital from a road traffic accident should be recorded in STATS19 but it appears that there are many HES records for which no STATS19 record can be found to correspond. It is well known that under-reporting in STATS19 is significant (DfT, 2006a) with studies indicating about twice as many casualties in road accidents as there are reported to and by the police, and under-reporting for particular types of vulnerable road user very much higher. Some police forces do not complete the postcode field and this reduces the matching rate. Where there are no valid STATS19 postcodes if the most probable correct matches found are included then the matching rate rises to 31%. It is probable the matching rate is higher for the police forces areas where the recording is most accurate.

Note that in the results table (Annex 2) the absence of the matches classified as high or very high in the years before 1999 is due to the absence of postcode collection in STATS19 prior to 1999. Also in 2004 an increase in very high matches is offset by a decrease in the medium category of likely matches. This indicates better quality postcode reporting by the police in recent years.

This paper is a description of the matching process rather than an explanation of any discrepancy between HES and STATS19 figures. It is hoped the matched database may provide evidence for researchers to explain any differences in accident rates. As an example of the type of analysis now possible on the matched database the following chart shows the numbers of serious injured from the STATS19 variable of severity of injury for matches thought to be highly likely to be correct.

Figure 5: *Number of matched records with high or very high confidence (top) showing those classified with a serious injury from STATS19, by year.*

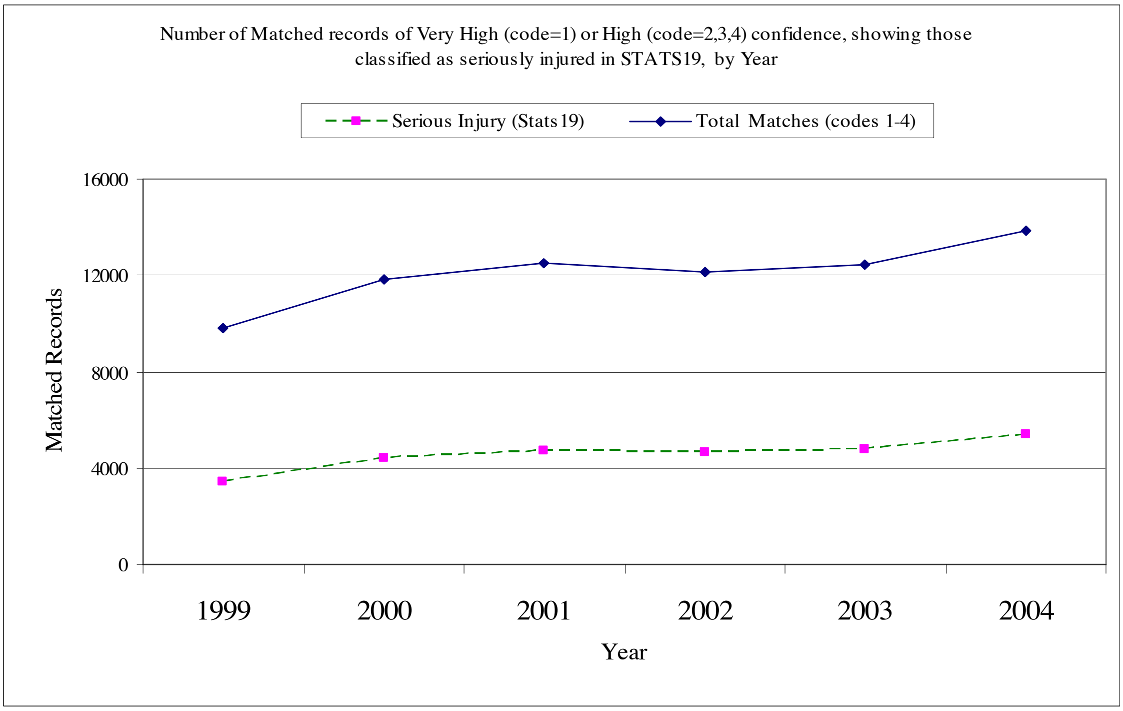


Figure 5 shows trends for the most highly probable matches and those matches for which the injury is reported as serious by STATS19. In theory all the matched records should have a severity of injury classification of "seriously injured". The difference between the two lines consists of accidents classified in STATS19 as not serious, i.e. those where the police record the injury as slight, even though the individual has been admitted to hospital. The proportions of seriously injured within the ‘very high’ matches are constant over time, very close to 38% in all years apart from 35% in 1999 in 2004. Why there are a large proportion of slightly injured is the type of question researchers need to examine.

6. Conclusions and future work

The one-to-one matching of the hospital admissions data (HES) and the road accident data (STATS19) is a practical illustration of how to match two datasets lacking a common matching key variable, (such as name and address). With the increasing awareness of the potential benefits to be gained through matching disparate databases but containing common units this is likely to be an increasingly common demand. The project illustrates how the ambiguous matches can be resolved by drawing up a set of matching rules and the convenience of using SAS/SQL for its implementation.

The implementation of this work provides an example of successful collaboration between different government departments in combining administrative data for public benefit. The resulting database provides a resource to researchers analysing the cause and circumstances of road accidents and may even be used within projects to help to understand national trends and to ultimately improve road safety – on average 9 people are killed in road crashes in Britain each day (DfT, 2006b).

References

Broughton J., Keigan M. (2007) Linking STATS19 and Scottish hospital in-patient data for the SafetyNet project, UK Transport Research Laboratory.

DfT (2006a) Road accident casualties: a comparison of STATS19 data with Hospital Episode Statistics,. Department for Transport, UK

DfT. (2006b) Road Casualties, Great Britain (RCGB) . Department for Transport, UK

Gill L. (2001) Methods for Automatic Record Matching and Linking and their use in National statistics, National Statistics Methodological Series No.25, UK Office for National Statistics

Gill M., Goldacre M. J., Yeates D. (2006) Changes in safety on England’s roads; analysis of hospital statistics, *British Medical Journal*.

Ward H., Robertson S., Townley K., Pedler A. (2007), Reporting of road traffic accidents in London: matching police STATS19 with hospital accident and emergency department data. UK Transport Research Laboratory.

ANNEX 1: HES-STATS19 matching rules

HES - STATS19 MATCHING RULES

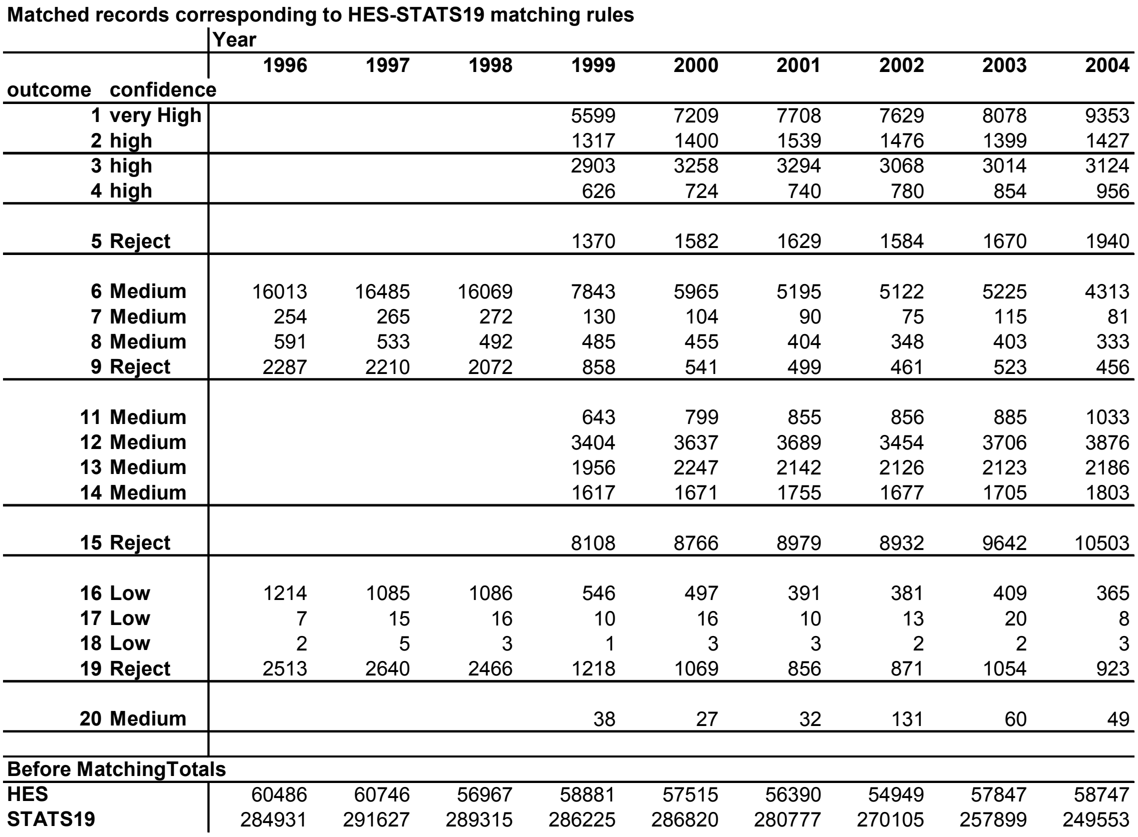
SHA (HA, acc) [blocking variable]	Sex	Age	Date (1)	Postcode	Casualty class	Casualty type	LAD (home, acc)	Matching confidence (2)	Outcome Code
Stage 1 - Exact match on SHA, age and date									
a) Valid postcode for matching (have letter in first char position)									
Exact	Exact	Exact	Exact	Exact - all chars	NA	NA	NA	Very high	1
Exact	Exact	Exact	Exact	Match first 2 chars (letters)	NA	NA	NA	High	2
Exact	Exact	Exact	Exact	Match any 2 of first 3 (if AA9 form) Match any 3 of first 4 (if AA99 form)	NA	NA	NA	High	3
Adjacent SHA	Exact	Exact	Exact	Exact - all chars	NA	NA	NA	High	4
Exact	Exact	Exact	Exact	No match	NA	NA	NA	Reject	5
b) No valid postcode for matching									
Exact	Exact	Exact	Exact	NA	Match	NA	NA	Medium	6
Exact	Exact	Exact	Exact	NA	No match	Match	NA	Medium	7
Exact	Exact	Exact	Exact	NA	No match	No match	Match	Medium	8
Exact	Exact	Exact	Exact	NA	No match	No match	No match	Reject	9
Stage 2 - Fuzzy match on SHA, age and date									
a) Valid postcode for matching (letter in first char position on both sources)									
Adjacent	Exact	STATS19 ± 3 (if 0,5) STATS19 ± 1 (else)	HES + 2	Exact - all chars	NA	NA	NA	Medium	11
Adjacent	Exact	STATS19 ± 3 (if 0,5) STATS19 ± 1 (else)	HES + 2	Match first 2 chars (letters)	NA	NA	NA	Medium	12
Adjacent	Exact	STATS19 ± 3 (if 0,5) STATS19 ± 1 (else)	HES + 2	Match any 2 of first 3 (if AA9 form) Match any 3 of first 4 (if AA99 form)	NA	NA	NA	Medium	13
Adjacent	Exact	STATS19 ± 3 (if 0,5) STATS19 ± 1 (else)	HES + 2	No match	NA	NA	NA		15
b) No valid postcode for matching									
Adjacent	Exact	STATS19 ± 3 (if 0,5) STATS19 ± 1 (else)	HES + 2	NA	Match	Match	NA	Low	16
Adjacent	Exact	STATS19 ± 3 (if 0,5) STATS19 ± 1 (else)	HES + 2	NA	Match	No match	Match	Low	17
Adjacent	Exact	STATS19 ± 3 (if 0,5) STATS19 ± 1 (else)	HES + 2	NA	No match	Match	Match	Low	18
Adjacent	Exact	STATS19 ± 3 (if 0,5) STATS19 ± 1 (else)	HES + 2	NA	Only one match for the three variables				19
Stage 3 - final sweep (records not already matched only)									
Any	Exact	Exact	Exact	Exact match for postcode sector	NA	NA	NA	Medium	20

Notes

(1) 'Exact' match on date means HES date of admission same as STATS19 date, or one day later

(2) If duplicates remain at any matching level, then include all records with matching confidence set to 'unacceptable'

ANNEX 2: Matched records corresponding to HES-STATS19 matching rules [7]



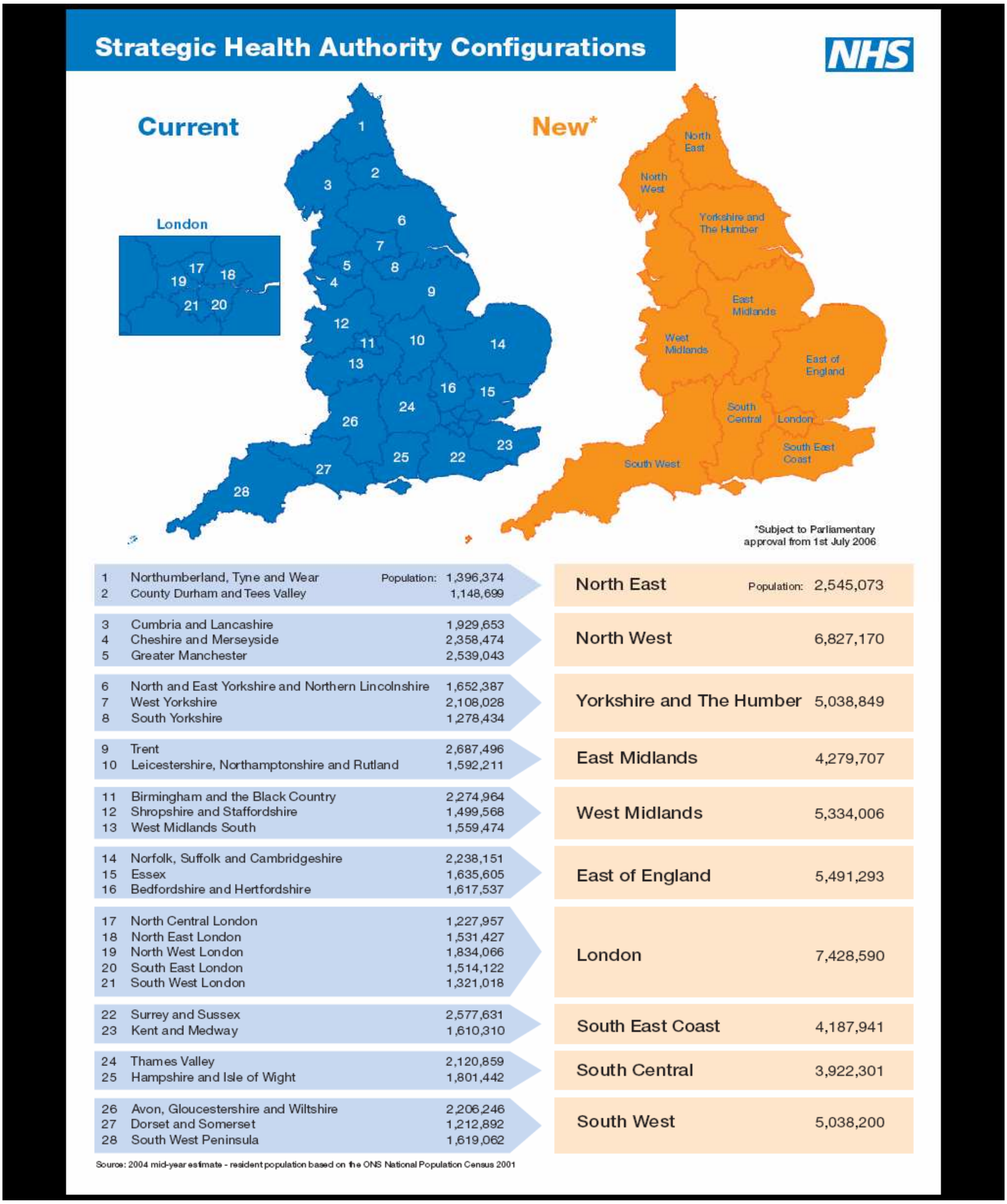
Matched records corresponding to HES-STATS19 matching rules

		Year								
		1996	1997	1998	1999	2000	2001	2002	2003	2004
outcome	confidence									
1	very High				5599	7209	7708	7629	8078	9353
2	high				1317	1400	1539	1476	1399	1427
3	high				2903	3258	3294	3068	3014	3124
4	high				626	724	740	780	854	956
5	Reject				1370	1582	1629	1584	1670	1940
6	Medium	16013	16485	16069	7843	5965	5195	5122	5225	4313
7	Medium	254	265	272	130	104	90	75	115	81
8	Medium	591	533	492	485	455	404	348	403	333
9	Reject	2287	2210	2072	858	541	499	461	523	456
11	Medium				643	799	855	856	885	1033
12	Medium				3404	3637	3689	3454	3706	3876
13	Medium				1956	2247	2142	2126	2123	2186
14	Medium				1617	1671	1755	1677	1705	1803
15	Reject				8108	8766	8979	8932	9642	10503
16	Low	1214	1085	1086	546	497	391	381	409	365
17	Low	7	15	16	10	16	10	13	20	8
18	Low	2	5	3	1	3	3	2	2	3
19	Reject	2513	2640	2466	1218	1069	856	871	1054	923
20	Medium				38	27	32	131	60	49
Before MatchingTotals										
HES		60486	60746	56967	58881	57515	56390	54949	57847	58747
STATS19		284931	291627	289315	286225	286820	280777	270105	257899	249553

[8] Note effect of quality of Postcode on matching outcome.
In STATS19 postcode has been collected only since 1999, explains the lack of any matches before.
Increase of Very High category to 9353 in 2004 suggests improvement in STATS19 Postcode reporting.

Increase in 2004 very high category probably due to improved postcode recording on police records.

ANNEX 3: Maps of the 28 Strategic Health Authorities in England. [8]



Strategic Health Authority Configurations

*Subject to Parliamentary approval from 1st July 2006

		Population:		Population:
1	Northumberland, Tyne and Wear	1,396,374	North East	2,545,073
2	County Durham and Tees Valley	1,148,699		
3	Cumbria and Lancashire	1,929,653	North West	6,827,170
4	Cheshire and Merseyside	2,358,474		
5	Greater Manchester	2,539,043		
6	North and East Yorkshire and Northern Lincolnshire	1,652,387	Yorkshire and The Humber	5,038,849
7	West Yorkshire	2,108,028		
8	South Yorkshire	1,278,434		
9	Trent	2,687,496	East Midlands	4,279,707
10	Leicestershire, Northamptonshire and Rutland	1,592,211		
11	Birmingham and the Black Country	2,274,964	West Midlands	5,334,006
12	Shropshire and Staffordshire	1,499,568		
13	West Midlands South	1,559,474		
14	Norfolk, Suffolk and Cambridgeshire	2,238,151	East of England	5,491,293
15	Essex	1,635,605		
16	Bedfordshire and Hertfordshire	1,617,537		
17	North Central London	1,227,957	London	7,428,590
18	North East London	1,531,427		
19	North West London	1,834,066		
20	South East London	1,514,122		
21	South West London	1,321,018		
22	Surrey and Sussex	2,577,631	South East Coast	4,187,941
23	Kent and Medway	1,610,310		
24	Thames Valley	2,120,859	South Central	3,922,301
25	Hampshire and Isle of Wight	1,801,442		
26	Avon, Gloucestershire and Wiltshire	2,206,246	South West	5,038,200
27	Dorset and Somerset	1,212,892		
28	South West Peninsula	1,619,062		

Source: 2004 mid-year estimate - resident population based on the ONS National Population Census 2001

[8] http://www.statistics.gov.uk/geography/england_health.asp

Use of administrative data in SBS survey

Renata Tumėnienė
Statistics Lithuania, Gedimino ave 29, LT-01500 Vilnius, Lithuania
e-mail: renata.tumeniene@stat.gov.lt

Abstract: In Statistics Lithuania the SBS survey is based on a statistical questionnaire and administrative data. Administrative data is used since 1999. The main sources of the administrative data are: Annual financial statement of enterprises, Annual profit taxes of enterprises (Individual enterprises), Natural persons who are engaged in an economic activity in their own right (Natural persons), State social insurance fund board (SODRA). These data are used to update the active enterprise list, to estimate small businesses and nonresponse enterprises and to create a final survey frame.

This paper documents treatment of administrative data and combination with the data of the SBS survey.

Keywords: SBS survey, Administrative data, Editing methods.

1. Introduction

The aim of this paper is to provide a description of the process of using administrative data for the SBS survey.

Performance of the SBS survey is regulated by: Republic of Lithuania law on the amendment of the law on statistics (23 December 1999 no VIII-1511), COUNCIL REGULATION (EC, EURATOM) No 58/97 of December 1996 concerning Structural Business Statistics. The amendment of this regulation is adopted on 2008 March 11 and Resolution of the Government of the Republic of Lithuania No 569 of 16 May 2001 on the implementation of the European System of Accounts. The objective of the SBS survey is to full fill the regulation requirements that are to prepare indicators about structure of the businesses according to the NACE classification, size classes and regions.

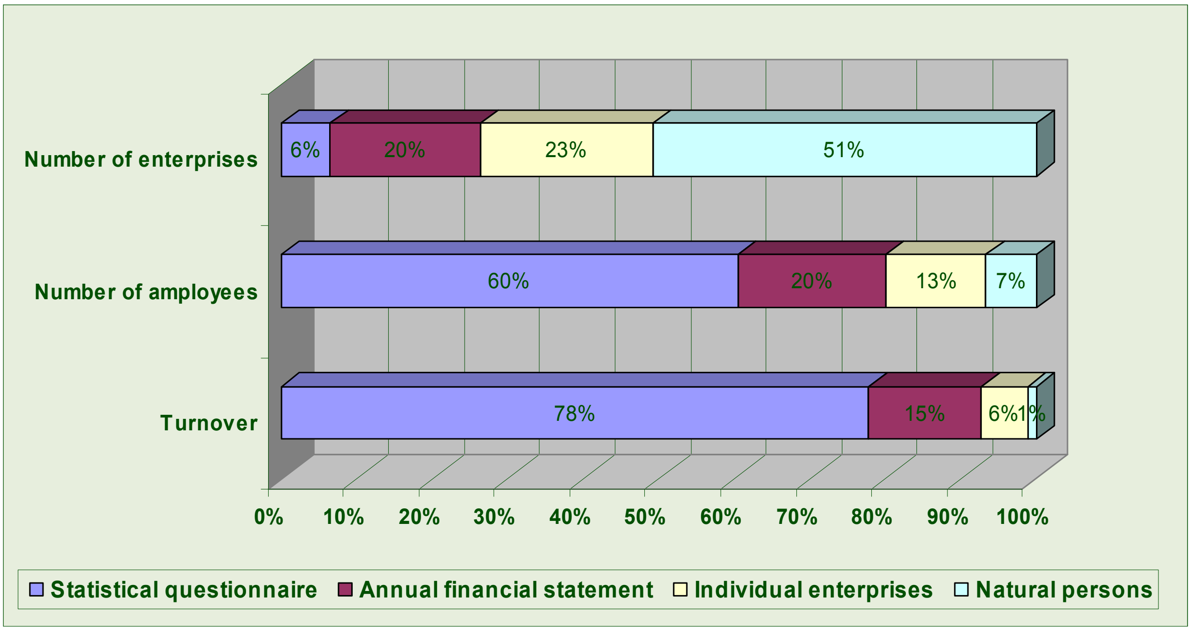
In Lithuania SBS survey is a census survey and consists of data of statistical questionnaire and of three sources of administrative data: Annual financial statements of enterprises, Individual enterprises and Natural persons. Data of SODRA is used for the imputation of the number of employees and salary. All administrative sources are used to update the active enterprise list, to estimate small businesses and nonresponse enterprises and to create a final survey frame. Different data sources can be merged by unique enterprise identification code. Number of enterprises collected via statistical questionnaire covers 6% of the total SBS survey enterprises while turnover of these enterprises covers about 80% of the total SBS turnover. The following sections will describe the editing methods of the annual financial statements and the combination of these statements with the statistical questionnaire. The editing procedure of Individual enterprises is not discussed here, because the editing procedure of these enterprises is very similar to the editing procedure of the financial statements. The Natural persons are not edited, because we get information about their income only.

2. SBS survey calendar for the reference year 2005

The time table of the SBS survey for the reference year 2005:

- September – October 2005 – list of active enterprise is created
- November 2005 – the enterprises for the survey is selected
- December 2005 – February 2006 – statistical questionnaire is revised and confirmed
- March 2006 – statistical questionnaires are sent to the enterprises
- May 2006 – enterprises delivers complete statistical questionnaire to the territorial statistical offices
- June – July 2006 – data is entered to the statistical data base and checked for the errors
- August 2006 – preliminary data are calculated
- September 2006 – preliminary data are delivered to the National accounts division
- October – November 2006 – preliminary data are prepared and delivered to the EUROSTAT
- January 2007 – administrative data are received
- January – May 2007 –administrative data are edited and missing variables of the statistical questionnaire are estimated
- June 2007 – definitive SBS data are prepared and sent to the EUROSTAT
- July 2007 – data are prepared and provided to the users

Figure 1: *Distribution of the main SBS survey indicators according to the administrative sources*



3. Editing methods of annual financial statements

Due to errors of various types data of annual financial statement must pass an editing procedure. Annual financial statement consists of Profit and Loss account and Balance sheet. Data editing procedure is automated and implemented by SAS software. A correct record must pass an edit rule, a linear equality based on accounting identities. Incorrect records must be corrected using various mathematical methods. Development of editing methods has been a continuous process.

The following editing methods are applied for the annual financial statement: **Edit rule, Sign checking, Locating the error, Outlier detection, Re-scaling, Donor based editing**.

The **Edit rule** determines whether a record is correct or not. It is a logical condition or a restriction to the value of a data item which must be met if the data is to be considered correct. During this method all incorrect records are flagged for further treatment.

Example: Bellow is the table of the Profit and Loss account which displays how the edit rule works.

Table 1: *Profit and Loss account*

VARIABLE CODE	*EDIT RULE*	*VARIABLE SIGN*	*NAME OF THE VARIABLE*
X_1	+	+	SALES
X_2	-	+	COST OF SALES
Y_1	Sum(X_1: X_2)	+/-	GROSS PROFIT (LOSS)
X_3	-	+	OPERATING EXPENSES
Y_2	Sum(X_1: X_3)	+/-	OPERATING PROFIT (LOSS)
X_4	+	+/-	OTHER ACTIVITIES
X_5	+	+	INCOME OF FINANCIAL AND INVESTING ACTIVITIES
X_6	-	+	EXPENSES OF FINANCIAL AND INVESTING ACTIVITIES
Y_3	Sum (X_1: X_6)	+/-	ORDINARY PROFIT (LOSS)
X_7	+	+	EXTRAORDINARY GAIN
X_8	-	+	EXTRAORDINARY LOSS
Y_4	Sum(X_1: X_8)	+/-	PROFIT (LOSS) BEFORE TAXATION
X_9	-	+	CORPORATE INCOME TAX
Y	Sum (X_1: X_9)	+/-	NET PROFIT (LOSS)

Variables from X_1 to X_9 must satisfy the following linear equation:

$$X_1 - X_2 - X_3 + X_4 + X_5 - X_6 + X_7 - X_8 - X_9 = Y \qquad (1)$$

Then the value of error is calculated:

$$e = (X_1 - X_2 - X_3 + X_4 + X_5 - X_6 + X_7 - X_8 - X_9) - Y \qquad (2)$$

The record is considered correct if the value of error is equal to zero. After the edit rule is applied all the correct records are flagged with the corresponding Flag.

For the remaining incorrect records the **Sign checking** method is applied. With this method variables which can have either a positive or a negative value are checked whether the sign is correct.

Example: In Table 1 the variable X_4 (Other activities) can have either a positive or a negative value (SIGN column indicates what sign can gain the corresponding variable). With this method we change the sign of the variable X_4 to the reverse sign:

$$X_4^* = -X_4 \tag{3}$$

After the sign of the variable is changed the edit rule is applied to check whether a record is correct or not:

$$(X_1 - X_2 - X_3 + X_4^* + X_5 - X_6 + X_7 - X_8 - X_9) – Y = 0 \tag{4}$$

If the (4) equation is true then the sign of the variable X_4 is changed to the reverse sign X_4^*. All the incorrect records which were corrected with this method are flagged with the corresponding Flag.

For the rest of the remaining incorrect records the **error is located** to a certain part of annual financial statement by the use of subtotals.

Example: In Table 1 using subtotals Y_1, Y_2, Y_3 and Y_4 a certain erroneous variables can be detected by applying the following equations:

$$\begin{aligned} &X_1 - X_2 = Y_1 \\ &X_1 - X_2 -X_3 = Y_2 \\ &X_1 - X_2 -X_3 + X_4 + X_5 - X_6 = Y_3 \\ &X_1 - X_2 -X_3 + X_4 + X_5 - X_6 + X_7 - X_8 = Y_4 \end{aligned} \tag{5}$$

Thus, the following equations can be derived:

$$\begin{aligned} &Y_1 – X_3 + X_4 + X_5 - X_6 + X_7 - X_8 - X_9 = Y \\ &Y_2 + X_4 + X_5 - X_6 + X_7 - X_8 - X_9 = Y \\ &Y_3 + X_7 - X_8 - X_9 = Y \\ &Y_4 - X_9 = Y \end{aligned} \tag{6}$$

And the following conditions can be tested:

$$\begin{aligned} &Y_1 – X_3 + X_4 + X_5 - X_6 + X_7 - X_8 - X_9 - Y = 0 \\ &Y_2 + X_4 + X_5 - X_6 + X_7 - X_8 - X_9 - Y = 0 \\ &Y_3 + X_7 - X_8 - X_9 - Y = 0 \\ &Y_4 - X_9 - Y = 0 \end{aligned} \tag{7}$$

If first condition of the (7) is not true then it is assumable that the error is located in variables X_1, X_2. If second condition is not true then it is assumable that the error is located in variables X_1, X_2, $X3$ and so on.

Also the other conditions of the Profit and Loss account can be tested to identify erroneous variables. For instance:

If $X_1 - X_2 = Y_1$ and $Y_1 - X_3 \neq Y_2$ and $Y_2 + X_4 + X_5 - X_6 = Y_3$ then it is assumable that variable X_3 is erroneous.
If $Y_1 - X_3 = Y_2$ and $Y_2 + X_4 + X_5 - X_6 \neq Y_3$ and $Y_3 + X_7 - X_8 = Y_4$ then it is assumable that error is located in variables X_4, X_5, X_6 and the like.

Outlier detection: the incorrect set of variables is compared to the distribution of corresponding variables of the correct records in the respective activity. This method is used to identify and correct big errors in one variable. In this method the values of all set of variables (correct and incorrect) are presented in relation to turnover (X_1):

$$S_i = \frac{X_i}{X_1} \qquad (8)$$

Then the distributions of these ratios are calculated and 1st ($D1$) and 9th ($D9$) deciles are selected as a threshold values. Suspicious values out of this target range may contain an error. The relative error is calculated:

$$S_e = \frac{e}{X_1} \qquad (9)$$

If a value of ratio (8) is out of target range, it is tested whether the value moves inside the target range after adjusting it by the error e.
When value of error is positive, for the negative variables the following conditions are tested:

$$S_i < D_1(S_i) \text{ and } D_1 \le S_i + S_e \le D_9 \qquad (10)$$

When both conditions are true, the error is adjusted to that particular variable: $X_i^* = X_i + e$.
For positive variables the following conditions are tested:

$$S_i > D_9(S_i) \text{ and } D_1 \le S_i - S_e \le D_9 \qquad (11)$$

When both conditions are true, the error is adjusted to that particular variable: $X_i^* = X_i - e$.

When value of error is negative, for negative variables the following conditions are tested:

$$S_i > D_9(S_i) \text{ and } D_1 \le S_i + S_e \le D_9 \qquad (12)$$

And for the positive variables the following condition are tested:

$$S_i < D_1(S_i) \text{ and } D_1 \le S_i - S_e \le D_9 . \tag{13}$$

Commonly this method allows correcting inaccurate typing mistakes. When the operator who enters the data to the computer can by accident type too much or to less figures then this method is very effective. All variables corrected with outlier detection method are marked with the corresponding flag.

After all the editing methods above-named are applied, the remaining incorrect records are divided into two groups determined by their relative error. A relative error of ± 5% of turnover is used as a threshold. When a record contains a relatively small error, less than ± 5% of turnover, the incorrect set of variables are **re-scaled**. The incorrect set of variables is multiplied by a scaling factor to the level of the record. The error is distributed to all variables belonging to the incorrect set. The scaling factor is the error divided by the sum of the incorrect set of variables (*E*):

$$k = \frac{e}{\sum_{i \in E} |X_i|} \tag{14}$$

Every incorrect variable is multiplied by the scaling factor:

$$\begin{aligned} X_i^* &= (1-k) \cdot X_i, \ \ when \ X_i \ge 0 \\ X_i^* &= (1+k) \cdot X_i, \ \ when \ X_i < 0 \end{aligned} \tag{15}$$

For the rest of the incorrect records containing error bigger than ± 5% of a relative error a **donor unit** is determined and the incorrect part is estimated by a data structure of a corresponding variable of a donor unit. Past information is used as a primary donor. Nearest neighbor is used as the donor if past information is not available.
The incorrect set of variables is estimated by a data structure of the same enterprise of a previous year:

$$X_i^* = Y \cdot \frac{X_i^{past}}{Y^{past}}, \tag{16}$$

where *Y* is a corresponding subtotal of an annual financial statement.

Nearest neighbor is used as the donor if past information is not available. It is selected from a group consisting of correct records in the respective activity. The distance measure between two variables is:

$$D_i = MIN\left\{ \sum_{k \in F} \left| X_{ik} - X_{ik}^{near} \right| \right\} \tag{17}$$

where *F* is a set of variables selected for comparison.

The incorrect set of variables is estimated by a data structure of a nearest neighbor:

$$X_i^* = Y \cdot \frac{X_i^{near}}{Y^{near}}, \tag{18}$$

where *Y* is a corresponding subtotal of an annual financial statement.

The donor editing method is more suitable for the balance sheet, as for the Profit and Loss account the logical edits are more relevant. The Logical edits are based on the accounting identities and erroneous variables are logically calculated.

4. Combination of administrative data with the data of statistical questionnaire

When all the data of administrative sources are corrected then these data are combined with the data of statistical questionnaire. Annual financial statement of enterprises includes only part of necessary variables for the statistical questionnaire. Statistical questionnaire consists of 10 sections (about 500 variables). Three sections (Assets; Equity and liabilities; Sales, expenses, profit) are estimated by using Annual financial statement, one section (Employees) by State social insurance fund board and remaining six sections are estimated by structural coefficients of the enterprises or by data structure of the donor units.

There are 2 types of forms of Profit and Loss account and Balance sheet: short form and full form. Full form contains about 200 variables while short form contains only about 55 variables. About 60% of the data of the long form is used for the estimation of the three sections above-named of the statistical questionnaire and about 80% of the data are used from the short form.

Combination of administrative data with the data of statistical questionnaire is divided into two parts. One part consists of estimation of the variables which are available in administrative data and the other part consists of variables which are not available in administrative data.

For the variables which are available in administrative data the following data combination procedure is applied:

1) Variables of the annual financial statement which directly corresponds with variables in statistical questionnaire are directly transferred to the statistical questionnaire.

2) For the evaluation of the missing variables of the statistical questionnaire which do not have directly correspondence with annual financial statement the optimal model based estimation method is detected and missing variables are estimated.

3) Variables of the statistical questionnaire are estimated using mathematical methods and structural coefficients.

4) For the estimation of the missing variables it is used:

4.1) Structural coefficients of the data of the same enterprise from the previous year.

4.2) Donor values (nearest neighbor method) or the structural coefficient of the donor data.

4.3) Structural coefficient calculated from the data of the respondent enterprises (statistical questionnaire) grouped by various classes.

The other combination procedure is applied for the variables which are not available in the administrative data:

1) Enterprises which do not deliver annual financial statement to the tax authorities, but their turnover from the other sources are known, variables needed for the statistical questionnaire are estimated as follows:

1.1) Previous year data of the same enterprise are multiplied by the alteration coefficient of the turnover.

1.2) Values of donor (nearest neighbor by turnover) data are imputed.

When the combination procedure is over, the final data base of the statistical survey is formatted.

5. Advantages and disadvantages of using administrative data in SBS survey

The best advantage of using administrative data for the SBS survey is a reduction of a response burden for the small enterprises. It also allows significantly reduce the costs of the survey, to update a list of active enterprises, to have a data by every enterprise. The administrative data gives good estimates for the most important variables, so the better precision we can have.

The best disadvantage of using administrative data is the scope of information. It is too small for the SBS needs. Lot's of variables need to be imputed and the risk of introducing model assumption errors occurs.

Integration of different data sources in the international migration statistics in Hungary

Éva Gárdos, Annamária Sárosi, Áron Kincses
Hungarian Central Statistical Office (HCSO), Budapest, Keleti K. u. 5-7, 1024
e-mail: eva.gardos@ksh.hu

Abstract: In 2002 a new voluntary statistical survey was introduced in the Hungarian international migration statistics. "The form of the acquisition of the Hungarian Citizenship" serves to collect data that complete those coming from administrative data source (population register). As the 23-27% of the new citizens do not fill in the statistical questionnaire the linkage of the two data sources is performed by using RASH-method.

Keywords: international migration statistics, acquisition of citizenship, administrative register, statistical population survey, RAS-method.

1. Introduction

The aim of this paper is to provide an informative description on the purpose, method and way of the linkage of an administrative register and a statistical population survey in the Hungarian migration statistics.

Regular publication of migration statistics in Hungary began in 1993. The HCSO is responsible for the compilation of that. The first applied major data sources were the population register and the register of residence (settlement) permits. Later on they were augmented with the register data on people acquired Hungarian citizenship and on refugees, respectively.

The yearly number of immigrants who acquire the Hungarian citizenship varies between 1-10 thousands. In the period of 1993-2006 the number of new citizens was all together 106 707, that is about 43% of the immigrants and more than one percent of the whole population. It is important to know why these people want to live in Hungary, what family background they have, what are their social conditions etc. However, there are quite a few pieces of information on them in the administrative data sources. From the year of 2002 a statistical data collection contributes to the data set of new citizens.

2. Data sources

Register of personal data and addresses (Population Register)

In the recent years the population register provides the administrative data on the naturalized people.

In Hungary a person may have a place of residence and additionally a place of stay. Place of residence ("permanent place of residence") is the address of the dwelling where the person lives. Place of stay ("temporary place of residence") is the address where a person stays longer than 3 months without an intention to leave finally a place of residence.

On the basis of people's declarations on permanent and temporary place of residence the Hungarian population register includes the following categories:

- Hungarian citizens having permanent residence (domicile) in Hungary,
- Hungarian citizens having permanent residence abroad (living abroad or living temporarily in Hungary) who asked to be registered,
- foreigners with permanent residence permits (including refugees),
- EEA citizens with residence permits.

Thus, it does not comprise every person entering or leaving the country. Foreigners staying in Hungary temporarily (i.e. foreigners with a residence visa or a "temporary" residence permit, foreigners with a certificate entitling for temporary stay, foreign diplomats and asylum seekers) and the overwhelming part of the Hungarian citizens staying abroad are not included. Moreover, the population register covers very limited information on the included people, not making possible to explore even the fundamental characteristics of the migrant population. Consequently the Hungarian migration statistics are developed on the basis of many administrative and statistical data sources and the population register is merely one of them.

Considering those who apply for the Hungarian citizenship their application is accepted by the Office of Immigration and Nationality (OIN). Following a positive decision the applicant has to take oath. The registrar of the settlement where the new citizen lives informs the population register in an electronic format on the data of the new Hungarian citizen. The Population Register Office forwards the data to the statistical office.

The register contains data as follows: names, mother's name, date of birth, place of birth, country of birth, sex, citizenship, address, date of registration, date of log out of the register, cause of registration, cause of log-out, family status, and data of acquisition of the Hungarian citizenship. These data can be considered of rather good quality, because the population register is the base of parliament elections and referendums.

Statistical survey on people acquired Hungarian citizenship

Connected to taking oath the applicant is asked to fulfil statistical forms on himself/herself and on the minor child(ren) coming with. The data provision is voluntary. The survey contains the data as follows: names, mother's name, date of birth, place of birth, country of birth, sex, family status, mother tongue, number of children, previous citizenship, educational attainment, reason of application, address, economic activity and occupation before entering Hungary, current economic activity and occupation and the date of the acquisition of the Hungarian.

The questionnaires are spread among the local governments by the HCSO and the completed ones are received by the Demographic Competence Centre of the HCSO. The data are entered directly to a central data base.

The completing-ratio of the fields in the statistical questionnaire is 100% or almost that. The lowest ratio is measured in the case of the educational attainment: 89.7%.

3. Comparison of the two data sources

Before linking of the two datasets it was investigated if they really comprise data on the same people. This was especially important before the first merging. The closeness of

the relationship between the two databases was measured with the linear correlation coefficient and the elasticity coefficient. (Hunyadi, L. – Vita, L. 2004)
The linear correlation coefficient is as follows:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}},$$

where x_i, and y_i are the values of the variables in the two different data sources while $\overline{x}$, and $\overline{y}$ are the related averages and -1 < r_{xy} < +1.
The elasticity coefficient is as follows:

$$E = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \frac{\overline{x}}{\overline{y}},$$

that indicates that in the given relationship, how much change of the variable y is caused if the variable x changed by 1%.

In the year 2002 altogether 2451, that is 73% of the new Hungarian citizens filled in the statistical form compared to the 3369 people registered in the administrative data source (Kincses, Á. 2003). It may be considered a quite good proportion taking into account, that the statistical data provision is voluntary. In the following years the survey produces records of the similar proportion compared to the registered cases.

Table 1: *The proportion of the completed statistical forms compared to the number of registered cases*

Year	Number of		Coverage ratio (%)
	statistical forms	registered cases	
2002	2451	3369	72.75
2003	4046	5261	76.91
2004	3965	5432	72.99
2005	7542	9870	76.41
2006	4509	6101	73.91

Three common variables were used in the comparison: previous citizenship, county of the place of residence and completed age.

1. In the case of the previous citizenship only the European countries were considered due to the low number of cases outside Europe and due to their less reliability than the European ones. The coefficients are as follows: r_{xy} = *0.9971* and *E=1.11*.

Table 2: *People acquired Hungarian citizenship by previous citizenship and sex, 2002*

Country of the previous citizenship	Male		Female		Together	
	Survey	Register	Survey	Register	Survey	Register
Romania	700	1020	835	1218	1535	2238
Yugoslavia	209	265	178	222	387	487
Ukraine	135	141	184	199	319	340
EU	10	14	7	10	17	24
Other European	47	75	70	140	117	215
Together	**1101**	**1515**	**1274**	**1789**	**2375**	**3304**
Others	**47**	**41**	**29**	**24**	**76**	**65**
Altogether	**1148**	**1556**	**1303**	**1813**	**2451**	**3369**

2. Considering the frequencies by the county of the place of residence the ratio of the covered register cases by the survey records is very different running through the counties. The proportion varies between 42% and 100%. Nevertheless the coefficients show closed relationship, $r_{xy} = 0.9851$ and $E=0.9278$.

Table 3: *People acquired Hungarian citizenship by county of place of residence and sex, 2002*

County	**Male**		**Female**		**Together**	
	Survey	**Register**	**Survey**	**Register**	**Survey**	**Register**
Budapest	322	393	361	468	683	861
Baranya	31	66	25	67	56	133
Bács-K	39	39	29	33	68	72
Békés	45	67	61	69	106	136
Borsod-A-Z	42	41	43	53	85	94
Csongrád	104	122	109	128	213	250
Fejér	55	73	70	88	125	161
Győr-M-S	48	77	53	89	101	166
Hajdú	44	87	49	96	93	183
Heves	18	22	22	33	40	55
Komárom-E	35	40	40	54	75	94
Nógrád	4	8	10	15	14	23
Pest	136	220	161	267	297	487
Somogy	15	35	22	47	37	82
Szabolcs-Sz-B	89	116	126	141	215	257
Jász-N-Sz	19	18	20	21	39	39
Tolna	26	33	22	30	48	63
Vas	32	33	29	35	61	68
Veszprém	19	30	32	42	51	72
Zala	25	36	19	37	44	73
Together	**1 148**	**1556**	**1 303**	**1813**	**2 451**	**3369**

3. Looking at the differences of the two data bases by the distribution of age groups the coverage rate is the lowest among the 0-14 year old children where the value is merely 42%. It is to be caused by the fact that the parents much less frequently

fill in the form for their minor sons or daughters than for themselves. As a consequence the average age in the register is 37.93 years while in the dataset of the statistical survey is 40.58. Thus, the difference is over 2.5 years.
The closeness of the relationship between the two data sets in this case is less than that was by the previous citizenship or by the county of the Hungarian address: $r_{xy} = 0.9503$ and $E=0.8927$.

Table 4: *People acquired Hungarian citizenship by age group and family status, 2002*

Age group	Never married		Married		Widowed		Divorced		Together	
	Survey	Register	Survey	Register	Survey	Register	Survey	Register	Survey	Register
0–14	185	466	0	0	0	0	0	0	185	466
15–19	88	148	0	0	1	0	0	0	89	148
20–24	90	119	19	33	0	0	0	1	109	153
25–29	168	195	183	249	0	1	9	13	360	458
30-–39	142	178	491	632	5	4	50	54	688	868
40–49	28	24	213	282	6	9	41	38	288	353
50–59	8	4	159	203	25	23	27	37	219	267
60–X	15	15	357	431	107	164	34	46	513	656
Sum	724	1 148	1 422	1 830	144	202	161	189	2 451	3 369

Summarizing the above results it can be stated that the two data sets are not independent, moreover they are highly correlated. We may suppose that the records of the statistical survey are covered by the cases in the administrative data source. With a trial for a record linkage 96.2% of the 2002 survey cases were found in the administrative data source, however this test was performed before having received all the administrative records. Thus, it can be supposed that the real ratio is even higher.

4. Linking of data arising from two different data sets

As the data base of the statistical survey will never cover the 100% of the cases involved into the register it seems reasonable to use the RAS method for linking the two data sets, rather than apply a record linkage. (Stoyan, G. – Takó, G, 1993, Kincses, Á, 2003, 2004) Let us consider a two-dimensional table of the statistical survey. One of the variables is common in the two data sets. The variable of the column is the common variable in the two data sets and the variable of the row is included only the data base of the statistical survey. The elements of the table are as follows:

$$\begin{array}{ccccc|c}
a_{11} & a_{12} & . & . & a_{1n} & a_{1.} \\
a_{21} & . & . & . & a_{2n} & a_{2.} \\
. & . & . & . & . & . \\
. & . & . & . & . & . \\
. & . & . & . & . & . \\
a_{m1} & a_{m2} & . & . & a_{mn} & a_{m.} \\
\hline
a_{.1} & a_{.2} & a_{.3} & . & a_{.n} & a
\end{array}$$

a_{ij} denotes the element in the cross of the row i and of the column j.

$a_{.j} = \sum_{k=1}^{m} a_{kj}$, $\forall j \in \{1,2,\ldots\ldots,n\}$, and $a_{i.} = \sum_{l=1}^{n} a_{il}$ $\forall$ $i \in \{1,2,\ldots\ldots,m\}$, and: $a = \sum_{b=1}^{m}\sum_{c=1}^{n} a_{bc}$ $\forall b,c \in \mathrm{N}$.

The RAS method modifies the elements of the table above in a way that the inner proportions will remain the same and at the same time the table will fit to the register data set.

As the first step that the column values will be changed to the ones in the register. The new values will be denoted as $b_{.j}$ (j = 1,2,….,n) and the grand total will be b.

....	.	.	.	.	.
....	.	.	.	.	.
....	.	.	.	.	.
....	.	.	.	.	.
....	.	.	.	.	.
....	.	.	.	.	.
$b_{.1}$	$b_{.2}$	$b_{.3}$	.	$b_{.n}$	b

Let us change a_{ij} to $a_{ij}^{'}$ so that the column sums will remain and the change of the elements of the table will be proportionate. Thus, the new elements will be as follows: $a_{ij}^{'} = \frac{b_{.j}}{a_{.j}} \cdot a_{ij}$ $\forall$ $i \in \{1,2,\ldots\ldots,m\}, \forall j \in \{1,2,\ldots\ldots,n\}$, and they fulfil the following equation: $\sum_{i=1}^{m} a_{ij}^{'} = \sum_{i=1}^{m} \frac{b_{.j}}{a_{.j}} \cdot a_{ij} = \frac{b_{.j}}{a_{.j}} \sum_{i=1}^{m} a_{ij} = \frac{b_{.j}}{a_{.j}} \cdot a_{.j} = b_{.j}$. This means that the column sums are as expected and the inner elements are proportionate.

The row sums will be as follows:

$$b_{i.} = \sum_{j=1}^{n} a_{ij}^{'} = \sum_{j=1}^{n} \frac{b_{.j}}{a_{.j}} \cdot a_{ij} \ \forall \ i \in \{1,2,\ldots\ldots,m\}, \forall j \in \{1,2,\ldots\ldots,n\}.$$

The sum of the totals equals to the value of the grand total:

$$\sum_{i=1}^{m} b_{i.} = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{b_{.j}}{a_{.j}} \cdot a_{ij} = \sum_{j=1}^{n} \frac{b_{.j}}{a_{.j}} (\sum_{i=1}^{m} a_{ij}) = \sum_{j=1}^{n} \frac{b_{.j}}{a_{.j}} \cdot a_{.j} = \sum_{j=1}^{n} b_{.j} = b\,.$$

In the way described above a projection was carried out that keeps the inner relationships invariant combining the pieces of information of administrative and of statistical data sources.

5. An example of the merged data

In order to demonstrate the enriched data content following the merge of data coming from the two different sources the distribution of the new Hungarian citizens by educational attainment and age groups will be shown in the following. (www.ksh.hu).

Table 5: *Number of naturalized people by educational attainment and age group, 2002-2006*

	Age group							
Educational attainment	0–14	15–24	25–29	30–39	40–49	50–59	60–X	Together
	2002							
Uncompleted elementary	426	17	3	0	0	0	51	497
Completed elementary	40	122	54	101	61	65	172	615
Completed secondary	0	142	219	491	173	113	260	1398
Completed university or college	0	20	182	273	119	89	176	859
Together	466	301	458	865	353	267	659	3369
	2003							
Uncompleted elementary	578	10	5	5	5	10	59	672
Completed elementary	74	210	63	141	69	84	266	907
Completed secondary	0	278	352	678	285	167	399	2159
Completed university or college	0	31	331	532	222	120	287	1523
Together	652	529	751	1356	581	381	1011	5261
	2004							
Uncompleted elementary	619	33	5	16	5	5	49	732
Completed elementary	103	224	60	163	86	55	234	925
Completed secondary	0	283	367	713	340	153	365	2221
Completed university or college	0	38	350	554	188	136	288	1554
Together	722	578	782	1446	619	349	936	5432
	2005							
Uncompleted elementary	1088	30	10	10	0	10	99	1247
Completed elementary	139	385	109	256	128	148	445	1610
Completed secondary	0	453	617	1508	531	292	801	4202
Completed university or college	0	69	647	1005	373	193	524	2811
Together	1227	937	1383	2779	1032	643	1869	9870
	2006							
Uncompleted elementary	662	24	6	6	0	12	67	777
Completed elementary	98	214	79	177	61	73	354	1056
Completed secondary	0	250	395	877	284	186	601	2593
Completed university or college	0	37	360	592	189	100	397	1675
Together	760	525	840	1652	534	371	1419	6101

Table 6: *Distribution of naturalized people by educational attainment and by age group (%), 2002-2006*

	Age group							
Educational attainment	0–14	15–24	25–29	30–39	40–49	50–59	60–X	Together
	2002							
Uncompleted elementary	91.42	5.65	0.66	0.00	0.00	0.00	7.74	14.75
Completed elementary	8.58	40.53	11.79	11.68	17.28	24.34	26.10	18.25
Completed secondary	0.00	47.18	47.82	56.76	49.01	42.32	39.45	41.50
Completed university or college	0.00	6.64	39.74	31.56	33.71	33.33	26.71	25.50
Together	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	2003							
Uncompleted elementary	88.65	1.89	0.67	0.37	0.86	2.62	5.84	12.77
Completed elementary	11.35	39.70	8.39	10.40	11.88	22.05	26.31	17.24
Completed secondary	0.00	52.55	46.87	50.00	49.05	43.83	39.47	41.04
Completed university or college	0.00	5.86	44.07	39.23	38.21	31.50	28.39	28.95
Together	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	2004							
Uncompleted elementary	85.73	5.71	0.64	1.11	0.81	1.43	5.24	13.48
Completed elementary	14.27	38.75	7.67	11.27	13.89	15.76	25.00	17.03
Completed secondary	0.00	48.96	46.93	49.31	54.93	43.84	39.00	40.89
Completed university or college	0.00	6.57	44.76	38.31	30.37	38.97	30.77	28.61
Together	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	2005							
Uncompleted elementary	88.67	3.20	0.72	0.36	0.00	1.56	5.30	12.63
Completed elementary	11.33	41.09	7.88	9.21	12.40	23.02	23.81	16.31
Completed secondary	0.00	48.35	44.61	54.26	51.45	45.41	42.86	42.57
Completed university or college	0.00	7.36	46.78	36.16	36.14	30.02	28.04	28.48
Together	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	2006							
Uncompleted elementary	87.11	4.57	0.71	0.36	0.00	3.23	4.72	12.74
Completed elementary	12.89	40.76	9.40	10.71	11.42	19.68	24.95	17.31
Completed secondary	0.00	47.62	47.02	53.09	53.18	50.13	42.35	42.50
Completed university or college	0.00	7.05	42.86	35.84	35.39	26.95	27.98	27.45
Together	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Using the linkage of the two data sources it was revealed that the distribution of the naturalized people by the educational attainment is rather stable and those among the new Hungarian citizens having completed tertiary education is higher than the total population in Hungary: almost 30% vs. less than 10% (Census, 2001).

References

Stoyan G., Takó G. (1993) *Numerikus módszerek 1* (Numeric methods), Budapest.

Hunyadi L., Vita L. (2004) *Statisztika közgazdászoknak* (Statistics for economists), KSH, Budapest.

Kincses, Á. (2003) A magyar állampolgárság megszerzésének statisztikájához tartozó adatforrások leírása. (Description of the data sources for the statistics of acquisition of the Hungarian citizenship - manuscript).

Kincses Á., (2004) Népmozgalom; A magyar állampolgárság megszerzése 2002-2003. (Population movement; Acquisition of the Hungarian citizenship 2002-2003. HCSO) KSH, Budapest.

6

Register based statistics

Integration of administrative data in Poland

Janusz Dygaszewicz, Agnieszka Prochot
Central Statistical Office of Poland, Warsaw, al. Niepodległosci 208
e-mail: j.dygaszewicz@stat.gov.pl, a.prochot@stat.gov.pl

Abstract: The paper includes a brief view on Polish administrative resources and possibility of their integration. Some issues concerning legal basis was mentioned. Also historical development of register based statistical researches was included. Lately some effort concerning administrative sources has been made and about 300 registers was identified and elaborated. Part of them still under deep researches and probably will be suitable for future censuses. Actual activity concerning preparation of principles of the new methodology of integration of public registers has been described. Finally, some remarks have been noted as conclusions.

Keywords: public registers, methodology, integration, census, Poland, data sources

1. Introduction

The aim of this paper is to provide condensed information regarding integration of Polish public administrative data for statistical purposes. Now in Poland exist several hundred registers, but only 3 of them are pointed as the base administrative registers. There are: population identification register called PESEL, economy units and enterprises identification register called REGON and land identification and administration units register called TERYT connected with housing and dwellings registers. This 3 base registers deliver unified identification systems across other different registers and make it enable to integrate almost all administrative sources. It is important to note, that administrative data records can never fully replace data collection by sample surveys, but that these two methods complement each other. Use of administrative data sources is strictly connected with quality of statistics. In order to adjust the statistics to quality standards the work conducted within the Polish Official Statistics is focused on the more extensive use of administrative data sources. Meeting of users' needs, cost effectiveness and non-excessive burden on respondents are those areas where the wider use of the administrative data sources is a priority for Polish Official Statistics.

2. The legal basis for integrating administrative data

The Polish Statistical Law (issued on 29 June 1995 on official statistics) guarantees that the official statistical services have a right to use the administrative records for statistical purposes - the information on the use of administrative records in the defined scope, form and time is specified each time in the Programme of Statistical Surveys of Official Statistics which is annually determined by the Council of Ministers, in a way of the Regulation.

The legal basis for statistical data safety, including administrative data, is comprised by following legal regulations:

1) Law issued on 29 June 1995 on official statistics (Journal of Laws of 1995, No. 88, item 439, with later amendments)
2) Law issued on 29 August 1997 on personal data protection (unified text: Journal of Laws of 2002, No. 101, item 926, with later amendments)
3) Internal order No. 10 by the President of the Central Statistical Office issued on 21 June 2001 on implementing Rules and procedures of handling statistical data.

3. Historical development of register based statistics

The work on the use of administrative data sources for statistical purposes is a continuous process. The Central Statistical Office of Poland (CSO) started work on the use of the administrative data sources as soon as the first, significant administrative system - General Electronic System of Population Registration (polish acronym PESEL) has been implemented in Poland (The system has been established by state authorities on the base of the Law issued on 10 April 1974 on population register and identification cards (unified text: Journal of Laws of 2006, No. 139, item 993 with later amendments). The data from this system first time were used for the National Population and Housing Census in 1988.

In 90's, due to the significant development of techniques of data processing, the number of computerised information systems of public administration considerably increased, and within the CSO the intensive work has been started on identification of these data sources. The co-operation between statistics and the public administration bodies responsible for information systems has been intensified and the use of administrative data sources for statistical purposes has increased. The identified systems were evaluated as useful for the particular statistical survey or as the data sources for many statistical surveys.

This second group covers, among others, the following administrative systems:

Table 1: *Administrative systems covered by Phare'2003 Project*

The name of the system	*Name of body responsible for administrative system*
Population Register System: – General Electronic System of Population Registration - PESEL, – Registration files on the level Gminas' offices	Ministry of Interior Affairs and Administration Gminas' offices, e.g.Commune/Municipal offices
Tax System data sets	Ministry of Finance
Social Insurance System	Social Security Service
	Agricultural Social Insurance Fund
Central Register of Insured Persons	National Health Fund
System of Social Assistance	Ministry of Labour and Social Policy
System Concerning Registered Unemployment	
Geodesy System data sets	Office of Geodesy and Cartography
Real Estate Tax Register	Gminas' offices, e.g.Commune/Municipal offices
Integrated Administration and Control System IACS	Agency for Restructuring and Modernization of Agriculture

Intensive co-operation with administrative bodies was continued during the years 2004 – 2006, when the CSO within the Phare' 2003 project (Twinning Covenant between Poland and Sweden) conducted the work on upgrading of the quality of Polish statistics. The purpose of the one of project component (concerning extension of the use of administrative data sources for statistical purposes) was a development of methods to enable extended use of administrative information systems for statistical purposes and also establishment of general principles and the rules in this area. The work covered systems mentioned above.

The results were achieved in co-operation with representatives of the bodies responsible for administrative registers.

The experts from Statistics Sweden, Statistics Finland and Statistics Denmark gave their support during the whole time of the duration of the project.

4. Current activities

Until now, about 300 administrative systems have been identified.

The following organization of work on the use of the administrative data sources has been established within the Polish Official Statistics:

- One of the tasks of the Programming and Coordination of Statistical Surveys Division is to coordinate and conduct work on the use of administrative data sources for statistical purposes. Within the framework of Programming and Coordination of Statistical Surveys Division the Administrative Data Sources Section operates. This unit has been established in 2000 year and its work concerning, among others, running and maintaining the Metainformation System of Administrative Data Sources.
- The tasks of the divisions of the CSO include, among others, work on identification of existing and currently developed administrative information systems and preparation of propositions for their utilization as sources for statistics.
- All the Regional Statistical Offices are obliged to co-operate with bodies of public administration operating in the voivodship within the scope of creation and utilizations of administrative data for statistical purposes.
- Additionally, within the Regional Statistical Office in Warsaw, a specialist unit - the Centre of Administrative Data Sources operates.

These tasks have been established by internal orders by the President of the Central Statistical Office of Poland, introducing the Internal regulation of the Central Statistical Office as well as the statutes of the Regional Statistical Offices.

At present, a very intensive work on the use of administrative data sources is carried out in connection with the Census of Agriculture which will be conducted in 2010 and the National Population and Housing Census – in 2011.

5. Research work

At the first stage, the work on the use of administrative data sources concerns the identification and description of administrative systems as the potential sources for statistics. Information on systems are collected in the Metainformation System of

Administrative Data Sources - SMA, in one of its elements - a database for standardized description of the administrative sources, which is the base of knowledge on administrative data sources. It provides such information like:

- the name of the administration body running the system,
- legal basis which defines principles of establishment and running the system, principles of data dissemination and also regulates the information scope of the system,
- aim of the system
- information scope of the system
- structure of the system (task, functions) that is important for statistics

and also level of computing system's implementation, the quality of data sets including a frequency of their updating and presence of identifiers which make possible integration the particular system with others or with statistical surveys.

Such information make possible to conduct preliminary evaluation of usefulness of administrative data sources for statistics.

The SMA system has also a database of the description of administrative concepts and classifications.

In order to examine in which degree the administrative systems are consistent with the Polish Official Statistics system in the range of information scope, definition of concepts and classifications, the IT tools was created - PiK system. This system enables comparisons and examination of consistency between statistics and administrative data sources.

Such work has already been conducted for the purposes of censuses. Knowledge of public registers is the crucial condition of the register based censuses. PiK system was provided with necessary information and a result of conducted work was evaluation of degree of consistency the administrative concepts with statistical concepts as well as concepts from Recommendation for the 2010 Censuses of Population and Housing (prepared by United Nations Economic Commission for Europe and Eurostat).

6. Establish a new methodology

The next steps concern the establishment or modifications of statistical surveys methodology taking into account the use of variables from administrative sources: the reduction of the number of variables collected in traditional way and, in the same time, enlarging the scope of variables in the surveys through the use of administrative registers. The additional variables can be used to create new and to update existing sampling frames. After the providing statistics with administrative data the work on transformation of administrative data sets into statistical data sets can be conducted. In order to ensure the legal basis for providing the Polish Official Statistics with administrative data, the information on the use of administrative records in the defined scope, form and time for the purpose of particular survey is specified each time in the Programme of Statistical Surveys of Official Statistics as it was mentioned before. All necessary works concerning integration of administrative data have been conducted on the base of special regulations. In the census case a special law will be adopted.

6. Conclusions

At present, within the Polish Official Statistics, the administrative data sources are used as the direct sources of data for statistical surveys as well as for creation and updating sampling frames. The linking data sources have not yet been utilized in statistical surveys.

However, there are the plans for the creation of an integrated administrative data sources system. This system will be provided with data from the different administrative systems and with data from statistical surveys – at first - for the census purposes and it will enable to describe the unit on the level of microdata.

The IT tool for investigation of methodological consistency (definition of concepts and classifications that are use in the administrative systems and the statistical surveys) will be created through further development of the PiK system.

The methodology for quality assessment of administrative data sources will be elaborated – the quality reports for each variable will be prepared.

Good cooperation between statistical services and other administrative register-keepers are crucial for integration of administrative data for statistical purposes.

Challenges of the register based census in Austria with special focus on effort and impact of including small register bases

Eva-Maria Reiner
Statistik Austria, Guglgasse 13, 1110 Wien
e-mail: eva-maria.reiner@statistik.gv.at

Abstract: For the register based census 2010 in Austria a register based census test was performed with reference date of 31st October 2006. An overview about appearing problems and applied solutions is given. A first issue more deeply discussed is the creation and use of an anonymous personal identification number. Secondly, a special focus is put on analysing the influence and gain of some small register bases. These data sources cost a lot of time and effort to be delivered to Statistics Austria, so knowledge about the necessity of them is essential for future register based censuses.

Keywords: register based census, data editing

1. Introduction

In 2010 the first register based census in Austria will take place. In preparation for this census a register based census test took place in October 2006. Both delivering and receiving the data from the various sources, as well as editing and analysing the received data was quite time consuming, so a first result was published in April 2008. In the following an overview of some challenges we had to face is given. This is mainly done by analysing the influence of some small social insurance register data, which were ten different data sets. Before that, a general part about the difficulties concerning data linkage is given.

2. Personal identification number

The first challenge of the register based census test was to find a unique identification number for each person. In Austria there are some different numbers used, which are nearly person-unique and which exist for almost every person. The main ones are the social security number and the population register number. Unfortunately due to the assigning procedure it is possible that a person could have more than one number in both cases. Also the protection of data privacy has to be taken into account. For this purpose a special branch specific personal identification number (bPIN OS), introduced for e-government procedures in Austria, was used.

2.1. Creation of the branch specific personal identification number

The bPIN is an anonymous personal identification number, which has the advantage that two different data owners have different bPINs for a particular person. For linkage, the

Registration Authority, which is part of the Austrian Data Protection Commission, has to be involved.

The bPIN is created from the identification number of the central population register in a rather sophisticated and non-invertible way: First the persons are linked with the identification number of the population register, using first name, surname and date of birth provided by a data owner. From this identification number, a “root figure” (Stammzahl) is created for every person by the Registration Authority. Then, for every data owner, this root figure is encoded to an area-specific person code. Hence two different data owners do not have the same bPIN for the same person. Moreover, if bPINs are transmitted, they are concatenated with a time stamp and encoded with a public key procedure.

For the record linkage, the Registration Authority has to equip data with both the bPIN of the Statistics Austria (bPIN OS) and of the data owner, each encoded with public key of the appropriate area. Only the Statistics Austria is able to decode the incoming bPIN OS from different sources, and hence data can be linked. This way the data security standards have been met.

2.2. Further record linkage

Almost every data source contained data sets without valid bPIN OS. There are two possible explanations for that. Either the person was not part of the Austrian population on the reference day, but part in another register due to various reasons. Or the data quality was not good enough to assign a bPIN OS to this person, e.g. because of a wrongly written name.

Here a record linkage procedure was implemented, using addresses, date of birth, sex and nationality as linking parameter. Before linking the data, some editing and standardising procedures, mainly for addresses and nationalities took place. After that there were several stages of data linkage, using different subgroups of linking attributes, but always the date of birth as “hard fact”. Only if there was a unique match, a bPIN OS was added to the data set with the missing identification number. So about 70.000 single population register data sets could be linked with other data. This are about half of the potential non-active register entries discussed further in section 3.3. Residence Analysis.

3. Editing and implementing of small register data

For the register based census test 2006 every administrative data which could be possibly interesting was requested from various data owners. After collecting and merging all data sources one of the questions that arose was how much information we gained from some small registers.

In the following section the influence of data delivered by special social insurance institutions is analysed.

3.1. Background

In Austria there is a compulsory insurance for employed people and a free insurance for certain groups of the population, e.g. non-working relatives of insured persons.

All major social insurance institutions and funds are members of the Main Association of Austrian Social Security Institutions (MA). For all of these members one collective data set was delivered to Statistics Austria including all data for the reference date. This delivery contained demographic and geographic information like sex, date of birth, citizenship, address, status of employment and means of subsistence.

Some special groups of public employed and self employed members of certain chambers (e.g. lawyers, medical doctors) are not members of MA. So they had to deliver data on their own. As some of them did not even have a database, and the legal situation was not clear in the beginning, to name just a few of the problems, this process turned out to be quite complicated. A huge effort was needed for the whole process of data requesting and receiving. Firstly, a lot of time was invested in communication. According to our records, in the time span from June 2006 until December 2007, 136 emails and 40 letters have been sent and 76 phone calls have been made, which were just the most important ones. Secondly, some of the data was delivered after the dead line or not at all, respectively. Thirdly, data was delivered in many different file formats and coding varied from one register to the other. Only data editing to create usable data sets for further processing took a man-month.

3.2. Benefit of data delivery

For the register based census test 64.856 data sets on person level with a valid person key were delivered by these special social insurance institutions. Further 26.523 data sets contained information about co-insured persons with a valid key. This amount is approximately 1% of the total population of Austria.

To determine the status of employment 66.144 data sets were delivered, as one person could have two places of work (e.g. two different pharmacies).

To give an idea about the sizes, more than 8.000.000 data sets were delivered by the MA, for example.

Both of this data was used as redundant information to confirm persons' main residence in Austria on the reference data as well as for various demographic attributes. Furthermore it was used as primary information to determine means of subsistence and status of employment.

To analyse the effect of the data sources of special insurance institutions, data editing and calculating processes were simulated without these sources and those results were compared with the original results of the register based census test.

3.3. Residence analysis

The first aim of the register based census is, as of any census, to determine the population of the country, which is defined as the persons with main residence in Austria on the reference day. As base register the central population register was used. This was installed in 2001 during the process of the last traditional census.

After linking all other data sources to the population register using bPIN OS, and accomplishing the record linkage procedures for data sets with missing bPIN OS, the persons of the register were analysed.

Data sets with no connection to other registers could be non-active register entries and therefore not to be counted at the census. So finding a sign of life from other registers was a main challenge of the register based census.

Comparing the results with or without data of those special social insurances, there were only 40 persons, or less than 0.001 % of the total population, who were only confirmed by them. This is due to the fact that as another register for comparison the tax register was used as well as the family allowance register. The small amount of difference could be persons who where not linked with other registers because of missing bPIN OS.
So for this aspect of register based census there is no significant surplus value.

Table 1: *Changing from not currently active to currently active by Länder and gender*

	Active without extra data	Percentage of total	Active due to extra data	Percentage of total	Active with extra data
		Total currently active population			
Total	4025101	99.87	5289	0.13	4030390
Burgenland	138371	99.97	44	0.03	138415
Carinthia	263589	99.93	184	0.07	263773
Lower Austria	781044	99.94	456	0.06	781500
Upper Austria	689209	99.63	2531	0.37	691740
Salzburg	260049	99.91	233	0.09	260282
Styria	587501	99.93	414	0.07	587915
Tyrol	341323	99.90	330	0.10	341653
Vorarlberg	169239	99.93	115	0.07	169354
Vienna	794776	99.88	982	0.12	795758
		Male			
Total	2177196	99.81	4070	0.19	2181266
Burgenland	76248	99.95	36	0.05	76284
Carinthia	143511	99.89	163	0.11	143674
Lower Austria	425156	99.91	374	0.09	425530
Upper Austria	377407	99.54	1727	0.46	379134
Salzburg	138119	99.86	199	0.14	138318
Styria	320939	99.89	358	0.11	321297
Tyrol	185984	99.84	301	0.16	186285
Vorarlberg	92502	99.89	104	0.11	92606
Vienna	417330	99.81	808	0.19	418138
		Female			
	1847905	99.93	1219	0.07	1849124
Burgenland	62123	99.99	8	0.01	62131
Carinthia	120078	99.98	21	0.02	120099
Lower Austria	355888	99.98	82	0.02	355970
Upper Austria	311802	99.74	804	0.26	312606
Salzburg	121930	99.97	34	0.03	121964
Styria	266562	99.98	56	0.02	266618
Tyrol	155339	99.98	29	0.02	155368
Vorarlberg	76737	99.99	11	0.01	76748
Vienna	377446	99.95	174	0.05	377620

3.4. Current activity status

A current activity status was determined for all persons living in Austria on the reference date. Additionally information of persons working in Austria, without living here, was gathered mainly for the census of local units. In the following section only persons with main residence in Austria are compared. There is a hierarchic concept for the current activity status, as defined in the census recommendations. On the top level the population is divided into a current active part (i.e. the labour force) and a part which is not currently active (i.e. persons not in the labour force). Comparing the data with and without the special social insurance data processed for determine whether a person is currently active or not, leads to the result in Table 1. As one can see, around 5.300 persons or 0.13% of the population changed from non-active to active. As data will be analysed on level NUTS II (in Austria called Länder), and for male and female separately, data is displayed in these groups.

Table 2: *Changes from "others" to "pension receivers" by Länder and Gender*

	Pension receivers without extra data	Percentage of total	Pension receivers due to extra data	Percentage of total	Pension receivers with extra data
Total currently active population					
Total	1743440	99.93	1286	0.07	1744726
Burgenland	67185	99.99	4	0.01	67189
Carinthia	125307	99.98	22	0.02	125329
Lower Austria	351510	99.99	38	0.01	351548
Upper Austria	288113	99.64	1046	0.36	289159
Salzburg	101244	99.98	25	0.02	101269
Styria	266201	99.98	40	0.02	266241
Tyrol	126687	99.98	22	0.02	126709
Vorarlberg	64180	99.98	13	0.02	64193
Vienna	353013	99.98	76	0.02	353089
Male					
Total	778284	99.97	196	0.03	778480
Burgenland	30485	99.99	4	0.01	30489
Carinthia	58938	99.97	17	0.03	58955
Lower Austria	160160	99.99	23	0.01	160183
Upper Austria	130468	99.96	57	0.04	130525
Salzburg	45677	99.98	8	0.02	45685
Styria	123510	99.98	21	0.02	123531
Tyrol	59931	99.98	13	0.02	59944
Vorarlberg	29022	99.97	9	0.03	29031
Vienna	140093	99.97	44	0.03	140137
Female					
	965156	99.89	1074	0.11	966230
Burgenland	36700	100.00	0	0.00	36700
Carinthia	66369	99.99	5	0.01	66374
Lower Austria	191350	99.99	14	0.01	191364
Upper Austria	157645	99.38	988	0.62	158633
Salzburg	55567	99.97	16	0.03	55583
Styria	142691	99.99	14	0.01	142705
Tyrol	66756	99.99	5	0.01	66761
Vorarlberg	35158	99.99	4	0.01	35162
Vienna	212920	99.99	28	0.01	212948

The biggest influence of the extra data could be observed in Upper Austria, where the active male population rose by 0.45%, but even here this was less than half a percent of the active population.

Not included in Table 1 are persons who were classified as contributing family workers based on an estimation model.

Looking only at the currently active population, 429 persons changed from unemployed to employed persons.

Within the not currently active population, about 1.300 persons were classified as pension receivers instead of “others”, which is approximately 0.1% of pension recipients. All changes are in table 2.

Also in this status there are the biggest changes in Upper Austria. Especially for the female population there are almost 1.000 persons more receiving pensions instead of being classified as “others”. This is more than 0.5 %, but still less than 1% of this group.

Analysing the effect on status of employment of the population in employment, one can see only small changes, as there are around 270 more self-employed persons, and around 800 more classified as employees or self-employed persons, instead of not classifiable. There were no significant differences among regions and sexes, so the results for all employed persons can be seen in Table 3.

Table 3: *Changes in Status of Employment*

	Status of Employment without extra data				
	Employees	Employers and Own-account workers	Contributing family workers	Persons not classifiable by Status	Total
Status of Employment with extra data					
Employees	3302004	19	-	697	3302720
Percentage of Employees Total	99.98	0.00	-	0.02	100.00
Self-employed	273	418506	-	103	418882
Percentage of Self-employed Total	0.07	99.91	-	0.02	100.00
Contr. fam. workers	-	-	13363	-	13363
Percentage of family workers Total	-	-	100.00	-	100.00
not classifiable	-	-	-	290135	290135
Percentage of not classifiable Total	-	-	-	100.00	100.00

On micro level such clarifying can be of high interest, for the whole population it is within the range of expected inaccuracy because of data quality and differences in definition.

3.5. Effect of not delivered data on activity status

In a further step it was attempted to analyse the potential influence of data not implemented, because it was too late or not at all delivered.

Parallel to the register based census test a sample survey was performed like a traditional census for 10.000 households all over Austria. These survey data were assigned with a bPIN OS too, so they could be linked with the register based census on personal level. An analysis of the differences concerning the status of employment was performed. A majority of the persons who answered to be currently employed where covered by register data, also some of those who were not delivered by the special social insurance institutions.

As an example for not delivered data and the potential influence persons potentially belonging to the chamber of lawyers were analysed. These professions are included in NACE categories together with many others, so the descriptions filled in by the interviewers are used for further analysis, combined with status of employment “self-employed”. The currently active population according to register and survey data was compared with the not currently active population according to register data, but currently active according to survey data.

Of the group of persons who where self-employed and filled in as description of work “lawyers” only four persons were classified as currently active, whereas twelve, so three times as many, were classified as non active according to the administrative data sources used. By contrast, all pharmacists were classified as active by processed register data.

So not using all the information leads to a systematic error. As some of the professions are under represented or not at all included in all available sources (e.g. artists or contributing family workers) there is the need for some estimation anyway.

But as the delivered data of small special social insurance institutions proved to be of good quality, using this data could be more reliable and should be included in future register based censuses.

3.6. Place of work - Local unit

Companies have to report the place of work for all employees starting from 2007 together with the annual pay slip. The MA data only includes address information on enterprise level, where as most of the special social security institutions delivered information on local unit level.

Using the 2007 data, a further analysis on the place of work (influence on commuting statistics and census of local units) will follow, and then the influence of the information of these data source will be analysed as well.

3.7. Family status, marital status and other demographic attributes

Unfortunately, most information on co-insured persons was delivered without a personal key by special social insurance institutions. Most of the 26.300 persons with a valid key were included either in the child allowance register or in the data about the co-insured persons, which was delivered via social insurance institutions belonging to MA. So taking into account the effort of data editing and the small gain (4.600 persons), it was not used for the register based census test for determine the family status.

For all demographic attributes information was gathered by many different administrative sources according to the principle of redundancy. So for example instead of missing values coming from one register, information from another register can be used. Using different sources for one attribute, leads to the necessity of defining rules, which have to take in account differences in data quality as well as a perspective of content. Along with analysis of various data source a set of rules was developed and extended. For the attributes sex and age the population register itself already offered good information for 99% of the population with main residence in Austria. Using the other sources for data controlling and replacing of missing data, a valid sex or age could be determined for all persons living in Austria on the reference date. The country of citizenship was also well represented in four major data sources (population register, MA, Unemployment Register, child allowance register), although there appeared a time-lag problem concerning naturalisation, where the country of citizenship was not updated.

For the legal marital status the situation was different, as it is not covered fully in any register. As it was just lately implemented in the population register, there is valid information for only 3% of the population. In a stepwise process valid marital status were defined, using other register sources as well as information about the family status. For 4.000 persons, or 0.05% of the population, a valid marital status was only delivered by the special social insurance institutes analysed.

Taking into account the high missing rates for this attribute any available administrative source should be used also in future. Those data included valid marital status for about 46.000 persons, these are 0.5% of the population. How up to date this information is, has not been analysed so far, as some of these data were implemented only very recently.

Another attribute delivered by some special social insurance institutions was the number of children. As this attribute was not delivered by the MV or any other source this was too little information to be included in the register based census test.

4. Conclusion

Using the bPIN as personal identification number worked for most of the population, although there will be a need for further record linkage procedures, which proved to be of good quality.

In summary there was no significant benefit from those ten data sources concerning residence analysis. So for just determining the population, even at a small scale, it is not necessary to use these sources.

For attributes like the marital or the current activity status information from these special social insurance institutions proved to be of good quality. Some information, e.g. about the place of work or de facto marital status was very detailed, so it would be recommendable to use this data for future register based censuses.

A form of standardisation for the data supply and preparation is recommended, as this would accelerate the process of data editing, which could be, at least partly, automated.

References

Fiedler R., Schodl P. (2008) Data imputation and estimation for the Austrian register-based census, *UN/ECE Work Session on Data Editing,* http://www.unece.org/stats/documents/2008.04.sde.htm.

Lenk M. (2007) Practical Guidelines Data Integration- The Principle of Redundancy - Austrian Register Based Census, *CENEX Paper WP2*, http://cenex-isad.istat.it.

Wallgren A., Wallgren B. (2007) *Register-based Statistics - Administrative Data for Statistical Purposes*, John Wiley & Sons, Ltd.

Conferences of European Statisticians (2006) *Recommendations for the 2010 Censuses of Population and Housing*, United Nations, http://www.unece.org.

Main Association of Austrian Social Security Institutions (2007) *Well insured - Social Security in Austria*, http://www.sozialversicherung.at.

Registerzählungsgesetz (2006), *BGBl. I Nr. 33/2006,* http://www.ris2.bka.gv.at.

Linking register of construction works with census in Estonia

Kai Kaarna
Statistics Estonia, 15 Endla Str, 15174 Tallinn, Estonia,
e-mail: kai.kaarna@stat.ee

Abstract: In Estonia we have some experiences with linking buildings and dwellings from Census and register. There was no unique identifier for buildings in both databases and available address characteristics were used.

Keywords: linking, Census of Population and Housing, register, building

1. Introduction

The aim of this paper is to provide our experiences of linking.

In autumn 2005 Statistics Estonia (SE) started to plan the next Population and Housing Census, which will be conducted in 2011.

For independent evaluation of the quality and usage of the Register of Construction Works data in statistics, Statistics Estonia together with the Ministry of Economic Affairs and Communications (MEAC) carried out the project "Preparation for the 2011 Census: Quality evaluation of the Register of Construction Works".

The activities concerning that project were connected with the linking of two different databases, comparing the data in them and mapping activities which are necessary to convert to register-based capitation.

There was possibility to link buildings and dwellings from the last Census and from Register of Construction Works (RCW) by addresses. Addresses were divided into following components: state, county, town or rural municipality, settlement, street or farm name, number of building and number of dwelling. For record linking there were no unique identifiers and we had had to link records by the available variables. But that could be affected by errors.

2. Design of sample

The basis for linking was a random sample of 4,700 buildings from the 2000 Population and Housing Census database. The layers had been formed considering the location of the building (county and settlement type). All the buildings were divided into 47 layers, from each 100 buildings were randomly selected to be used in the survey.

Of the buildings in the random sampling, the sample of dwellings was taken as follows: from the buildings with up to 4 dwellings all dwellings were taken into the sample and from the rest of buildings every tenth dwelling. The sample of dwellings included 6,193 dwellings.

In Estonian villages there have been historically used both farm names and street names. In the beginning of 20 century there were only farm names and every farm was identifiable. Most of Estonian villages are scattered and there are no street names and numbers of the buildings. In Soviet time farms were reorganized and historical farm names were not used and have been forgotten and don't correspond to the present houses. Because of historical changes, there are some buildings addressed by farm names or lately by street names with number of buildings but some are in the register only by name of village.

3. Linking process

In the linking process our attempt was to find for all items of the sample a "partner" from RCW if possible.

There were three stages. At first an attempt was made to locate an automatic response to each building from the sample amongst the RCW, thereafter the buildings not linked were checked manually, one by one, and if possible were linked. At the same time the reasons for unlinking were studied and new regulations were made for automatic linking. Then a new automatic linking attempt was made.

All the dwellings of the linked buildings were used for linking dwellings.

In Estonian villages there have been historically used both farm names and street names. In the beginning of 20 century there were only farm names and every farm was identifiable. Most of Estonian villages are scattered and there are no street names and numbers of the buildings. In Soviet time farms were reorganized and historical farm names were not used and have been forgotten and don't correspond to the present houses. Because of historical changes, there are some buildings by farm names or lately by street names with or without number of buildings but some are in the register only by name of village.

Rules in the first stage:

1. In cities buildings from different databases were connected by the following characteristics: county, settlement, street (coded in the Census) and building's number. 3,302 pairs were generated, of which almost half were RCW outbuildings.
2. Two rules were used to link buildings in rural areas (rest of the settlement types) automatically with RCW buildings by addresses. Equivalents were tried to be found by RCW addresses as follows:
   a. If the building was situated in a settlement as town, small town, village, city without municipal status, the following characteristics were used for linking the buildings: county, rural municipality, settlement, street or if it is missing street/farm, number of building. Altogether 563 different buildings were linked on the basis of this rule.
   b. In addition the following rule was used for linking villages and buildings: county, rural municipality, settlement, street or if it is missing street/farm (in RWC the characteristic is in text format). In addition 158 more buildings were linked.

4. 1:1 or not

In some cases several "partners" were found to an address from Census. There were many records belonging to the register, which matched by address with same record from the census. Obviously sometimes among these buildings were cotes, sheds etc. We

checked if there was living space in the partner and then matched the Census-address with that record. We decided that there is unique match if in RCW can be found only one linkable building with living space.
49% of the buildings (2,304 buildings) were linked uniquely using the first programme.

5. Manual linking

While linking the buildings manually we were discovered by regions many different types of errors that caused unlinking. The most commonly the reason was in writing stile of the texts in addresses. Hence abbreviations, but also quotation marks, first name expansion, special letters, dash and space differences in street/farm names were the reasons for unlinking. These reasons were taken into consideration for generating new rules for linking buildings by addresses.

6. Uniform address-standard

In Estonia standardized address data system (ADS) has been developed now but it was not used in these databases. Since up to the summer 2007 the uniform address-standard has been absent in the country. ADS could solve these linking problems, but only if it will be used in both (register and Census data).

7. New automatic linking

It was decided to apply some new rules and to carry out the third linking. The rules applied well and as a result of the third linking we succeeded to link, within the whole sampling, 68% (3,188) of buildings.

Table 1: *Building linking by stages*

Result	I automatic		manual		II automatic		Sample, N
	N	%	N	%	N	%	
Equivalent is a building with dwelling(s)	2,184	69	444	14	2,977	95	3,146
Equivalent is several buildings with dwelling(s)	104	56	21	11	173	93	187
Equivalent is buildings with no dwelling(s)	16	33	17	35	38	78	49
Not linked	0	0	540	41	0	0	1,318
Sample	2,304	49	1,022	22	3,188	68	4,700

Using the expansion factors (weights) we expanded the results for the whole Census database.

Table 2: *Building linking by counties (on the assumption of buildings), weighted*

County	Building linking by counties				Buildings in sample	Total
	Equivalent is a building with dwelling(s) %	Equivalent is several buildings with dwelling(s) %	Equivalent is buildings with no dwelling(s) %	Not linked %	N	N
Harju	68.5	7.0	1.7	22.8	600	39,355
Hiiu	71.6	2.7	1.5	24.2	200	3,328
Ida-Viru	60.2	3.4	1.3	35.0	300	13,289
Jõgeva	27.5	0.5	0.2	71.7	300	10,210
Järva	50.0	1.5	-	48.5	300	8,910
Lääne	60.4	2.4	0.4	36.9	300	7,309
Lääne-Viru	42.0	1.9	0.9	55.3	300	14,863
Põlva	15.9	0.6	0.3	83.2	300	9,700
Pärnu	61.3	3.5	1.5	33.7	300	17,788
Rapla	70.0	2.8	1.9	25.4	300	9,544
Saare	66.9	4.9	1.6	26.5	300	10,312
Tartu	42.3	4.1	0.4	53.2	300	21,096
Valga	60.7	3.9	0.2	35.1	300	8,637
Viljandi	82.7	4.7	2.8	9.8	300	13,207
Võru	21.2	2.2	0.3	76.3	300	10,146
Total	54.9	3.8	1.1	40.2	4,700	197,694

Table 3: *Buildings linking by counties (on the assumption of dwellings), weighted*

	Building linking by counties				Sample (dwellings)	Total (dwellings)
County	Equivalent is a building with dwelling(s) %	Equivalent is several buildings with dwelling(s) %	Equivalent is buildings with no dwelling(s) %	Not linked %	N	N
Harju	78.9	10.0	0.8	10.3	967	224,763
Hiiu	74.7	2.7	1.3	21.4	224	5,003
Ida-Viru	85.0	4.2	2.0	8.8	539	85,859
Jõgeva	42.0	0.6	0.3	57.1	334	17,951
Järva	63.6	2.1	-	34.3	374	18,558
Lääne	69.9	2.2	0.4	27.5	365	15,145
Lääne-Viru	57.5	3.4	0.7	38.4	384	33,256
Põlva	31.8	0.9	0.5	66.8	384	15,656
Pärnu	72.7	4.4	1.0	21.9	383	40,127
Rapla	72.0	4.7	2.5	20.7	365	17,551
Saare	72.2	5.8	1.2	20.8	338	16,453
Tartu	63.0	7.5	0.4	29.2	408	64,660
Valga	70.4	3.9	0.2	25.5	366	17,440
Viljandi	83.6	4.5	2.2	9.7	377	25,954
Võru	39.8	6.8	0.4	53.0	385	18,891
Total	71.9	6.4	1.0	20.7	6193	617,267

Weighted results are not so good (Table 2): we can link totally 55% of buildings and in some counties even less than 25%. In one layer (villages in county of Võru) there were no linked buildings at all. This area is a periphery with especially little villages.

We were able to link at least buildings for 72% of dwellings totally (Table 3). In these cases only one part of the address – the number of dwelling – did not match in both data-bases. In county of Ida-Viru we linked buildings for 85% of dwellings and in county of Viljandi for 84% but in county of Põlva only for 32% and in county of Võru for 40%.

The equivalents could be found for 43-44% of dwellings (in some cases in register there were no parts as dwellings but there have been made changes in the register) and for 28% of dwellings at least buildings could be linked although the dwellings could not be linked (Table 4).

Table 4: *Buildings and dwellings linking, weighted*

Buildings	Dwellings	Dwellings, %
Equivalent is a building with dwelling(s)	Linked with RCW part of building	43.4
	Linked with RCW building (building is not divided in parts)	0.2
	Dwelling is not linked, but RCW is divided in parts	28.3
Equivalent is several buildings with dwelling(s)	Also the building is unlinked	6.4
Equivalent is buildings with no dwelling(s)	Also the building is unlinked	1.0
Not linked	Also the building is unlinked	20.7
Total (N)		617,267
Sample (N)		6,193

References

TF2005 project "Preparations for the 2011 Population Census: evaluation of the quality of the Register of Construction Works" in Estonia, *Final Report* (2008).

European Commission

Insights on Data Integration Methodologies — ESSnet-ISAD workshop, Vienna, 29-30 May 2008

Luxembourg: Office for Official Publications of the European Communities

2009 — 236 pp. — 21 x 29.7 cm

ISBN 978-92-79-12306-1
ISSN 1977-0375

KS-RA-09-005-EN-N

Publications Office
Publications.europa.eu