ESSnet on DATA INTEGRATION

Report on WP1
State of the art on statistical methodologies for data integration

ISTAT, CBS, GUS, INE, SSB, SFSO, EUROSTAT

ESSnet on Data Integration

Contents

Preface

Miguel Guigó

Instituto Nacional de Estadística – INE, Spain

The following document is the result of the tasks carried out under the name Work Package 1 (WP1), a part of the ESSnet on Data Integration project. The goal of WP1 is to provide a review on the state of the art concerning data integration procedures, and serve as a guide for producers of official statistics within the ESS in order to get an adequate theoretical background on the subject.

The report has been designed as a tool to correctly identify and accurately define a problem of integration of multiple sources of data; then, compare the methods available, their features, and their ability -or not- to solve the current problem; and finally, choose the alternative that best fits the characteristics of the information to be combined, being aware of the issues that can arise.

There are three key ideas to keep in mind relating to the procedures shown in this document: first, their scope is basically the information enhancement that is achievable at the statistical unit level; second, they are based on the theory of probability, which allows to perform a complete set of quality tests and measures; and last, they are intended to implement automated processes, capable of dealing with large datasets often handled by NSIs and other government institutions. Notwithstanding the foregoing, some issues discussed in other scientific domains - such as computing -, or relating to manual treatment of data or estimates at the aggregate level have been also taken into account.

Three methodological areas are presented and developed: i) record linkage, ii) statistical matching, and iii) micro-integration processing. While the two first are regarded as strictly data integration techniques, in the sense that they are methods to gather information from two or more different data sources for the same statistical production process, the third aims to improve the quality of the data obtained from combined sources by correcting errors and ensuring the reliability of the outcomes.

The text has been conceived as an update and a completion of the corresponding report that was issued by the ESSnet Statistical Methodology Project on Integration of Survey and Administrative Data (ESSnet - ISAD); nevertheless, it has been prepared to be read for itself, though it presupposes a previous background on Statistics and reading the ESSnet - ISAD document is recommended.

The Report of WP1: State of the art on statistical methodologies for data integration consists of 6 sections. Chapter 1 provides an overview of the record linkage procedure for both the standard (Fellegi-Sunter) and alternative (Bayesian) approaches, and the most prominent problems concerning the former along with the most suitable solutions and recent developments -mainly, 2007 onwards-. Chapter 2 gives an overall picture of several issues that have been barely faced by the literature on statistical matching, such as data drawn from complex sample survey designs, uncertainty on the density distributions associated to this method, or some non-parametric procedures. Chapter 3 and 4 analyze the impact of data integration techniques on other methodological areas such as, respectively, ecological inference and statistical disclosure control. Chapter 5 provides a formal definition of micro integration processing together with both a description of the concepts involved in micro-integration and the differences with related fields -macro-integration or editing and imputation-, also including a framework of process steps and methods involved in micro-integration. Last, a comprehensive bibliography on record linkage and statistical matching, beyond the references already mentioned throughout those five chapters, has been included at the end of the document.

It is important to stress that, in order to ensure the internal coherency and clarity of the text, a unique term for each procedure has been used along the report, regardless of the fact that some of them can be found under different names in the scientific literature; e.g. record linkage is sometimes also known as ”object identification” ”entity resolution” or ”merge-purge”, whilst statistical matching could be also known as ”data fusion” or ”synthetical matching”, etcetera.

Finally, the following report should be regarded as just the first step in order to make data integration techniques applicable in real cases: the rest of the deliverables and outputs of the ESSnet DI project such as i) the document on methodological developments on problems which are still open (WP2), ii) the report on case studies (WP4) and iii) the software applications provided for record linkage (RELAIS) and statistical matching (StatMatch) (WP3) are the additional tools that can ensure a successful performance of the integration process.

Special thanks are due to Marcin Szymkowiak for the efforts in transforming all the files in this document in LaTeX.

Chapter 1

Literature review update on record linkage

Summary: The goal of record linkage procedures is to identify pairs of records which belong to the same entity; by record is meant a set of data which has been gathered on a unit and arranged in the form of fields or variables; records, in their turn, are gathered into databases. We introduce a formal procedure to find records belonging to the same unit being from either different or the same source – that is, database. These procedures are based on probabilistic instead of deterministic criteria and rely on the equivalence of values when comparing those from two different records on a field-by-field basis; and then, on the probability of agreement between values given the true – and unknown – status of the pair of records – that is, they actually do belong to the same entity or they actually do not. Both standard and alternative approaches for probabilistic record linkage are discussed: the former is widely known as the Fellegi-Sunter theory and the latter is represented by the Bayesian approach. Some other related issues common to both alternatives, such as reducing the number of comparisons and dealing with risk of data disclosure are also illustrated in subsequent sections.

Keywords: record linkage, microdata, Fellegi-Sunter, E-M algorithm, Bayesian models, efficient blocking.

1.1 Introduction

Nicoletta Cibella[a], Miguel Guigó[b], Mauro Scanu[a], Tiziana Tuoto[a]

[a] *Istituto nazionale di statistica – Istat, Italy*

[b] *Instituto Nacional de Estadística – INE, Spain*

1.1.1 The concept and aim of record linkage

Probabilistic record linkage consists of a sequence of procedures addressed to determine whether a pair of records from two different sets A and B belong to the same entity or not. For what concerns the statistical production process, this entity is typically a person, a household, a business, an establishment or, broadly speaking, any kind of statistical unit that is present in a set of microdata.

The immediate target of these procedures is to enrich the information already held on such units in a database that is maintained for statistical purposes, by means of adding new data on the same individuals from other sources; and the reason for its use is the absence of a unique and error-free identifier which would permit to merge this information in an automated, massive and low-cost way, while dodging the risk of committing mistakes.

Since the so-called records are made up of a set of data collected on an individual, arranged in different fields (let us regard them as ”columns”) holding different possible values and sorted togheter in the same ”row” for the same individual, it is feasible to use those data in order to link records, by means of comparing fields that are common to A and B – provided that the information stored in these fields has been first properly treated in both A and B to make comparisons possible. This set of common fields will be known from now on as *key variables.*

1.1.2 Characterizing the probabilistic approach

The whole record linkage process can be regarded as a workflow, some of whose stages – see Cibella, Tuoto et al. (2009) for a complete view – define, depending on the specific solutions used, whether a probabilistic approach has been adopted or not. So, (1) the choice of the comparison method, a comparison function, and (2) the decision rule which states if a pair should be considered as a match given a function value, together with (3) the evaluation of results, can be regarded as those basic stages.

The first, though not unique, feature of probabilistic record linkage is that, in order to determine whether a pair of records (a, b), which have been brought together, belongs to the same unit or not, all the key variables are simoultaneously compared. This is due to the fact that data stored in both records is assumed to contain errors that could either result in non-coincident values for matching records or *vice versa*, for each variable considered separately (for some examples, see Winkler, 2006a). Furthermore, every record in a set A is compared to each record on set B. This ignores, in principle, the alternative of using a hierarchical algorithm that first examines one piece of information and then discards definitely a subset of candidate records from B to be linked to $a \in A$. As a consequence, a probabilistic record linkage procedure has to handle one set Ω of elements which are pair of records, $\{r\} = \{(a, b) : a \in A, b \in B\}$, with Ω typically made up of the Cartesian product AxB.

Some procedures, though, have been developed in order to reduce that amount of comparisons, see Baxter, Christen and Churches (2003) and Michelson and Knoblock (2006) or Goiser and Christen (2006) for a critical approach; they will be more widely developed in the following sections.

A second feature is the way in which similarity between records is assessed; that is, in other words, the value associated to each pair $r = (a, b)$. Given an element (record) with known values for K variables, its closeness to another element could be measured, for example, in terms of a distance δ, where δ is a function $\delta(a, b)$, be it Euclidean or whatever. Probabilistic record linkage first associates to each pair a comparison value $\gamma^{ab} = \gamma(a, b)$ (we will denote it simply by γ), which is a vector of K components $\gamma = (\gamma_1, \ldots, \gamma_K)$, one for each key variable to be compared. The value of γ_k, a k-th component of γ, would be $\gamma_k = 1$ when information on both records exists and agrees on the k-th field, and $\gamma_k = 0$ otherwise (see ESSnet on ISAD, 2008, section 1). Another possible set of values could be selected for γ_k considering the outcomes "information on both records exists and agrees", "information on both records exists and disagrees" or "information is lost in any or in both records".

The key point to be stressed, though, concerning the way that probabilistic record linkage measures closeness between records, does not lie on the space Γ of comparisons that contains the possible values assigned to γ. Once $\gamma(a, b)$ is obtained, a value related to each γ is then calculated, say a function $\phi(\gamma)$, expressed in terms of probabilities. That means in its turn to assign such values to each pair (a, b) in order to assess the similarity between a and b. Thus, by means of using that measurement, it is feasible to implement a variety of tools provided by Theory of Probability and Statistical Inference, be them parameter estimation, hypothesis-testing, classification and discriminant analysis, logistic regression, Bayesian estimates, etcetera (for a detailed overview see Herzog et al., 2007).

1.1.3 Alternative methods for probabilistic record linkage

1.1.3.1 The basic Fellegi-Sunter approach

The early contribution to modern record linkage dates back to Newcombe *et al.* (1959) in the field of health studies, followed by Fellegi and Sunter (1969) where a more general and formal definition of the problem is given. Following the latter approach, let A and B be two partially overlapping files consisting of the same type of entities (individuals, households, firms, etc.) respectively of size n_A and n_B. Let Ω be the set of all possible pairs of records coming from A and B, i.e. $\Omega = \{(a,b) : a \ \in A, b \in B\}$. Suppose also that the two files consist of vectors of variables (X_A) and (X_B), either quantitative or qualitative, and that (X_A) and (X_B) are sub-vectors of k common identifiers, called key variables in what follows, so that any single unit is univocally identified by an observation x. Moreover, let γ^{ab} designate the vector of indicator variables regarding the pair (a,b) so that $\gamma^{ab}_j = 1$ in the j-th position if $x^A_{a,j} = x^B_{b,j}$ and 0 otherwise, $j = 1, \ldots, k$. The indicators γ^{ab}_j will be called *comparison variables.*

Given the definitions above we can formally represent record linkage as the problem of assigning the couple $(a,b) \in \Omega$ to either one of the two subsets M or U, which identify the matched and the unmatched sets of pairs respectively, given the state of the vector γ^{ab}. This assignment can be modelled by a new variable C, which assumes the value 1 for the pairs in M and 0 otherwise.

Probabilistic methods of record linkage generally assume that observations are independent and identically distributed according to appropriate probability distributions. Following Fellegi and Sunter (1969), the bivariate random variable C is latent (unobserved), and it is actually the target of the record linkage process. The comparison variables $\boldsymbol{\gamma}_{ab}$ follow distinct distributions according to the pair status. Let $m(\boldsymbol{\gamma}_{ab})$ be the distribution of the comparison variables given that the pair (a,b) is a matched pair, i.e. $(a,b) \in M$, and $u(\boldsymbol{\gamma}_{ab})$ be the distribution of the comparison variables given that the pair (a,b) is an unmatched pair, i.e. $(a,b) \in U$. These distributions are crucial for deciding the record pairs status, as explained in WP1 of the

ESSnet on Isad (2008).

1.1.3.2 Drawbacks of the Fellegi-Sunter procedure

Decision rules based on the Fellegi-Sunter approach are problematic for different reasons.

1. **Constraints on multiple matches.** Most of the times it is mandatory that each record in file A links to at most one record in file B. The Fellegi-Sunter approach is not able to manage this constraint. It is necessary to apply an optimization procedure to the record linkage results. The interactions between these two procedures and the effects on the record linkage quality have not been investigated yet.

2. **Information on frequency of rare and frequent states of the key variables.** Apart from some nad̆ve approaches described in Fellegi and Sunter (1969) and Winkler (1995), it is often ignored the fact that equalities on a rare state of a key variable supports the idea that the two records refer to the same unit more than when the equality is on a very frequent case.

3. **Model assumption 1**. The Fellegi and Sunter approach assumes that the available observations are the $n_A \times n_B$ pairs of records given by the Cartesian product of the two files A and B. The statistical model is a mixture model that assumes that comparisons on a key variable for different pairs are statistically independent. This is never true, and may influence the results of the record linkage process.

4. **Model assumption 2**. Usually it is a common practice to choose simple models of interaction of key variable comparisons, when the key variables are more than one (i.e. always). By far, the most used model in practice considers the conditional independence of key variable comparisons given the linkage status (true link or not). The appropriateness of this model should be verified in practice, because it does not always hold. Anyway, usual statistical tests fail to give reasonable results.

These problems are tackled by a Bayesian procedure that will be described in Section 1.3. This method will organize differently the data for record linkage, introducing explicitly a possible error model for the key variables used in record linkage, and constraining the unknown parameters to have values according to appropriate prior distributions that reflect the (possible) available knowledge on the amount of overlap between A and B, the amount of error in the key variables and the frequency distributions of the key variables in the population.

1.2 Advances in the Fellegi-Sunter theory

Miguel Guigó

Instituto Nacional de Estadística – INE, Spain

The paragraphs below will show several core issues on the Fellegi-Sunter (FS) model for probabilistic record linkage, which strategies are discussed as follows: the method for calculating a global value of closeness as a starting point is introduced in section 1.2.1; section 1.2.2 deals with statistical inference and parameter estimation; a second method for getting global values based on computer science is proposed in section 1.2.3; some relevant assumptions and procedures within the FS scope are reviewed, together with some critical points of view, in sections 1.2.3.1 and 1.2.3.2.

1.2.1 Comparing common information

Let us assume, following ESSnet on ISAD (2008) and Cibella, Tuoto et al. (2009) that two partially overlapping datasets A and B (in the sense that they are supposed to hold information on at least a subset of individuals whose data are recorded in both A and B) have been previously pre-processed and harmonized, and a group of K common identifying attributes or key matching variables have been chosen in order to compare records in pairs (a, b) one for each set.

Their degree of similarity or disimilarity then depends on a multidimensional (K-dimensional) scale that should be reduced to a single value named ”global weight” or ”composite weight” (Gu et al., 2003), made up of combining values corresponding to every attribute. Newcombe et al. (1959) and Newcombe and Kennedy (1962) offered a solution coherent (though not formalized) with probabilistic and information theory by using log probabilities in the form

$$W\left(a,b\right)=\sum_{k}^{K} w_k; \quad w_k=\log_2 \frac{p_k}{p'_k}=\log_2 p_k-\log_2 p'_k,$$

where the composite weight W for a pair (a, b) is obtained by adding partial weights w_k corresponding to each attribute compared. In its turn, w_k is a log ratio where p and p' are the probabilities of agreement for respectively true matches and pairs accidentally brought together, in case of attributes actually agreeing; otherwise, they should be the probabilities of disagreement. Weights are therefore considered as odds ratios and formulated in terms of frequencies, which are obtained directly from the observed data for a variety of categories of an attribute - for example, different probabilities of agreement (and disagreement) depending of the frequency of each surname. The intuitive and appealing idea is that (1) since probabilities of agreement tend to be high in true matches and low in false matches, weights should be positive in case of coincidence, and (2) since the opposite is true for false matches, weights should be negative when attributes do not coincide.

This approach is partly shared in Copas and Hilton (1990), which goes further with the idea of probability ratios. Given a pair (a, b) of records, a hypothesis test of H_1 (both records relate to the same person or entity) against H_0 (they relate to different entities) can be formulated by taking into account the distributions

$$p(a,b) = P(a,b\,|match); \quad p_a = \sum_b^n p_{ab}; \quad p_b = \sum_a^n p_{ab};$$

where p_{ab} is the probability of selecting a pair a, b belonging to the same individual, and p_a , p_b are the marginal probability distributions for, respectively, a and b of being selected. Then the test can be performed through

$$\frac{p_{ab}}{p_a p_b}$$

provided that a "study file" is available, consisting of a set of matched record pairs. Then, for a given data field, each record can take one of n known values labelled $1, 2, \ldots, n$, and a double-entry table can hold the frequencies or probabilities for each pair of values of being selected when its true status is a match. Moreover, a wide range of models, based on the statistical behaviour of the errors with respect to the correct values, can be fitted in order to calculate those probability ratios.

Two important features must be stressed in this approach. First, the so obtained weights are value-related rather than field-related. This seems that the agreement on a specific comparison field results in different weights depending on whether that common value is rare or not. Second, several degrees or levels of agreement can be achieved and assessed for each field instead of a dichotomous pattern of complete agreement/disagreement. Nevertheless, the need of a study file containing a set of records with their true status known must be also considered as an important constraint in order to implement this approach in practice.

Fellegi and Sunter (1969) added a theoretical background to the idea by Newcombe et al. proposing a composite weight as a function of $\boldsymbol{\gamma} = \boldsymbol{\gamma}(a, b)$ which can be written in the form

$$R\left(\boldsymbol{\gamma}\right)=\frac{m\left(\boldsymbol{\gamma}\right)}{u\left(\boldsymbol{\gamma}\right)}=\frac{\Pr\left(\boldsymbol{\gamma}/r\in M\right)}{\Pr\left(\boldsymbol{\gamma}/r\in U\right)},$$

$$W=\log_2 R\left(\boldsymbol{\gamma}\right)=\log_2\frac{m\left(\boldsymbol{\gamma}\right)}{u\left(\boldsymbol{\gamma}\right)}=\log_2 m\left(\boldsymbol{\gamma}\right)-\log_2 u\left(\boldsymbol{\gamma}\right)$$

where $m(\boldsymbol{\gamma})$ and $u(\boldsymbol{\gamma})$ are conditional density functions, which give the probability of each value $\boldsymbol{\gamma}$ when, respectively, the pair $r=(a,b)$ belongs to the subset of true matched pairs (M) or not (U), provided that $M\cap U=\emptyset$ and $M\cup U=\Omega$, the space of all possible pairs.

Please note that, in this general form, it does not matter in which way values are given to $m(\boldsymbol{\gamma})$, since eventually a value in the form of a probability ratio will be assigned to $m(\boldsymbol{\gamma})$ and thus to each pair of records, $r=(a,b)$. So, as we stated above, the measurement of closeness between two records is given - before taking logs - through a value in a range $(0,+\infty)$. However, the loss of information due to the fact that only complete agreements are reported, is discussed in section 1.2.3.

The reason for adopting such a weight is that $R(\boldsymbol{\gamma})$ can be subsequentially handled as a likelihood ratio whose likelihood functions can be expressed more broadly as $L(\gamma;\theta_1)$ for $m(\boldsymbol{\gamma})$ and $L(\gamma;\theta_0)$ for $u(\boldsymbol{\gamma})$, being θ_1, θ_0 the parameters corresponding, respectively, to the hypotheses $r\in M$ and $r\in U$ (non-observable events), thus allowing to perform a hypothesis test with maximum discriminant power to check whether a pair is more likely to belong to the same entity or not. The Fellegi-Sunter approach also provides criteria to establish acceptance or rejection values; that will be further discussed in section 1.2.3.2.

In order to get specific expressions which be also feasible to handle, Fellegi and Sunter introduce the Conditional Independence Assumption (CIA) for the joint distributions $m(\boldsymbol{\gamma})$ and $u(\boldsymbol{\gamma})$, stating that they can be written as a product of the probability functions $m_k(\gamma_k)$, $u_k(\gamma_k)$ since the behaviour of each γ_k (agreement or not in the k-th field) does not depend on the information contained in the remaining data fields. This assumption, largely discussed, has been rejected by some subsequent approaches that will be introduced in sections 1.2.2 and 1.2.3.1. Then

$$m\left(\boldsymbol{\gamma}\right)=\prod_{k}^{K}m_k^{\gamma_k}\left(1-m_k\right)^{1-\gamma_k}$$

and

$$u(\boldsymbol{\gamma}) = \prod_{k}^{K} u_k^{\gamma_k} (1 - u_k)^{1-\gamma_k}$$

are typically made up of Bernoulli distributions $p_k(\gamma_k) = p_k^{\gamma_k}(1 - p_k)^{1-\gamma_k}$ where p_k is the probability of $\gamma_k = 1$ and its complement, the probability of $\gamma_k = 0$, thus obtaining

$$R(\boldsymbol{\gamma}) = \prod_{k}^{K} \left(\frac{m_k}{u_k}\right)^{\gamma_k} \left(\frac{1 - m_k}{1 - u_k}\right)^{1-\gamma_k}$$

$$\log R(\boldsymbol{\gamma}) = \sum_{k}^{K} \{\gamma_k (\log m_k - \log u_k) + (1 - \gamma_k)(\log[1 - m_k] - \log[1 - u_k])\}$$

The whole model has as unknown parameters $m_1 \ldots m_K$, $u_1 \ldots u_K$, and, since the former are conditional probabilities, $Pr(r \in M)$ and $Pr(r \in U)$, say π and $1 - \pi$. One of the key issues of the FS methodology is then the estimation of those parameters. At this point, the following alternatives can be taken into account:

- To consider or not additional assumptions on the model specification. These usually refer to the number of expected pairs that really match, based on the actual sizes N_A and N_B of the datasets to be merged, and expected N_U and N_M regarding that M and U are subsets of $A \times B$ (or Ω).
- To make use of external files and then handle them as training data with known matching status, or just the data collected from the files A and B themselves, be them the entire files or a training sample.

Both alternatives have been widely adopted and several arguments can be given in favor or against. The results of earlier studies on the same population, in which the true status of pairs has been clerically reviewed, can be extremely useful when available, since they can provide accurate and reliable estimates. On the other hand, inconsistent standards applied for different clerks in different batches may drive to disappointing and deceptive results. Moreover, the assumption of stability in the proportions and other parameters (Winkler, 1999) through different registers or even different populations is highly risky; for example, the frequencies and proportions observed in some fields containing information such as name or surname can dramatically vary depending on the selected site. And, of course, a clerical review specificaly made for a new linkage problem when previous studies are not available can be too expensive and time consuming.

The option of using external training data has been chosen in recent years and for several purposes by a handful of new methods such as machine learning or information retrieval (see Winkler, 2000, and Goiser and Christen, 2006). The latter, using the entire files, has been adopted within the standard FS scope.

1.2.2 Estimating unknown parameters

As it can be deduced from paragraphs above, probabilistic record linkage based on the FS procedure intends to discover whether two records do really belong to the same unit through a model that includes, in its turn, a set of conditional distributions in which the true status of the records must be known. From the probabilistic approach, this can be viewed as a problem related to statistical inference.

Fellegi and Sunter (1969) propose two methods for estimating unknown probabilities m_k, u_k, π, using field value frequencies at A and B.

1.2.2.1 Field value frequencies (I)

The first method assumes that, given a key variable – or *matching field* – it can take, say, J true and error-free different values, with true frequencies

$$
\begin{array}{ll}
f_1^{(A)}, \dots\ f_J^{(A)} & \sum_{j=1}^{J} f_j^{(A)} = N_A \\[2ex]
f_1^{(B)}, \dots\ f_J^{(B)} & \sum_{j=1}^{J} f_j^{(B)} = N_B \\[2ex]
f_1^{(M)}, \dots\ f_J^{(M)} & \sum_{j=1}^{J} f_j^{(M)} = N_M
\end{array}
$$

Then, the probability of agreement is defined for each field value according to: first, the relative frequencies of that value in M (for true matched pairs), or A and B otherwise; second, the probabiltiy that none of the true values is missing nor has been misreported[1], that is, the absence of errors. The

[1] The FS approach also introduces the case where the field value has genuinely changed over time though records in A and B actually belong to the same individual. Anyway, a broad set of similar events can be ignored here without loss of generality.

importance of this proposal is that it deals with probabilities of agreement or disagreement that, even for the same matching field, could differ from one value to another. This leads to the possibility of building value-specific weights for each key variable, which will be discussed at section 1.2.2.4.

1.2.2.2 Field value frequencies (II)

The second proposal is based on the fact that some unconditional probabilites can be directly estimated, starting from the idea that the probability $P(\gamma)$ can be expressed as

$$P(\gamma) = P(\gamma/r \in M)P(r \in M) + P(\gamma/r \in U)P(r \in U)$$

as equally happens to each $P(\gamma_k)$ separately under the CIA. Then, the procedure uses some events – called configurations – related to the probability of γ_k to be an agreement or not while the remaining γ_h hold different values; once their expected proportions are expressed, the conditional probabilities m_k, u_k, π, can be derived from a system of equations[2]. The importance of the statement on $P(\gamma)$ made above, is that it introduces the use of conditional probabilities, and thus the Bayesian perspective, to be introduced at section 1.3.

1.2.2.3 E-M algorithm

Jaro (1989) gives a solution for estimating the set of unknown parameters via maximum likelihood from the sample provided by the current observations, starting from the E-M (expectation-maximization) algorithm initially developed by Dempster, Laird and Rubin (1977), that was conceived for "incomplete data" models where, as actually happens in the FS approach, a subset of variables cannot be directly observed; in this algorithm, the values of the unobserved variables are also estimated together with the rest of parameters, in a model of "complete data". In probabilistic record linkage applications, this variables correspond to the true statuts of r, say $g_r = 1$ when pairs match or $g_r = 0$ otherwise, with probabilities

$$P(\gamma_r = 1) = P(r \in M) = \pi, \quad P(\gamma_r = 01) = P(r \in U) = 1 - \pi$$

the likelihood function

[2] Provided that $K > 2$.

$$L\left(\boldsymbol{\gamma}; m, u, \pi\right) = \prod_{r\in\Omega} \left[m\left(\boldsymbol{\gamma}_r\right) P\left(r \in M\right)\right]^{g_r} \left[u\left(\boldsymbol{\gamma}_r\right) P\left(r \in U\right)\right]^{1-g_r} =$$

$$= \prod_{r\in\Omega} \left[\pi \prod_{k}^{K} m_k^{\gamma_{k,r}} \left(1-m_k\right)^{1-\gamma_{k,r}}\right]^{g_r} \left[\left(1-\pi\right) \prod_{k}^{K} u_k^{\gamma_{k,r}} \left(1-u_k\right)^{1-\gamma_{k,r}}\right]^{1-g_r}$$

and the log likelihood

$$\log L = \sum_{r\in\Omega} g_r \left\{\log\pi + \sum_{k}^{K} \left[\gamma_{k,r}\log m_k - \left(1-\gamma_{k,r}\right)\log\left(1-m_k\right)\right]\right\} +$$

$$+ \sum_{r\in\Omega} \left(1-g_r\right) \left\{\log\left(1-\pi\right) + \sum_{k}^{K} \left[\gamma_{k,r}\log u_k - \left(1-\gamma_{k,r}\right)\log\left(1-u_k\right)\right]\right\}$$

The solution is achieved iteratively; initial estimates $\hat{m}_k^{(0)}$, $\hat{u}_k^{(0)}$, $\hat{\pi}^{(0)}$ can be arbitrarily chosen and $\hat{g}_r^{(p+1)}$ at the p+1-th step are obtained by means of calculating their expectation (E) given $\hat{m}_k^{(p)}$, $\hat{u}_k^{(p)}$, $\hat{\pi}^{(p)}$; and then L is maximized (M) calculating the corresponding values of $\hat{m}_k^{(p+1)}$, $\hat{u}_k^{(p+1)}$, $\hat{\pi}^{(p+1)}$, setting their partial derivatives equal to 0.

So the following phases: (1) the calculation of a global or composite weight for each pair of records based on a likelihood ratio $R(\boldsymbol{\gamma})$ using conditional distributions $m(\boldsymbol{\gamma})$, $u(\boldsymbol{\gamma})$; (2) the estimation of these conditional distributions along with the unconditional distribution P$(r{\in}M)$ via the EM algorithm; and (3) a decision criterion to consider r as a link or not, also proposed by Fellegi and Sunter (1969), make up the cornerstone of the FS scope for probabilistic record linkage.

1.2.2.4 Frequency-based weight scaling

One of the limitations that have been pointed out on the FS approach with respect to the former statement by Newcombe et al. is, along with the CIA, that only field-specific weights, instead of value-specific weights, are taken into consideration.

The FS composite likelihood ratio is the sum of field-specific weights which measures the contribution of agreements or disagreements depending on the relative importance of each key variable as a whole; thus, chance agreements as a consequence of an error are regarded as more feasible in, e.g., a dichotomous variable such as ”gender” (with only two different values) that in a matching field such as surname. Nevertheless, the fact that the probability of agreement between records, when they actually do not match, differs from a value to another, is not taken into account. So, an agreement in the second name within the Central Population Register of Spain for a value such as "Rodríguez" –which frequency is extremely high– is weighted the same as an agreement in the name "Lucini" – which is extremely uncommon.

Yancey (2000) extends the frequency-based approach in the FS model in order to calculate the m and u value-specific probabilities for each field, under the CIA. For a given matching field, we can denote the event "fields agree on both records" $\gamma = 1$ as G, and "both records take the j-th value" as G_j. Then, the event "fields agree on both records and take the j-th value" has the conditional probabilities

$$P(G_j/M) = P(G_j \cap G/M) = P(G_j/G, M)P(G/M)$$

$$P(G_j/U) = P(G_j \cap G/U) = P(G_j/G, U)P(G/U)$$

assumed that the pair is, respectively, a true match or a non-match. Therefore, the value-specific agreement weight should be

$$\frac{m\left(G_j\right)}{u\left(G_j\right)} = \frac{P\left(G_j/M\right)}{P\left(G_j/U\right)} = \frac{P\left(G_j/G, M\right)}{P\left(G_j/G, U\right)} \frac{P\left(G/M\right)}{P\left(G/U\right)}$$

which is the traditional binary agreement weight premultiplied by a probability ratio, which can be estimated once the former is calculated via the EM algorithm. The result is an adjusted weight that takes into account the frequency of each different value of the key variable.

Zhu et al. (2009) propose to improve the record linkage performance by means of a value-specific frequency factor in order to adjust the field-specific weight.

$$W = \sum_{k=1}^{K} \left\{ S_k^{\gamma_k} \log_2 \left(\frac{m_k}{u_k} \right)^{\gamma_k} + \log_2 \left(\frac{1 - m_k}{1 - u_k} \right)^{1-\gamma_k} \right\},$$

$$S_k = \left(\frac{N_k/J_k}{f_k} \right)^{\frac{1}{2}} = \sqrt{\frac{A_k}{f_k}}$$

where W_k is the general scaling factor; N_k is the total number of the values for the field (be them different or not), J_k is the number of unique values for the field, and f_k is the specific frequency of the current value; A_k is then the average frequency for the field. Note that only f_k varies from a current value to another, while the rest of the elements of S_k are constant for each field. As a result, scarce values below the average frequency will result in a high scaling factor and *vice versa*. The original FS weight, however, does not have to be scaled in each and every case, but in the ones corresponding to the most uncommon values, below a chosen cut-off percentile.

In order to evaluate the performance of the frequency-based weight scaling, since it is not equal to a formal likelihood ratio anymore, the former procedures for calibrating false-match rates must be replaced by a specificity (SPEC) and sensitivity (SENS) analysis, together with a positive predictive value (PPV), through a comparison of the results against a *gold standard* of clerically reviewed records. Once record pairs have been identified as false-positives (FP) – when the pair has been declared as a link and actually records do not match, false-negatives (FN), true-positives (TP) and true-negatives (TN), SPEC = TN/(TN+FP), SENS = TP/(TP+FN) and PPV=TP/(TP+FP).

1.2.3 Approximate field comparators

As shown in section 1.2.1 similarity and dissimilarity between records is generally measured on the basis of mere agreements or disagreements on the values of the key variables, given a pair $r = (a,b)$; and thus, the K-dimensional vector γ is typically made up of zeroes or ones. This space of possible comparisons is often regarded as excessively restrictive (Yancey, 2005; Winkler, 2006b), since dichotomous variables do not permit to use values related to partial agreements.

While quantitative data can provide a distance between values, $\delta(a, b)$, such partial agreements are specially difficult to handle in case of comparing fields that contain strings of characters; a situation, however, that occurs very often in record linkage applications. In a major statistical operation such as a census, Porter and Winkler (1997) and Herzog et al. (2007) report that names and surnames may contain typographical errors (transcription, keypunching, etcetera) or legitimate variations[3] that could affect 20%-30% of true matched pairs of records – whose γ_k value would be then computed as ‘"0". Computer science has come to aid statistics in this issue via *string comparators*. A string comparator gives values of partial agreement between two strings, usually mapping the pair into the interval [0,1] in order to subsequently modify the usual weights of the record linkage procedure.

[3] For example, adopting spouse's surname after marriage. This latter case would not result, of course, in any kind of partial agreement.

Jaro (1972) introduced the first string comparator metric, based on an algorithm. Let L_a and L_b be the lengths of two character strings a,b; c, the number of *common* characters – agreeing characters within half of the length of the shorter string; t, the number of transposed characters. A *transposition* happens when a character is common to both strings but it is placed at different positions. Then

$$\Phi\left(a,b\right)=\Phi_1\frac{c}{L_a}+\Phi_2\frac{c}{L_b}+\Phi_3\frac{c-t}{c}$$

$$\Phi_1+\Phi_2+\Phi_3=1\ \ ;\ \ 0\leq\Phi\leq1$$

Bigrams (length-two strings, see ESSnet on ISAD, 2008, section 1) can also be used to build string comparators. The bigram function returns the total number of common bigrams in both strings divided by the average number of bigrams in the two strings (Porter and Winkler, 1997). Therefore, $0 \leq \Phi_{\mathrm{b}} \leq 1$ still holds.

Bilenko and Mooney (2002), Bilenko et al. (2003) or Winkler (1990, 1994, 2004), among others, propose or refers several enhancements to this basic approach. The latter also give procedures for adjusting new weights w'_k for each key variable using Φ_k. This can be done by means of mere substitution, a linear combination, etcetera; w'_k should be then still within the range $0 \leq w'_k \leq +\infty$ though, since w_k or w'_k represents the ratio m_k/u_k in case of agreement or (1-m_k) / (1-u_k) in case of disagreement, it is possible to analyze how does the adjustment affects (increases or decreases) the probabilities for matched and unmatched pairs. An extreme case can illustrate this; a transformation of γ_k into γ'_k given Φ_k can consist of assigning "total agreement" ($\gamma'_k = 1$) when Φ_k is above an upper bound. Then, the frequencies for $\gamma'_k = 1$ will increase and the same will happen to m_k and u_k, while penalizing (1-m_k) and (1-u_k). Therefore this method, on the one hand, takes advantage of the additional information provided by approximate comparators, but on the other hand ignores the difference between a total and partial match. Anyway it is important to take into account that these methods (Porter and Winkler, 1997) are not statistically justified, required constant maintenance and values achieved are highly unstable.

An approach that intends to conciliate the use of approximate field comparators with the traditional FS-Jaro procedure can be found in Yancey (2004b, 2005). Once a measure of dissimilarity $\delta(a,b)$ such that $0 \leq \delta \leq 1$ is built, it is possible to obtain the variable γ, which typically varies from 0 to 1 (by means of, e.g., $\gamma = 1 - \delta$ and then get the field weight

$$W\left(x\right) = \log_2 \frac{m\left(x\right)}{u\left(x\right)} = \log_2 \frac{\Pr\left(\gamma = x/M\right)}{\Pr\left(\gamma = x/U\right)} \qquad 0 \leq x \leq 1$$

with extreme cases, $W(1)$ when $\gamma = 1$, and $W(0)$ when $\gamma = 0$, for complete agreement and disagreement, respectively. The parameters of the model can be then estimated by maximum likelihood via the EM algorithm as in the FS-Jaro model, with the particularity that γ could not be only equal to 0 or 1. Nevertheless, note that it is not necessary to obtain the variable γ since it is enough to directly associate a probability to each value of δ in order to get the field weights.

DuVall, Kerber and Thomas (2010) develop this procedure for the particular case where δ is the Levenshtein distance or *edit-distance* (see Bilenko et al., 2003), which measures the number of *edit operations* – inserts and edits – that transform a string a into another string b. Levenshtein (1966) provides an algorithm to compute the minimum number of edit steps that convert a to b, and max(L_a,L_b) is the maximum number of edit steps. These elements allow to build a standardized distance between field values,

$$\delta_L = \frac{L\left(a, b\right)}{\max\left(L_a, L_b\right)} \qquad 0 \leq \delta_L \leq 1$$

where $L(a,b)$ would be the minimum number of edit steps to transform a into b by means of the Levenshtein algorithm. The conditional distributions $m(\delta)$ and $u(\delta)$ and the corresponding $W(\delta)$ can be then calculated as an approximate comparator extension (ACE) of the FS method. Since it is important to check whether the distributions of the global score for true matches and non-matches are well separated, sample means and variances of such scores can be calculated in order to compare both distributions via a Welch two-sample test.

1.2.3.1 On the conditional independence assumption

As pointed out in the section on how weights are calculated, the FS model makes the assumption that linking variables are statistically independent given the true status of a pair, and then the distributions $m(\boldsymbol{\gamma})$ and $u(\boldsymbol{\gamma})$ can be expressed as products of the distributions of their respective components; this means that, first, the lack of agreement in a linking variable is not correlated with the lack of agreement in other variable when the pair is a match; and, second, that chance agreements are neither correlated among false matches. This is a critical point in the model and it has been critizised as too simplistic, but at the same time it is difficult to analyse the dependency between errors since their frequencies are seldom high – specially in case of chance agreement. Moreover, even existing evidence of correlation, sometimes it does not seem to yield poor results in practice (Winkler, 1989, 1994) or it is not feasible to remove any subset of variables given their short availability.

Following Thibaudeau (1993), the conditional probabilities $m(\gamma)$ and $u(\gamma)$ could be expressed, instead in terms of the well-known Bernoulli distributions, as a model based on a latent or non-observable random variable, which is obviously the true status C of each pair ($C = 1$ for true matches M, and $C = 0$ for non-matches U).

Frequencies and then relative frequencies and probabilities for each value γ can thus be expressed through a set of parameters whose values depend on C and on g_1, g_2 $\ldots g_K$ and, once estimated, represents precisely the behavior of each pair.

Let $v\left(\gamma\right) = v\left(\gamma_1, \gamma_2, \ldots \gamma_K\right) = v_M\left(\gamma_1, \gamma_2, \ldots \gamma_K\right) + v_U\left(\gamma_1, \gamma_2, \ldots \gamma_K\right)$ be the frequency of pairs with the value γ for the comparison vector, which are only observable in the aggregated form that the left-hand side of the equality shows. The right-hand side shows the count of pairs for each subset of true, respectively, matches and non-matches. In its logarithmic form it can be modelled as

$$\log v\left(C, \gamma_1, \gamma_2, \ldots \gamma_K\right) = \mu + \beta\left(C\right) + \sum_k \alpha_k\left(\gamma_k\right) + \sum_k \xi_k\left(C, \gamma_k\right)$$

where μ can be regarded as an average value, that is modified by an specific parameter β depending on the true status of the pair, a set of parameters α_k depending on the pattern of agreement, and another set of parameters ξ_k that retrieve the interactions between a field and the latent variable. The effect of both true status and agreements on the probability of each pair is based on the constraints

$$\beta\left(U\right) = -\beta\left(M\right) \; ; \; \alpha_k\left(0\right) = -\alpha_k\left(1\right) \; ;$$

$$\xi_k\left(U, \gamma_k\right) = -\xi_k\left(M, \gamma_k\right) \; ; \; \xi_k\left(C, 0\right) = -\xi_k\left(C, 1\right),$$

and related parameters are likewise estimated via the the EM algorithm shown in 1.2.2.3. Note that this basic model does not not tackle the problem of dependence between the comparison fields, thus it does not basically differ from the FS approach with respect to the conditional independence assumption. It is the use of additional terms in the form

$$\log v\left(C,\gamma_1,\gamma_2,\ldots\gamma_K\right)=\mu+\beta\left(C\right)+\sum_k\alpha_k\left(\gamma_k\right)+\sum_k\xi_k\left(C,\gamma_k\right)$$
$$+\left(1-C\right)\sum_{k<l}\eta_{kl}\left(\gamma_k,\gamma_l\right)$$

$$\eta_{kl}\left(\gamma_k,0\right)=-\eta_{kl}\left(\gamma_k,1\right)\ ;\ \eta_{kl}\left(0,\gamma_l\right)=-\eta_{kl}\left(1,\gamma_l\right)$$

what handles the interactions between any couple of comparison variables given the true status of the pair. The term (1 - C) is introduced due to the observed fact that significant correlations between probabilities of field agreements only arise in case of actual non-matches.

In fact, not only interactions between couples of variables but also those concerning three-variable groups or even more should be estimated, provided that the corresponding restrictions are also added. Nevertheless, due to limitations of the optimization algorithm, the most advisable strategy consists of first estimate the conditional independence model and then use the so-obtained values as a starting point to add a subset of variables with suspected interactions once the corresponding correlation matrix has been examined. Given that subset of K' variables, interactions between two-variable, three-variable and up to K'-variable groups should be estimated.

Tromp et al. (2008) build a model which assumes dependence between a couple of key variables – say the h-th and the l-th– whose pattern of agreements should be them strongly related, keeping the assumption on conditional independence for the remaining K-2 ones. A new set of parameters, say m_1, $\ldots m^*_h$, $\ldots m^*_l$, $\ldots m_K$, m_{hl}, u_1, $\ldots u^*_h$, $\ldots u^*_l$, $\ldots u_K$, u_{hl}, π, is considered. The probabilities m_{hl}, u_{hl}, correspond to those cases where, for a pair of records, both h-th and l-th fields agree; m^*_h, m^*_l, u^*_h, u^*_l, correspond to those where only one of both fields agree; 1- m^*_h- m^*_l - m_{hl} and 1- u^*_h- u^*_l - u_{hl} express disagreement in both fields.

A variant of the FS-Jaro likelihood function is used, removing m_h, $m_{l,}u_h$, $u_{l,}$ and their complements and placing the probabilities described above. Counters γ_h, γ_l are also substituted by indicators $\mathbf{I}(\gamma)$, with $\mathbf{I}(\gamma)=1$ when the corresponding configuration of agreements in h-th and l-th fields actually happens, or $\mathbf{I}(\gamma)=0$ otherwise. Then, parameters are estimated using the EM algorithm. A case study with data on childs from the Dutch perinatal registers shows that weights for agreement calculated under CIA were considerabily higher than those yielded taken into account condional dependence between two highly correlated variables.

1.2.3.2 Considering records as links and non-links

The ultimate goal of calculating a ratio $R(\boldsymbol{\gamma})$ is laying down a rule, based on statistical inference, to decide on the assumed status of each pair r for which such ratio has been obtained; Fellegi and Sunter give a detailed procedure that depends on the error rates that are expected to yield in terms of records wronlgy matched and wrongly discarded as matches (see ESSnet on ISAD, 2008).

Though applying in the record-linkage context the Neyman-Pearson lemma, which states that the ratio of two likelihood functions with alternative parameters – in our case $m(\boldsymbol{\gamma})$, $u(\boldsymbol{\gamma})$ – can give the best acceptance and rejection regions for a hypothesis – the records are matched or not – given a maximum affordable error, Fellegi and Sunter provide a specific theorem on the construction and properties of an optimal linkage rule. Three decisions A_1, A_2, A_3, are possible, depending on whether each pair is declared as: "*a link*" (A_1), "*a non-link*" (A_3), or "*a possible link*" (A_2) subject to a later clerical review. Two admissible error rates are taken into account:

$\mu = \sum_{\gamma} P\left(A_1/\boldsymbol{\gamma}\right) u\left(\boldsymbol{\gamma}\right)$ (1) for false matches wrongly declared as *links*, and
$\lambda = \sum_{\gamma} P\left(A_3/\boldsymbol{\gamma}\right) m\left(\boldsymbol{\gamma}\right)$ (2) for true matches wrongly declared as *non-links*[4].

Since the structure of the model does not supply a continuous and monotonically increasing or decreasing function $R(\boldsymbol{\gamma})$, values of R are just sorted from the highest to the lowest, say $R(\boldsymbol{\gamma})_{(1),}R(\boldsymbol{\gamma})_{(2)} \ldots R(\boldsymbol{\gamma})_{(L)}$; the corresponding values $u(\boldsymbol{\gamma})_{(h)}$ for each $R(\boldsymbol{\gamma})_{(h)}$ are selected to be included in (1), starting from $u(\boldsymbol{\gamma})_{(1)}$ and until the set-up value of m is achieved; then, if $u(\boldsymbol{\gamma})_{(n)}$ is the last item to be added, $R(\boldsymbol{\gamma})_{(n)}$ is the upper cut-off threshold. A similar procedure can be followed for (2), starting from $m(\boldsymbol{\gamma})_{(L)}$ to $m(\boldsymbol{\gamma})_{(n')}$and getting the lower cut-off threshold $R(\boldsymbol{\gamma})_{(n')}$.

The parameter estimation via the EM algorithm following the solution by

[4] Usually, a given value γ will correspond to a unique decision A_i, which conditional probability on γ will be then "0" or "1". The general approach shown above stands only for cases where μ and λ cannot be exactly achieved by adding particular $m(\boldsymbol{\gamma})$'s and $u(\boldsymbol{\gamma})$'s in each error rate, and then for boundary values of γ, one decision or another is randomly taken.

Jaro (1989), since provides directly a maximum-likelihodd estimation of the proportion π of true matches, can offer a more straightforward method to establish a unique bound, though ignoring maximum error rates. Tromp et al. (2008) merely sort all record pairs by descending total weight and then count backward the number of estimated matches; once that number has been reached, the corresponding weight is accepted as the threshold value.

Belin and Rubin (1995) consider the FS method extremely inaccurate in the sense that false-match rates are underestimated, after looking up empirical evidences from training data, clerically reviewed once the FS rule was used. However, though their approach seems to criticize just the decision rule and the estimation of false-match rates, it also questions the C.I.A. and introduces an alternative that can also considered as a starting point for the Bayesian approach. No matter what kind of weight W is adopted, even $R(\boldsymbol{\gamma})$, it is important to know more about the specific distribution of W.

They assume that its observed distribution is the result of merging either cases when pairs match and do not match. Moreover, they introduce a mixture model whose general form is

$$f\left(W/Z,\boldsymbol{\theta}\right)=f_1\left(W/\boldsymbol{\theta}_1\right)Z+f_2\left(W/\boldsymbol{\theta}_2\right)\,\left(1-Z\right)$$
$$P\left(Z/\pi\right)=\pi^Z\left(1-\pi\right)^{1-Z}$$

and could be assimilated to the equation showed in the paragraph on Jaro (1989), since Z_r is the true status of each pair (Z_r =1 when $r\in M$ and Z_r = 0 when $r\in U$) and π = Pr($r\in M$). That seems in principle, no difference with FS scope; but they consider feasible to make additional assumptions on the behaviour of $f\left(W_r/Z_r,\theta\right)$ instead of the traditional view –that should be based then on $m(\boldsymbol{\gamma})$ = P(γ/ Z_r =1) and $u(\boldsymbol{\gamma})$ = P(γ/ Z_r = 0) and the C.I.A. –, considering$f_1\left(W_r/Z_r=1,\boldsymbol{\theta}_1\right)$ and $f_2\left(W_r/Z_r=0,\boldsymbol{\theta}_2\right)$ two different but typically normal distributions or, if not, that it is possible to transform them into two normal distributions; the likelihood then becomes

$$L\left(W,Z;\boldsymbol{\theta},\pi\right)=\prod_{r\in\Omega}\left[f_1\left(W_r/\boldsymbol{\theta}_1\right)\pi\right]^{Z_r}\left[f_2\left(W_r/\boldsymbol{\theta}_2\right)\,\left(1-\pi\right)\right]^{1-Z_r}$$

from a (normal) mixture model where θ_1, θ_2 and π must be estimated to completely characterize the model; with θ_i = μ_i, σ_i^2; the difference between means should be large enough to identify both distributions separately.

In order to achieve normal distributions from the current distributions of the weights, they fall back on a family of well-known power transformations proposed by Box and Cox (1964), which have proven good practical results in a variety of applications (see for example Box, Jenkins and Reinsel, 1994, p.358); then

$$W'_r = \begin{cases} \left(W_r^{\lambda} - 1\right) / \lambda W_r^{\lambda - 1} & \lambda \neq 0 \\ \omega \lg W_r & \lambda = 0 \end{cases}$$

provided that transformations are not the same for f_1 and f_2. Therefore, two additional parameters λ and ω are needed for each distribution, though they are estimated in training samples previously reviewed, where the true status of the pairs is known, looking for the values that best fit to a normal distribution. It is also assumed that those are "global" parameters that remain stable for any scenario, provided that they are different for true matches and for non-matches.

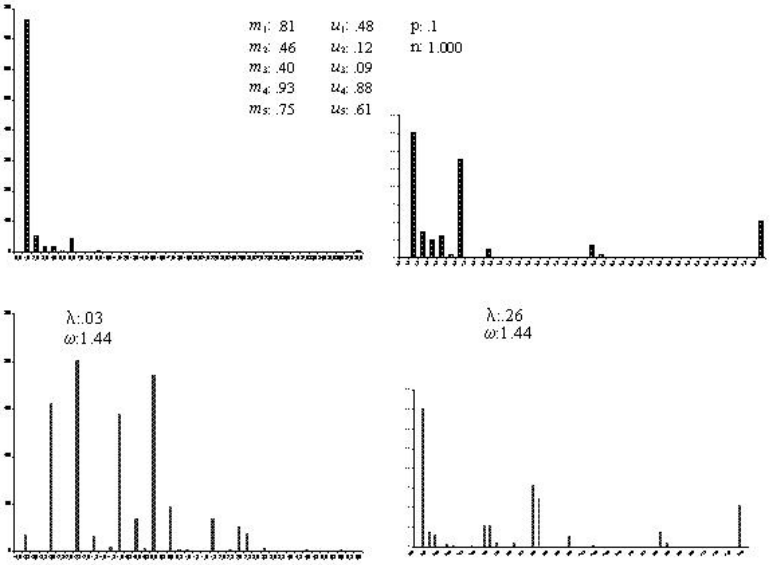


Figure 1.1. Frequency distributions of the global weight W, calculated as $R(\gamma)$ following the FS procedure, from a simulation of a training sample $n = 1{,}000$ obs., before (above) and after (below) a Box-Cox power transformation. The data generating process is a mixture model of the form $L = \prod_{r \in \Omega} \left[\pi \cdot m\left(\gamma_r\right)\right]^{Z_r} \left[(1-\pi) \cdot u\left(\gamma_r\right)\right]^{1-Z_r}$, under c.i.a. with known parameters m_k, u_k, K=5 and $\pi = .1$. Distributions of true matches (right) and non-matches (left) are shown separately. Power parameters λ are arbitrarily selected and ω is the common geometric mean.

Then, μ_i, σ_i^2, and π are estimated via the EM algorithm in a similar way that Jaro, calculating $\hat{Z}_r^{(p+1)}$ at the E step, and $\hat{\mu}_i^{(p+1)}$, $\hat{\sigma}_i^{2(p+1)}$, $\hat{\pi}^{(p+1)}$ at the M step. Standard errors of the parameters are also given via the SEM algorithm (see Meng and Rubin, 1991).

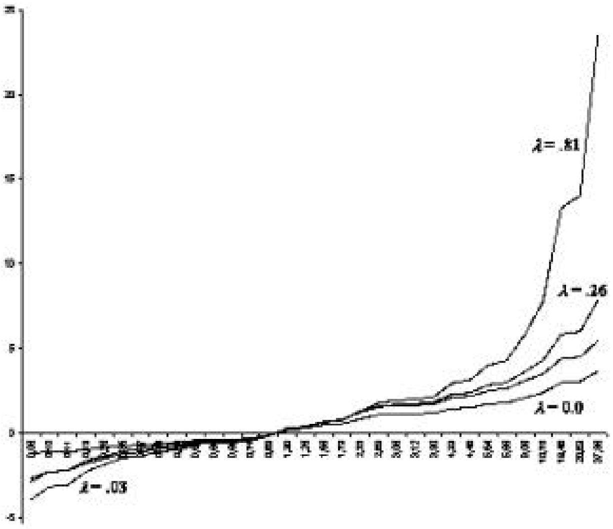


Figure 1.2. Several candidate transformations for the global weight W in terms of λ, the abscissa showing the original values of W. Note that $\lambda = 0$ results in lnW, as proposed in the traditional FS procedure.

The model then provides methods for estimating false-match rates given a cutoff and for each record pair given its weight. Neverthless, an important idea emerging from this approach is the suggestion that, derived from the (normal) mixture model given above and once obtained the estimates of θ_1, θ_2, $P(Z_r/W_r)$ can be found using Bayes's Theorem. This has been also stated above from the frequentist perspective by Fellegi and Sunter and gives the starting point for the Bayesian approach.

1.3 A Bayesian model

Nicoletta Cibella, Mauro Scanu, Tiziana Tuoto

Istituto nazionale di statistica – Istat, Italy

As remarked in Section 1.1.3.2 there are some problems that are not directly solved by the application of the Fellegi-Sunter procedure (constraints, use of frequencies of rare categories). Furthermore, they rely heavily on model assumptions that are not valid. A Bayesian procedure by Tancredi and Liseo (2010) tackles this problem. The following section summarizes it.

The Bayesian approach needs some additional notation when compared to the one used until now. As usual:

- N is the unknown size of the whole population of interest

- A and B are two subsets of the population
- X represents the vector of key variables used to link A and B, V represents the sets of possible values of the vector X. V consists of k different vectors of values (i.e. k corresponds to the product of the number of states of each key variable)

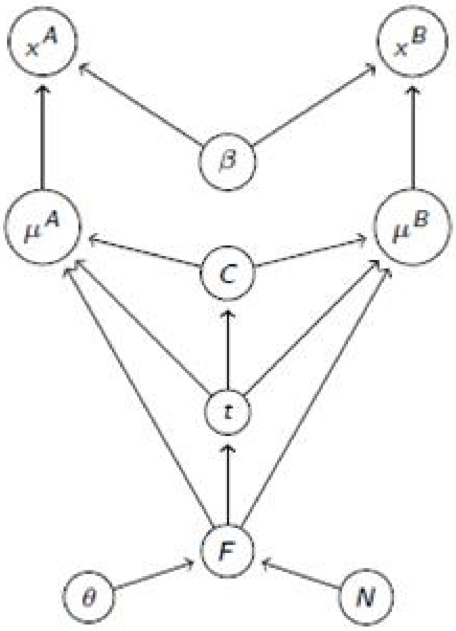
It is necessary to distinguish between variables and parameters.

The Bayesian model in Tancredi and Liseo (2010) considers the following variables:

μ_A: is the vector of true values for the key variables on the n_A records in A.
μ_B: is the vector of true values for the key variables on the n_B records in B.
X_A: is the vector of observed key variables on the n_A records in A.
X_B: is the vector of observed key variables on the n_B records in B.
C: is the matrix of the true status of each pair (a,b) as matches (c_{ab}=1) or non matches (c_{ab} =0).
F: the frequency distribution of X on the whole population of N units.

Note that only X is actually observed.

Tancredi and Liseo suggest modeling all these variables in a multivariate distribution that follows this graphical model:



The graphical model, known also as Bayesian network (Lauritzen, 1996), describes in a graphical way the set of dependencies and independencies between the variables: absence of an arrow corresponds to (conditional) independencies between the variables (for instance, in the picture above x^A is independent of x^B given β). The graphical model decomposes the multivariate distribution of the variables of interest in the different conditional distribution of each variable given their parents (i.e. those variables which are connected with the variable of interest by a direct arrow). Hence, the previous graphical model specifies the following multivariate distribution:

$$\begin{aligned} P\left(x^A, x^B, \mu^A, \mu^B, C, t, F, \beta, \theta, N\right) = {} & P\left(x^A, x^B \left|\beta, \mu^A, \mu^B\right.\right) P(\beta) P\left(\mu^A, \mu^B \left|C, t, F\right.\right) \\ & P(C\left|t\right.) P(t\left|F\right.) P(F\left|\theta, N\right.) P(\theta) P(N) \end{aligned}$$

Tancredi and Liseo suggest the following distributions for each factor of the previous multivariate distribution.

- X_A and X_B: these vectors depend on the corresponding true values according to the formula:

  $$p\left(x^i = v^i_j | \mu^i = v^i_{j'_i}\right) = \beta_i I_{\left\{v^i_{j_i} = v^i_{j'_i}\right\}} + (1-\beta_i)\,\psi_{j_i}, \quad i = 1, \dots, h$$

  where $\psi_{j_i} = 1/k_i$ (this corresponds to a simple version of the hit-miss model as described in Copas and Hilton, 1990) and β is the probability of measurement error for the key variable X_i for i=1,. . . ,h.

- μ_A and μ_B are assumed to be two independent simple random samples drawn from the population of unknown size *N:*

  $$p\left(\mu^A, \mu^B | F\right) = p\left(\mu^A | F\right) p\left(\mu^B | F\right)$$

  In principle

  $$p\left(\mu^S | F\right) = \frac{1}{\binom{n^S}{f_1^S, \dots, f_k^S}} \frac{\prod_{j=1}^{k} \binom{F_j}{f_j^S}}{\binom{N}{n_S}} \qquad S = A, B$$

  where $f^S = (f_1^S, .., f_j^S, .., f_k^S)$ are the unobserved true sample counts for each element of V.

  The above model may also be written using the latent structure that explicitly introduce the matching matrix C and the vector t.

1. The configuration matrix C is assumed to be constrained so that each record in A can be linked with at most one record in B and vice versa, i.e. C the sum of the values of C by row or by column can be at maximum equal to 1;
2. t is a vector $(t_1,\dots,t_k)$ of as many values as the categories of the set of key variables (k); t_j represents the number of matches among the units whose true value is equal to v_j, j=1,... k.

The number of different configuration matrix C is equal to

$$\binom{n_A}{T}\binom{n_B}{T}T!$$

where

$$T=\sum_{j=1}^{k}t_j=\sum_{ab}C_{ab}\leq\min\left(n_A,n_B\right).$$

The distribution for μ_A and μ_B is taken by randomly sampling units from the groups of units with the same true value v_j in different groups, i.e. the matches, the non matches, and the non sampled. Hence, a natural distribution is:

$$\begin{aligned}
p\ &\left(\mu^A,\mu^B|C,t,F\right)=\\
=&\ \frac{\prod_{j=1}^{k}\binom{F_j-t_j}{f_j^A-t_j,f_j^B-t_j,F_j-f_j^A-f_j^B+t_j}}{\binom{N-T}{n^A-T,n^B-T,N-n^A-n^B+T}}\frac{\prod_{j=1}^{k}t_j!\left(f_j^A-t_j\right)!\left(f_j^B-t_j\right)!}{T!\left(n^A-T\right)!\left(n^B-T\right)!}
\end{aligned}$$

- C is assumed to be a uniform random variable among the different possible configuration matrices:

$$p\left(C|t\right)=\left[\binom{n^A}{T}\binom{n^B}{T}T!\right]^{-1}$$

- t follows a multinomial distribution given T and the vector F of true frequencies in V, while T follows a hypergeometric distribution.

$$p\left(t|F\right)=p\left(t|T,F\right)p\left(T|F\right)=\left[\prod_{j=1}^{k}\binom{F_j}{t_j}/\binom{N}{T}\right]\binom{n^A}{T}\binom{N-n^A}{n^B-T}/\binom{N}{n^B}$$

- F follows a multinomial distribution with parameters N (i.e. the population size) and θ, i.e. the parameters of a superpopulation model.

  As far as the parameters in the model are concerned, apart from C which is already been defined as a uniform, Tancredi and Liseo suggest the use of the following standard prior distributions.

- N is assumed to follow a non-informative prior distribution:

$$p\left(N\right) \propto \Gamma\left(N - g + 1\right)/N!, \quad g \geq 0.$$

- θ is assumed to follow a hyper-Dirichlet distribution.
- β is a vector of uniform random variables.

This model can be used for different record linkage purposes: estimation of N (in this case it is preferable to marginalize the previous distribution with respect to C), estimation of C. This model can be in principle modified in order to visualize other parameters of interest, as a correlation coefficient between two variables of interest observed respectively in A and B.

Finally, β is a measure of the measurement error in the two occasions A and B.

1.4 Efficient Blocking

Gervasio-Luis Fernández Trasobares

Instituto Nacional de Estadística – INE, Spain

1.4.1 Background

Blocking can be regarded as a search for a set cover (o a set partition) of the target sets $\{A; B\}$:

$$\{A_i, B_i \quad i = 1, \ldots, m\}$$

$$\{A = \bigcup_{i=1}^{m} A_i \quad \text{(a set partition:} \quad A_i \cap A_j = \emptyset \quad \forall i \neq j)$$

$$\{B = \bigcup_{i=1}^{m} B_i \quad \text{(a set partition:} \quad B_i \cap B_j = \emptyset \quad \forall i \neq j)$$

then, the space of cross-products is not $A \times B$ anymore, more, but it is made of the corresponding $S = \bigcup_{i=1}^{m} A_i \times B_i$.

The most efficient set of subsets is achieved by means of minimizing their size, provided that as many record pairs belonging to M (true matches) as possible are still feasible:

$$\begin{array}{l} \min \{\text{Card}\{S\}\} \quad \wedge \\ \max \{\text{Card}\{M \bigcap S\}\} \end{array}$$

Standard blocking is usually implemented by means of sorting files on a variable, that can consist of several concatenated attributes, and then each block is specified by a key: $A_i = \{x \in A/V(x) = k_i\}$

Some other traditional blocking techniques rely on a subset of appropriately representative individuals $\{x_i/i = 1, \ldots, m\}$, so that blocks are built by means of a distance or similarity measure, with respect to those representative units: $A_i = \{x \in A/d(x, x_i) \leq w_i\}$ or $A_i = \{x \in A/S(x, x_i) \geq w_i\}$.

As a sum up of these procedures, they consist of a search for attributes that permit an efficient block specification, or to use a measure for similarity or distance between individuals, in order to choose the most representative among them for the whole data.

A new approach arises when block specification is made directly via Rules or Predicates: in such a case, procedures related to Machine Learning can be used, in order to fix the set of those Rules or Predicates that provide an efficient block structure.

Block	Rule or Predicate
A_i	$R_i = \{method, attribute, value\}$
$A_i = \{x \in A/V(x) = k_i\}$	$R_i = \{V(.), x, k_i\}$
$A_i = \{x \in A/d(x, x_i) \leq w_i\}$	$R_i = \{d(., x_i), x, w_i\}$
$A_i = \{x \in A/S(x, x_i) \geq w_i\}$	$R_i = \{S(., x_i), x, w_i\}$

As regards this latter approach, one of its most important drawbacks, as it has been pointed out, is the need of a previous training data. For the particular case of merging survey and administrative data, though, this may not be seen as a very important issue, since one main feature of this sort of statistical processes is its repetition on a regular basis. Experience acquired in previous years or surveys ought to be useful in order to specify a training data that properly suits the matter to deal with.

Alternatively, blocking could be considered as a matter of record classification, since every block represents those records to be brought together as part of the same class. Many of the procedures used for classification purposes are actually analogous to those already adopted in traditional blocking techniques, and some other are born of more recent approaches.

Comparative analysis among several blocking procedures can be found at references listed below: Christen P. (2007), Christen P. and Gayler R. (2008) and On B.W., Lee D., Kang J. and Mitra P. (2005).

1.4.2 Traditional techniques

1.4.2.1 Suffix Array-Based Blocking

Aizawa A. and Oyama K. (2005) proposes a method for fast detection of matched pairs of records. At a first step, blocks are generated from an index made of variable length tokens, and then a subset of them, according to a set of criteria, is selected. At the second step, automatically extracts the blocking keys from already known reliable links in the previous blocks and to obtain this way, the appropriate and definitive blocks.

1.4.2.2 Sorted Neighborhood Methods

Yan S.; Lee D.; Kan M.Y. and Giles C.L. (2007) proposes an adaptive algorithm which automatically modifies some of the parameters used in a Sorted Neighbourhood Method (SNM) algorithm. In this case, its sliding window size.

1.4.3 Rule-Based and Predicate-Based Techniques

1.4.3.1 Predicate-Based Formulations of Learnable Blocking Functions

Bilenko M.; Kamath B. and Mooney R.J. (2006) proposes a general framework for machine learning of blocking functions from general predicates.

1.4.3.2 Sequential covering algorithm to discover disjunctive sets of rules

Michelson M. and Knoblock C.A. (2006) proposes a method for sequential learning of disjunctive rules which gradually cover the subset of true matched pairs within the whole training data.

1.4.4 Modern classification techniques

1.4.4.1 Seeded Nearest Neighbour and Support Vector Machine Classification

Christen P. (2008) proposes a new two-step approach for automatic record linkage. In a first step, some examples with high quality data from training sets are automatically retrieved, with the purpose of bringing together record pairs to be compared. In a second step, the former examples are used in order to train a classifier based on a Vector Support Machine (VSM).

1.4.4.2 Efficient Clustering

Yin X.; Han J. and Yu P. (2006) proposes a procedure for hierarchical representation of similarities between objects and the corresponding calculation in an efficient way. This permits to build clusters that can therefore be used as blocks.

References on probabilistic record linkage

[1] Aizawa A. and Oyama K. (2005), *A fast linkage detection scheme for multi-source information integration*, Web Information Retrieval and Integration (WIRI'05) pages 30–39. Tokio. 2005.

[2] Baxter R., Christen P. and Churches T. (2003), *A Comparison of fast blocking methods for record linkage*, Proceedings of 9th ACM SIGKDD Workshop on Data Cleaning, Record Linkage, and Object Consolidation, August 2003: 25–27, Washington, DC, USA.

[3] Belin T.R. and Rubin D.B. (1995), *A method for calibrating false-match rates in record linkage*, Journal of the American Statistical Association, Volume 90: 694–707.

[4] Bilenko M. and Mooney R.J. (2002), *Learning to Combine Trained Distance Metrics for Duplicates Detection in Databases*, Technical Report AI-02-296, University of Texas at Austin, Feb 2002.

[5] Bilenko M., Mooney, R., Cohen W., Ravikumar P. and Fienberg S. (2003), *Adaptive Name Matching in Information Integration*, IEEE Intelligent Systems 18(5): 16–23 (2003). IEEE Computer Society.

[6] Bilenko M.; Kamath B. and Mooney R.J. (2006), *Adaptive blocking: Learnig to scale up record linkage*, IEEE International Conference on Data Mining (ICDM'06) pages 87–96. Hong Kong. 2006.

[7] Box G. E. P., Jenkins G. M. and Reinsel G. C. (1994), *Time Series Analysis, Forecasting and Control*, (3rd ed.) Prentice. Hall, Englewood Cliffs.

[8] Box G.E.P. and D.R. Cox (1964), *An analysis of transformations (with discussion)*, Journal of the Royal Statistical Society B.26: 211–246.

[9] Christen P. (2007), *Towards Parameter-free Blocking for Scalable Record Linkage*, Joint Computer Science Technical Report Series TR-CS-07-03, The Australia National University. Canberra. Australia.

[10] Christen P. (2008), *Automatic record linkage using seeded nearest neighbor and support vector machine classification*, ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD'08) pages 151-159. Las Vegas. Nevada. USA.

[11] Christen P. and Gayler R. (2008), *Towards scalable real-time entity resolution using a similarity-aware inverted index approch*, In AusDM'08. CRPIT vol. 87. Glenelg. Australia. 2008.

[12] Christen P.; Gayler R. and Hawking D. (2009), *Similarity-aware indexing for real-time entity resolution*, Technical Report TR-CS-09-01. School of Computer Science; The Australian National University; Canberra; Australia.

[13] Cibella N., Fernández G.-L., Fortini M., Guigó M., Hernández F., Scannapieco M., Tosco L., Tuoto T. (2009), *Sharing Solutions for Record Linkage: the RELAIS Software and the Italian and Spanish Experiences*, In Proc. Of the NTTS (New Techniques and Technologies for Statistics) Conference, Bruxelles, Belgium, 2009.

[14] Copas, J. R., and Hilton, F. J. (1990), *Record linkage: statistical models for matching computer records*, Journal of the Royal Statistical Society, A, Volume 153:287–320.

[15] Dempster A.P., Laird N.M. and Rubin D.B. (1977), *Maximum Likelihood From Incomplete Data via the EM Algorithm*, JournalEM Algorithm. Journal of the Royal Statistical Society Series B, 39(1):1–38.

[16] DuVall S.L., Kerber R.A., Thomas A. (2010), *Extending the Fellegi-Sunter probabilistic record linkage method for approximate field comparators*, Journal of Biomedical Informatics, Volume 43, Issue 1, February 2010: 24–30.

[17] ESSnet on ISAD (2008), *Report of WP1. State of the art on statistical methodologies for integration of surveys and administrative data*, http://cenex-isad.istat.it.

[18] Fellegi I.P. and Sunter A.B. (1969), *A theory for record linkage*, Journal of the American Statistical Association, Volume 64, 1183–1210.

[19] Goiser K. and Christen P. (2006), *Towards Automated Record Linkage*, Proceedings of the Fifth Australasian Data Mining Conference (AusDM2006), Sydney.

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

[21] Herzog T. (2004), *Playing with Matches: Exploring Data Quality Issues With an Emphasis on Record Linkage Techniques*, SOA 2004 New York Annual Meeting - 18TS, Playing With Matches: Applications of Record Linkage Techniques and Other Data Quality Procedures.

[22] Herzog T.N., Scheuren F.J. and Winkler W.E. (2007), *Data Quality and Record Linkage Techniques*, Springer Science+Business Media, New York.

[23] Jaro M.A. (1972), *UNIMATCH: a computer system for generalized record linkage under conditions of uncertainty*, Proceedings of the 1972 Spring Joint Computer Conference, Session: The computer in government: a tool for change: 523-530. American Federation of Information Processing Societies, New York.

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

[25] Lauritzen, S.L. (1996), *Graphical Models*, Clarendon Press, Oxford.

[26] Levenshtein V. (1966), *Binary codes capable of correcting deletions, insertions, and reversals*, Soviet Physics Doklady 10(8): 707–710. For an available reference, see Navarro, G. (2001) A guided tour to approximate string matching. ACM Computing Surveys. Vol. 33 (1): 31–88.

[27] Meng X. and Rubin D.B. (1991), *Using EM to Obtain AsymptoticEM to Obtain Asymptotic Variance-Covariance Matrices: the SEM Algorithm*, Journal of the American Statistical Association, 86, 899–909.

[28] Michelson M. and Knoblock C.A. (2006), *Learning blocking schemes for record linkage*, National Conference on Artificial Intelligence (AAAI-2006). Boston; 2006.

[29] Newcombe H. B. and Kennedy J. M. (1962), *Record linkage, making maximum use of the discriminating power of Identifying information*, Communication of the Association for Computing Machinery, Volume 5(11), pp. 563–566.

[30] Newcombe H. B., Kennedy J.M., Axford S.J. and James A.P. (1959), *Automatic linkage of vital records*, Science, 130: 954–959.

[31] On B.W.; Lee D.; Kang J. and Mitra P. (2005), *Comparative Study of Name Disambiguation Problem using a Scalable Blocking-based Framework*, In ACM/IEEE Joint Conf. on Digital Libraries (JCDL). Jun. 2005.

[32] Porter E.H. and Winkler W.E. (1997), *Approximate string comparison and its effect on advanced record linkage system*, Bureau of the Census, Statistical Research Division, Statistical Research Report Series, n. RR97/02.

[33] Tancredi, A. and Liseo, B. (2010), *A hierarchical Bayesian approach to matching and size population problems.*

[34] Thibaudeau Y. (1993), *The discrimination power of dependency structures in record linkage*, Survey Methodology, Volume 19, pp. 31–38.

[35] Tromp M., Méray N., Ravelli Anita C. J., Johannes B. R., Gouke J. B. (2008), *Ignoring dependency between Linking Variables an Its Impact on the Outcome of Probabilistic Record Linkage Studies*, Journal of the American Medical Informatics Association 2008 15: 654–660.

[36] Winkler W.E. (1989), *Methods for adjusting for lack of independence in the Fellegi-Sunter model of record linkage*, Survey Methodology, Volume 15, pp. 101–117.

[37] Winkler W.E. (1990), *String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage*, Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 354–359.

[38] Winkler W.E. (1994), *Advanced methods for record linkage*, Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 467–472.

[39] Winkler W.E. (1995a), *Matching and Record Linkage*, Business Survey Methods (Cox B.G., Binder D.A., Chinnappa B.N., Christianson A., Colledge M., Kott P.S. (eds)), pp. 355–384. Wiley, New York.

[40] Winkler W.E. (1999), *The state of record linkage and current research problems*, U.S. Bureau of the Census, Statistical Research Report Series, No. RR1999/04. U.S. Bureau of the Census, Washington, D.C.

[41] Winkler W.E. (2000), *Machine learning, information retrieval and record linkage*, Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 20–29.

[42] Winkler W.E. (2004), *Methods for Evaluating and Creating Data Quality*, Information Systems (2004), 29 (7): 531–550.

[43] Winkler W.E. (2006a), *Overview of Record Linkage and Current Research Directions*, U.S. Bureau of the Census, Statistical Research Division Report Series, n.2006/2.

[44] Winkler W.E. (2006b), *Data Quality: Automated Edit-Imputation and Record Linkage*, U.S. Bureau of the Census, Statistical Research Division Report Series, n.2006/7.

[45] Yan S.; Lee D.; Kan M.Y. and Giles C.L. (2007), *Adaptive Sorted Neighborhood Methods for Efficient Record Linkage*, Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries: 185–194. Vancouver; Canada.

[46] Yancey W.E. (2000), *Frequency-dependent probability measures for record linkage*, Statistical Research Report Series, n. RR2000/07. U.S. Bureau of the Census, Washington, DC.

[47] Yancey W.E. (2004b), *An Adaptive String Comparator for Record Linkage*, Statistical Research Division Report Series, n. 2004/02. U.S. Bureau of the Census, Washington, DC.

[48] Yancey W.E. (2005), *Evaluating String Comparator Performance for record linkage*, Statistical Research Division Report Series, n. 2005/05. U.S. Bureau of the Census, Washington, DC.

[49] Yin X.; Han J. and Yu P. (2006), *LinkClus: Eficient clustering via heterogeneous semantic links*, In VLDB'06, pages 427–438. Seoul, Korea, 2006.

[50] Zhu V.J., Overhage M.J., Egg J., Downs S.M., Grannis S.J. (2009), *An Empiric Modification to the Probabilistic Record Linkage Algorithm Using Frequency-Based Weight Scaling*, Journal of the American Medical Informatics Association, 2009, 16: 738–745.

Chapter 2

Literature review update on statistical matching

Summary: Statistical matching is the problem of construction of joint information of variables not jointly observed in one survey, but available in two different sample surveys. A review of the statistical matching methods is in ESSnet ISAD (2008). This state-of-the-art update considers previously neglected areas: the case of sample surveys drawn according to complex survey designs from a finite population; developments of the concept of uncertainty in statistical matching; use of nonparametric procedures.

Keywords: uncertainty in statistical matching, calibration, nonparametric methods

2.1 Introduction

Mauro Scanu

Istituto nazionale di statistica – Istat, Italy

Statistical matching is a data integration problem that consists of the following input:

a) two sample surveys A and B drawn from the same population;

b) an empty intersection of the observed units in the two samples;

c) a non-empty intersection of the sets of variables observed in the two sample surveys.

Item b) precludes the possibility to use record linkage methods for the integration of two data sets.

The objective in a statistical matching problem is the "construction of joint information" of a multivariate variable whose components are not jointly observed in either of the two sources, i.e. at least a pair of variables (Y, Z) is such that Y is observed only in the first and Z in the second sample survey, respectively. The term "construction of joint information" is rather general, and includes the following cases:

- a new data set of microdata, which contains the whole set of variables observed in the two sample surveys;
- an estimate of a parameter of the multivariate variable (*e.g.* contingency tables, correlation coefficients, . . .).

This problem has been analyzed in the deliverables of the ESSnet on Integration of Surveys and Administrative data. The first workpackage (ESSnet – ISAD, 2008) details the state-of-the art on statistical matching in four paragraphs, where it is possible to find information on the following issues:

- a description of the different approaches for tackling statistical matching,
- a definition of uncertainty in statistical matching, due to the absence of sample joint information on Y and Z;
- an illustration of how it is possible to assess the accuracy of a statistical matching method.

In the last years there have been some updates that deserve the attention of those working in a national statistical institute. These updates have been reviewed in the next paragraphs and refer to:

- the case the two sample surveys have been selected according to (possibly different) complex survey designs,
- the assessment of uncertainty in parametric (multinomial and normal distributions) and nonparametric settings,
- the use of nonparametric estimators for tackling the statistical matching problem when the conditional independence assumption holds.

Furthermore, attention is given to those applied areas which have a connection with the statistical matching problem: in this case, differences are highlighted and the solutions developed independently from statistical matching are reviewed.

2.2 Statistical matching when dealing with data from complex survey sampling

Marcello D'Orazio

Istituto nazionale di statistica – Istat, Italy

Statistical matching techniques allow integrating data sources referred to the same target population. In national statistical institutes often these data sources derive from complex sample surveys carried out on the same population. In practice, the available data are collected on a random sample of the target population drawn according to complex survey designs involving stratification, two or more selection stages (e.g. selection of a sample of Municipalities – the Primary sampling Units - and subsequent selection of a subsample of households within each sample Municipality), unequal probability sampling, etc.

When dealing with such data sources, statistical matching techniques can not ignore the sampling design and the different weights associated to each sample unit.

In literature there are relatively few statistical matching methods that tackle explicitly the sampling design and the corresponding sampling weights:

a) Renssen's *calibrations based approach* (Renssen, 1998)

b) Rubin's *file concatenation* (Rubin, 1986)

c) Wu's approach based on *empirical likelihood* methods (2004)

2.2.1 Renssen's calibration approach

This approach is based on a series of calibration steps of the survey weights in order to achieve consistency between estimates computed from the available data sources. Calibration here is intended as a technique to derive new survey weights, as close as possible to the starting ones, that fulfil some constraints set by the researcher.

In the standard framework of statistical matching, the two samples to match are denoted as $A = (X, Y)$ and $B = (X, Z)$, being w_{Ai} $(i = 1, \dots, n_A)$ and w_{Bj} $(j = 1, \dots, n_B)$ the weights associated to the units in A and B, respectively. The first step in the Renssen's procedure consists in calibrating survey weights in A and survey weights in B such that the new weights, $w_{Ai}^{(1)}$

and $w_{Aj}^{(1)}$, applied to the set of the common variables, X, allow to reproduce some known population totals:

$$\hat{t}_{xA} = \sum_{i=1}^{n_A} w_{Ai}^{(1)} x_i = t_{xU}, \quad \hat{t}_{xB} = \sum_{j=1}^{n_B} w_{Bj}^{(1)} x_j = t_{xU}$$

Note that if the population totals t_{xU} are not known for some of the common variables, Renssen suggests first to calibrate weights with respect to the know totals, then, for variables whose population total are not known, an estimate is derived using a combination of the sample estimates (pooled estimate):

$$\hat{t}_{xU} = \lambda \sum_{i=1}^{n_A} w_{Ai}^{(1)} x_i + (1-\lambda) \sum_{j=1}^{n_B} w_{Bj}^{(2)} x_j$$

with $0 \leq \lambda \leq 1$ (λ can be decided according to a subjective reasoning, otherwise some practical rules can be applied, a basic rule consists in setting $\lambda = n_A / (n_A + n_B)$). Hence a new calibration step of the weights $w_{Ai}^{(1)}$ and $w_{Bj}^{(1)}$ is carried out in order to derive new weights, $w_{Ai}^{(2)}$ and $w_{Bj}^{(2)}$, that allow to reproduce the pooled estimate in both A and B.

Note that these initial calibration steps may not be an easy task. For instance, when X is categorical, calibration can be carried out with respect to the marginal distributions, or to the joint distribution; a mixed situation is also allowed (just marginal distributions for some variables, and joint distribution for some other variables). Calibrating when there are both categorical and continuous variables can create some problems. Some authors suggest categorizing continuous variables, in particular when their distribution is skewed.

The calibrated weights $w_{Ai}^{(2)}$ and $w_{Bj}^{(2)}$ can be used to derive estimates from A and B. In particular, in case of categorical variables, under the Conditional Independence Assumption (CIA), the joint distribution $P(Y, Z)$ is estimated by:

$$\hat{P}^{(CIA)}(Y, Z) = \hat{P}^{(A)}(Y|X) \times \hat{P}^{(B)}(Z|X) \times \hat{P}(X)$$

Note that $P(X)$ can be estimated indifferently on A or on B.

In presence of auxiliary information represented by a third data source C, containing all the variables X, Y and Z, two alternative estimates of $P(Y, Z)$ can be derived. These estimates are derived directly from file C after a series of further calibration steps.

The simplest estimate can be obtained under the *incomplete two way stratification.* This approach consists in calibrating the weights w_{Ck} $(k = 1, \ldots, n_C)$ of the units in C by constraining them to reproduce in C the marginal distributions of Y and Z estimated from A and B respectively (after the initial calibrations carried out using the common variables).

A more complex estimate of $P(Y, Z)$ can be obtained under the *synthetic two way stratification.* Roughly speaking it consists in adjusting the $\hat{P}^{(CIA)}(Y, Z)$ using residuals computed in C between predicted and observed values for Y and Z respectively (for more details see Renssen, 1998).

2.2.2 Rubin's file concatenation

The approach proposed by Rubin (1986) consists in concatenating the two data sources A and B. A new data file, the concatenated data set, $F = A \cup B$, will contain $n_F = n_A + n_B$ units with missing values on Y and on Z (given the initial framework there are no units with Y and Z jointly observed). Before using this concatenated file as a single sample selected from the target population it is necessary to associate to each unit a new survey weight that expresses how representative it is.

Following Rubin (1986) the weight for the kth unit in the concatenated file is

$$w_{A\cup B,k} = \frac{1}{\pi_{Ak} + \pi_{Bk}}$$

where π_{Ak} is the probability that the kth unit is included in the sample A and π_{Bk} the probability that the unit is included in the sample B. Obviously, for each unit coming from file A, π_{Ak} is already known, while it necessary to compute the probability that this unit is included in the sample B. This probability can be derived if the sampling design used for selecting B is known and the corresponding design variables are available in A. The same happens, with reversed role, for the units belonging to file B. In other words the inclusion probabilities in the concatenated file can be computed if the design variables for both A and B are known for all the units in the concatenated file. Unfortunately this is not always the case.

It is worth noting that the expression to derive the weights proposed by Rubin is an approximation, the exact formula would be

$$w_{A\cup B,k} = \frac{1}{\pi_{Ak} + \pi_{Bk} - \pi_{A\cap B,k}}$$

being $\pi_{A\cap B,k}$ the probability that the kth unit is included in both A and B. If the two samples A and B are selected independently, it comes out that $\pi_{A\cap B,k} = \pi_{Ak}\,\pi_{Bk}$.

Rubin's approximation assumes $\pi_{A\cap B,k} = 0$. Unfortunately this assumption may not be true, in particular when the sampling designs involved in selecting A and B, allow unequal probability sampling and large units have a higher probability of being included into the sample (PPS sampling). This is likely to happen in sampling of enterprises. In these surveys it is common to stratify units according to some characteristics and to their size, and strata containing the largest units are censused ("take all" strata). Thus if a very large unit is in the take all strata of both the surveys it will have $\pi_{Ak} = \pi_{Bk} = \pi_{A\cap B,k} = 1$.

When it can not be assumed that $\pi_{A\cap B,k} = 0$, these probability has to be computed in order to derive the correct concatenated weights $w_{A\cup B,k}$.

As a further comment Rubin notes that before using the concatenated weights in order to derive the survey estimates, it is preferable to correct them to allow their sum to reproduce the population size N (this is usually a property of a sampling design, such as simple random sampling, stratified random sampling, etc.). This constraint can be fulfilled by using a simple ratio correction:

$$w'_{A\cup B,k} = w_{A\cup B,k}\frac{N}{\sum_{k=1}^{n_A+n_B} w_{A\cup B,k}}.$$

The difficulties in estimating the concatenated probabilities have seriously limited the applicability of the Rubin's approach. Recently, this approach has been successfully used by Ballin *et al.* (2008a and 2008b). These authors suggest a Monte Carlo approach in order to estimate the concatenated probabilities based on the ideas introduced by Fattorini (2006). Fattorini suggests estimating the inclusion probabilities by drawing M independent samples from the target population with the same sampling design. The inclusion probability of the kth unit can be estimated as the fraction out of the samples containing it out of the M independent drawings.

This approach has been applied by Ballin *et al.* (2008a) to estimate the $\pi_{A\cap B,k}$ when concatenating two sample surveys carried out on the Italian farms. These surveys (Farm Structural Survey and the survey on the economic structure of the farms) share the same sampling design (stratified random sampling but with different stratification criteria) but they are not independent. The samples are selected with the objective of reducing as much as possible their overlapping in order to reduce the response burden. Unfortunately, the overlapping can not be avoided at all because in both the surveys the largest farms have probability close or equal to 1 of being included into the sample.

The procedure suggested in Ballin *et al.* (2008a) consists in iterating M times the following procedure

(i) draw a sample from the target population using sampling design of the survey A

(ii) draw a sample target population using sampling design of the survey B;

then compute $X_{A\cap B,k}$ the number of times unit k is included at the same time in both the samples:

$$X_{A\cap B,k} = \sum_{t=1}^{M} I_t \left(k \in s_{A,t} \cap \quad k \in s_{B,t} \right) \quad k = 1, 2, \ldots, N$$

and estimate the probabilities $\pi_{A\cap B,k}$ through the following expression:

$$\tilde{\pi}_{A\cap B,k} = \frac{X_{A\cap B,k} + 1}{M + 1}, \quad k = 1, 2, \ldots, N$$

Obviously, this procedure based on the Monte Carlo experiments can be applied only if the whole sampling frame is available and contains all the design variables used to draw A and B.

Once the concatenated weights have been estimated, it is possible to use them to estimate the marginal/joint distribution of the X variables. On the contrary, methods to deal with missing values are needed to deal with the estimation of $P(X, Y)$, $P(X, Z)$ and $P(Y, Z)$.

For example, a possible approach to estimate $P(X, Y)$ consists in using just the units on which X and Y are fully observed (units coming from file A) weighing each observation using a new weight $w''_{A\cup B,k}$, obtained by calibrating the concatenated weights $w'_{A\cup B,k}$ for units in file A in order to reproduce the marginal distribution (or joint) distribution of the X variables estimated from the whole concatenated file using the weights $w'_{A\cup B,k}$.

An interesting application of Rubin's file concatenation approach is presented in Ballin *et al.* (2008b). Data of the Farm Structural Survey and the survey on the economic structure of the Italian farms are concatenated. The two surveys are not independent given that the samples are selected in order to avoid as much as possible that the same units are included in both the samples, so to reduce the response burden. Although this negative coordination, it is not possible to avoid at all that some farms are selected in both the surveys, this subset of farms are the largest ones that are included in both the samples with certainty (inclusion probability equal to 1). Hence, when concatenating the data of the two surveys there is a small subset of forms for which all the variables are available. This subset can be considered as a source of auxiliary information to better exploit the relationship between Y and Z. Unfortunately this subset is composed only of large farms and hence it a valuable source of auxiliary information limited just to this type of farms. The paper shows some alternative approaches, based on the usage of file concatenation and the corresponding weights, that using estimation methods developed to deal with missing values try to exploit the relationship among Y and Z for the whole population.

2.2.2.1 Empirical likelihood

A recent paper from Wu (2004) explores the usage of *empirical likelihood* (EL) to combine information from two sample surveys. Empirical likelihood methods extend the usage of methods based on likelihood (developed for the case of i.i.d. samples) to the case of complex samples drawn from finite populations (for a comprehensive review see Rao and Wu, 2008)

Empirical likelihood methods have been introduced for the case of simple random sampling. To tackle the case of general unequal probability sampling Chen and Sitter (1999) proposed the usage of a *pseudo empirical likelihood* (PEL) approach that assumes a two stage random mechanism:

1. the finite target population is an i.i.d. sample from a superpopulation and the log-likelihood is:

$$l_U(\mathbf{p}) = \sum_{k=1}^{N} \log(p_k),$$

with

$$p_k = P(y = y_k), \quad p_k > 0, \quad \sum_{k=1}^{N} p_k = 1$$

2. a random sample s is selected from U with unequal inclusion probabilities $\pi_k = P(k \in s)$ $(0 < \pi_k \leq 1)$. An estimate of the previous log-likelihood is obtained by using an estimator derived from the

Horvitz-Thompson estimator of a population total:

$$l_{HT}(\mathbf{p}) = \sum_{k \in s} \frac{1}{\pi_k} \log (p_k)$$

$l_{HT}(\mathbf{p})$ is the *pseudo log empirical likelihood* (PEL)

The theory is slightly different to allow the stratified unequal probability sampling (see Wu and Rao, 2008, for details).

It is worth noting that an improvement of the PEL approach consists in using it as an alternative to the weights calibration, by simply maximizing the PEL with the further constraint that the estimates of p_k reproduce know population totals for a set of auxiliary variables. Wu (2005) provides some algorithms that allow to derive the maximum pseudo empirical likelihood (MPEL) estimates of p_k.

Wu (2004) suggests using the PEL approach to combine data from two independent surveys A and B by maximizing the combined PEL:

$$l_{HT}(\mathbf{p}, \mathbf{q}) = \sum_{i \in A} \frac{1}{\pi_{Ai}} \log (p_i) + \sum_{j \in B} \frac{1}{\pi_{Bj}} \log (q_j)$$

with $p_i = P(y_A = y_{Ai})$ and $q_k = P(y_B = y_{Bj})$.

Unfortunately, Wu's objective is to combine estimates related to Y_A and Y_B (differences among of the same variable observed in different time periods, etc.).

A possible extension of the approach proposed by Wu to the framework of statistical matching is proposed in D'Orazio *et al* (2009). The interesting feature of this proposal is that it allows identifying different solutions of the combined PEL according to the different statistical matching approaches with complex samples presented by Rubin, Renssen and Wu. In order to compare the three statistical matching approaches under the framework of the PEL , D'Orazio *et al* (2009) carried out a limited simulation study (not all the features and all the approaches proposed by Wu are considered). Simulation results show that there is not a best approach. Slightly better results have been obtained under the Renssen schema but it does not outperform the other ones.

2.3 Uncertainty in statistical matching

The study of uncertainty is a recent approach to statistical matching. As stated in the WP1 document of the ESSnet-ISAD (ESSnet-ISAD, 2008), it relies on the analysis of the "uncertainty space", which is the set of all the possible (generally not unique) distributions of the random variables $(Y,Z|X)$ compatible with the available information, *i.e.* observed marginal distribution of (Y,X) and (Z,X).

Example: An example is discussed in Torelli et al (2008). The objective of the study was the joint analysis of variables observed in two agricultural surveys: FADN and FSS. This statistical matching study was focused on the variables:

- Y: Total number of cattle in the farm, with categories 1 (1 or more cattle) or 2 (no cattle)
- Z: Intermediate farm consumption, with categories Z = 1 (up to 4999 Euro), Z = 2 (5000-24999 Euro), Z = 3 (25000-99999 Euro), Z = 4 (100000-499999 Euro), Z = 5 (over 500000 Euro).

The common variables used for this statistical matching application were: Utilized agricultural area in hectares, European size units and Livestock unit coefficient.

The resulting contingency table for the two variables Y and Z is as follows.

	$Y = 1$		$Y = 2$	
	$\underline{\theta}_{1k}$	$\bar{\theta}_{1k}$	$\underline{\theta}_{2k}$	$\bar{\theta}_{1k}$
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

In other words, every cell is composed by a minimum and a maximum frequency. The width of these intervals is called "uncertainty". This kind of uncertainty reflects the fact that the two variables of interest have not been observed jointly. Anyhow, information from the two sample surveys FADN and FSS allows to say that, for instance, the relative frequency of farms with Z=1 and Y=1 is between 3% and 5%.

In the study of uncertainty in statistical matching it is necessary to deal with:

1) the description of the uncertainty space, i.e. the set of parameters admissible given the available data (in the example, the set of distributions for (Y, Z) with frequencies inside the intervals),

2) the evaluation of uncertainty, i.e., roughly speaking "how large" the uncertainty space is (in the example, how large are the intervals).

Some recent papers discuss statistical matching issues that can be referred to the two previous elements.

2.3.1 Study of the distributions of the uncertainty space

Gilula *et al.* (2009) introduce a Bayesian model to use auxiliary information on the association between the dichotomous variables Y and Z to weaken the conditional independence assumption (CIA) of Y and Z given X. This model can be useful to empirically analyse the space of the possible distributions compatible with the data at hand by introducing hypothesis on a parameter having a direct interpretation in practice.

They focus on the conditional probabilities $P(Y{=}i,Z{=}j|X{=}k){=}\ \theta_{ij|k}$ for $i,j = 0,1$ and k=1,...,K. If we assume the conditional independence, the multinomial model conditionally on X is given by $\theta_{ij|k} = \theta_{i|k}\theta_{j|k}$.They write a multinomial model that departs from the CIA by introducing a parameter λdescribing the association between Y and Z given X. The model is

$$P(Y=0,Z=0|X=k)=\theta_{00|k}+a, \quad P(Y=0,Z=1|X=k)=\theta_{01|k}-a,$$

$$P(Y=1,Z=0|X=k)=\theta_{10|k}-a, \quad P(Y=1,Z=1|X=k)=\theta_{11|k}+a,$$

where $a{=}\lambda\min\{\theta_{01|k},\theta_{10|k}\}$if $\lambda\in(0,1]$, *i.e.* there is a positive association between Y and Z given X, and $a{=}\lambda\min\{\theta_{00|k},\theta_{11|k}\}$if $\lambda\in[-1,0)$, *i.e.* there is a negative association between Y and Z given X.

They finally suggest $p(\lambda)\propto\frac{1}{(1+|\lambda|)^{\alpha}}$ as prior distribution for λ, that is a symmetric distribution centred on zero with the parameter α defining how informative it is. They also suggest inferring on λ conditional on the values of $\hat{\vartheta}_{i|k}$and $\hat{\vartheta}_{j|k}$that are the estimates of the conditional distributions of $Y|X$ and $Z|X$.

2.3.2 Assessing the uncertainty

In this context there are some updates with respect to the evaluation of uncertainty of statistical matching in the case of multivariate normal distributions and in a non parametric setting.

Uncertainty in the multinormal case

Raessler and Kiesl (2009) study uncertainty when the r.v.s Y, Z and X are multinormally distributed.

As remarked in D'Orazio *et al.* (2006), uncertainty in the case of multivariate normal distributions is related to the non estimability of the correlation coefficients ρ_{yz}. Given an estimate of the parameters ρ_{yx} and ρ_{zx} (that can be obtained by the available information), the set of all possible ρ_{yz} compatible with those values denotes the uncertainty of the matching process. The compatible ρ_{yz} are all those values in [-1,1] such that the resulting correlation matrix is definite positive. All feasible correlations form an ellipsoid.

A measurement of the uncertainty is given by the volume of the ellipsoid formed by all the feasible correlations, and, as Raessler and Kiesl (2009) show, it is proportional to the product of the length of its semi-axes given by $1/\sqrt{\lambda_i}$, where λ_i is the i-th (i=1,...,n) eigenvalue of the matrix C=$(1 - \rho_{zx}\rho_{xx}^{-1}\rho_{xz})^{-1}(\rho_{yy} - \rho_{yx}\rho_{xx}^{-1}\rho_{xy})^{-1}$

Uncertainty in the non parametric setting

Conti *et al.* (2009) deal with the problem of evaluating the uncertainty of a statistical matching problem in a non parametric setting. Uncertainty in this setting is still described by the class of the models compatible with the information arising from the data at hand, but they are not identified by a finite number of parameters. This implies that the description of uncertainty in this context is considerably more difficult.

The natural way to describe the class of distributions consists in using the Fréchet class. We recall that, a measure of uncertainty is nothing more than a suitable functional that quantifies "how large" is such a class. Then, conditionally on X, we have a set of plausible statistical models, namely the Fréchet class of all distribution functions $H(z,\, y|x)$ compatible with the univariate d.f.s $G(z|x)$, $F(y|x)$ that can be estimated from the data.

For every $(z,\, y)$, the pair of inequalities

$$L^x\left(F\left(y|x\right),G\left(z|x\right)\right) \leq H\left(z,y|x\right) \leq U^x\left(F\left(y|x\right),G\left(z|x\right)\right)$$

holds, where the bounds

$$L^x\left(F\left(y|x\right),G\left(z|x\right)\right) = max\left\{G\left(z|x\right)+F\left(y|x\right)-1,0\right\}$$

and

$$U^x\left(F\left(y|x\right),G\left(z|x\right)\right) = min\left\{G\left(z|x\right),F\left(y|x\right)\right\}$$

are themselves joint d.f.s with margins $G(z|x)$ and $F(y|x)$.

The set of d.f.s

$$H^x = \left\{H\left(z,y|x\right) : L^x\left(F\left(y|x\right),G\left(z|x\right)\right) \leq H\left(z,y|x\right) \leq U^x\left(F\left(y|x\right),G\left(z|x\right)\right)\right\}$$

is the Fréchet class of marginal d.f.s $G(z|x)$ and $F(y|x)$.

Taking the expectation with respect to the distribution of X, we obtain the unconditional Fréchet class

$$H = \{H(z,y) : E_x\left[L^x\left(F(y|x), G(z|x)\right)\right] \leq H(z,y) \leq E_x\left[U^x\left(F(y|x), G(z|x)\right)\right]\}$$

As a consequence of Jensen's inequality, the previous Fréchet class is narrower than the "naive" Fréchet class

$$\{H(z,y) : max\left(F(y) + G(z) - 1, 0\right) \leq H(z,y) \leq min\left(F(y), G(z)\right)\}$$

that does not use the common information X available on the two datasets A and B.

Given x, uncertainty can be measured by the following difference

$$\Delta^x(F,G) = \int \left[U^x\left(F(y|x), G(z|x)\right) - L^x\left(F(y|x), G(z|x)\right)\right] dF(y|x)\, dG(z|x)$$

An overall measure can be given by

$$\Delta(F,G) = \\ \int \left\{ \int \left[U^x\left(F(y|x), G(z|x)\right) - L^x\left(F(y|x), G(z|x)\right)\right] dF(y|x)\, dG(z|x)\, dQ(x) \right\} = \\ E_x\left[\Delta^x(F,G)\right]$$

where $Q(x)$ is the marginal distribution of X.

It is possible to estimate the extrema of the distributions, as well as Δ by means of the corresponding empirical distribution functions.

2.4 Nonparametric procedures for statistical matching

Mauro Scanu

Istituto nazionale di statistica – Istat, Italy

The first statistical matching applications consisted of the application of the distance hot-deck imputation method, where distance was based on the common variables available in the two files. For instance, Okner (1972) imputed the 1967 Survey of Economic Opportunity with individual income tax returns available on the 1966 Internal Revenue Income Tax. Statistical matching

Hot-deck methods were the main tool for statistical matching for decades. Singh *et al.* (1990) give an overview of methods based on the hot-deck family that can be applied in a statistical matching framework. This approach can be considered as *non-parametric*, because it is not needed any parametric distribution function for the r.v.'s X, Y, Z (as the normal distribution, used in Kadane, 1978, and in many other papers on statistical matching). Indeed, distance hot-deck can be formally described as belonging to at least two classes of nonparametric procedures (in the following, random hot-deck is described as a special case of the kNN random hot-deck method, as well as of the nonparametric regression method based on a kNN estimate of the nonparametric regression function, when k=1).

Assume that Y and Z are independent given X (the conditional independence assumption). Statistical matching can be tackled by imputing a value of the missing r.v. Z in each record in the data set A. Given that under the conditional independence assumption, Y does not have any information on Z when X is known, attention can be restricted to the following data sets:

$$\left(x_a^A\right), a = 1, \ldots, n_A,$$
$$\left(x_b^B, z_b^B\right), b = 1, \ldots, n_B,$$

for samples A and B respectively.

A family of nonparametric imputation techniques can be described as follows. For every x_a^A in A, let $b(a)$ =($b_1(a)$,..., $b_k(a)$) be the labels of its k donor records in B, on the basis of the n_B observations x_b^B, b = 1, . . . , n_B, and let $\mathbf{x}_{b(a)}^B$ be the corresponding vector $\left(x_{b_1(a)}^B, \ldots, x_{b_k(a)}^B\right)$. Next, the corresponding z-values $\mathbf{z}_{b(a)}^B = \left(z_{b_1(a)}^B, \ldots, z_{b_k(a)}^B\right)$ are considered. Finally, the missing value z_a^A is imputed by $\tilde{z}_a = g\left(\mathbf{z}_{b(a)}^B\right)$, $g(\cdot)$ being an appropriate function. Examples are the arithmetic mean of $\mathbf{z}_{b_j(a)}^B$, j=1,...,k, their median, or a randomly chosen value from $\mathbf{z}_{b_j(a)}^B$, j=1,...,k.

2.4.1 Choosing the donor records

By far, the most common selection technique of the k donors of a record a in A consists in taking its k nearest neighbours, $k \geq 1$, *i.e.* those record in

B with labels ($b_1(a)$,..., $b_k(a)$) such that:

$$d\left(x_a^A, x_{b_j(a)}^B\right) \leq d\left(x_a^A, x_{b_{j+1}(a)}^B\right), j = 1, \ldots, k,$$

and

$$d\left(x_a^A, x_{b_j(a)}^B\right) \leq d\left(x_a^A, x_b^B\right), \quad \text{for any} \quad b \notin \left(b_1(a), ..., b_k(a)\right),$$

where $d(.,.)$ is the Euclidean distance. Goel and Ramalingham (1989) suggest the use of the Mahalanobis distance, anyway the Mahalobis distance, as well as other distances, has been seldom considered, see D'Orazio *et al.* (2006) for an overview that includes also the case of a multivariate X.

2.4.2 kNN random hot deck

Once the k nearest neighbours of x_a^A and $x_{b(a)}^B$ are obtained, one could impute the missing z_a^A by randomly choosing a label $\tilde{b}\,(a)$ among $b_j(a)$, j = 1,. . . ,k, and in taking imputed values

$$\tilde{z}_a = z_{\tilde{b}(a)}^B, \quad a = 1, \ldots, n_a.$$

A generalized version of this approach is in Aluja-Banet *et al.* (2007). A value is taken at random assuming different probabilities of selection for the donor records: observations close to x_a^A have higher probabilities than those further away.

<u>Note:</u> When $k = 1$, this imputation method reduces to *distance hot deck*. Imputed data are obtained as:

$$\tilde{z}_a = z_{b_1(a)}^B, \quad a = 1, \ldots, n_a.$$

In other words, each record in A is matched with the closest record in B.

<u>Note:</u> This family of methods can be modified in order to avoid that a donor is selected more than once. The idea is that, if n_A=n_B and each donor is selected only once, the marginal distribution of Z is perfectly reproduced also in file A.

2.4.3 Methods based on nonparametric regression function

Let X and Z be linked by a nonparametric regression function:

$$Z = m(X) + \varepsilon,$$

where $m(x) = E(Z|X = x)$ is the regression function of Z given X, and $\varepsilon = Z - m(X)$ is the error term, such that $E(\varepsilon|X = x) = 0$ for every x. For the sake of simplicity, in the sequel we will further assume that the errors are homoscedastic, *i.e.* $E(\varepsilon|X = x) = \sigma^2$ independent of x. A simple idea to impute Z in sample A could consist of the following steps.

1. Estimate the regression function $m(x)$ by the sample B. From now on, such an estimator will be denoted by $\hat{m}^B(x)$.

2. Let $\hat{\varepsilon}_b^B = z_b^B - \hat{m}^B(x_b^B)$, b=1,...,n_b, be the corresponding residuals in B.

3. Impute the missing z_a^A by $\tilde{z}_a^A = \hat{m}^B(x_a^A) + \tilde{\varepsilon}^B$, a=1,...,n_a, where $\tilde{\varepsilon}^B$ is drawn at random among $\hat{\varepsilon}_1^B$, ..., $\hat{\varepsilon}_{n_b}^B$

The rationale of the previous steps is simple: at first, estimate the regression function and compute plausible values of the errors ε, then use these pieces of information for imputing Z in A. This approach belongs to the set of imputation methods known as *stochastic*.

When the residuals $\tilde{\varepsilon}^B$ are omitted, step 3 is substituted by:

3. Impute the missing z_a^A by $\tilde{z}_a^A = \hat{m}^B(x_a^A)$, a=1,...,n_a.

This imputation method is *deterministic*. The possibilities of estimation of the nonparametric regression function $m(.)$ are diverse. Two of them are sketched in the following lines. Take in mind that the efficiency of the estimators of the nonparametric regression function deteriorates when X or Z are multivariate.

2.4.4 kNN methods

The kNN imputation method consists in estimating the nonparametric regression function $m(.)$ by the kNN estimator. Formally, the regression function $m(.)$ is estimated by the average of Z corresponding to the k nearest neighbours of x. When $x = x_a^A$:

$$\hat{m}^B\left(x_a^A\right) = \frac{1}{k}\sum_{j=1}^{k} z_{b_j(a)}^B, \quad a = 1, \ldots, n_A..$$

This estimate can be used for both the deterministic and the stochastic imputation method, as defined above.

The key point in using the kNN estimator is the choice of the parameter k that determines the amount of smoothing of z_b^Bs data. It plays a role similar to the bandwidth for kernel smoothers.

Note: It can be shown (Paass, 1985; Cohen, 1991) that distance hot deck is also equivalent to impute missing data through the kNN method, with k=1. Such a procedure seems to be at first sight a deterministic technique, because residuals $\hat{\varepsilon}_1^B$, ..., $\hat{\varepsilon}_{n_b}^B$ are null whenever x is equal to any of the n_B values x_b^B observed in B. As a matter of fact this method imputes at the same time both the regression function and the residual.

2.4.5 Local polynomial regression

A different estimator of the nonparametric regression function $m(x)$ is the *local polynomial estimator* (Fan and Gijbels, 1996). Suppose that $m(x)$ possesses p+1 derivatives, and denote by $m^{(j)}(x)$ its jth derivative, j=1,. . . ,p+1. The nonparametric regression function is approximated locally by a polynomial of order p:

$$m(t) \approx m(x) + m^{(1)}(x)(t-x) + \ldots + \frac{1}{p!}m^{(p)}(x)(t-x)^p =$$
$$\beta_0 + \beta_1(t-x) + \ldots + \beta_p(t-x)^p.$$

The polynomial is local because the parameters $\beta_0, \ldots, \beta_p$ depend on x. These parameters can be estimated by the weighted least squares method, *i.e.* can be found minimizing the quantity:

$$\sum_{b=1}^{n_B}\left(z_b^B - \sum_{j=0}^{p}\beta_j\left(x_b^B - x\right)^j\right)^2 K_h\left(x_b^B - x\right)$$

where $K_h(t) = h^{-1}K(t/h)$, $K(.)$ is a nonnegative function and h is a smoothing parameter (bandwidth) determining the size of the neighbourhood of x used in estimating $m(x)$.

Local polynomial estimators have been proved as particularly useful and efficient as well. Their merits are thoroughly discussed in Fan and Gijbels (1996). In particular, when p=0 the local polynomial estimator reduces to the Nadaraya-Watson estimator. When p=1, the local polynomial estimator reduces to the *local linear estimator*. This has several advantages if compared to the Nadaraya-Watson estimator, which can be extremely inefficient when x is close at the extremes of its range and needs to assume that $V\left(\varepsilon\left|X=x\right.\right)$ is independent of x.

2.4.6 Matching noise

Under the conditional independence assumption, statistical matching accuracy can be evaluated by the *matching noise*, i.e. the distance between the actual distribution of Z given X and the distribution of the r.v. that generates the imputations $\tilde{Z}$ given X. If these two distributions are "similar", the imputed data set A can be representative of $(X,\ Z)$ and, under the conditional independence assumption, of $(X,\ Y,\ Z)$. Preliminary evaluations are in Paass (1985).

In Marella *et al.* (2008) the matching noise that affects kNN method is determined. It is proved that $\mathbf{X}^{B}_{\mathbf{b}(a)}\left|X^{A}_{a}\right.$ converges in distribution to a k-dimensional vector whose elements are equal to X^{A}_{a}. Hence, stochastic kNN, distance hot deck and selection of a random element from the k nearest neighbours tend asymptotically to be matching noise free, while deterministic kNN (8) is unavoidably biased.

Conti *et al.* (2008) proves similar results also for imputations based on the local linear regression estimator. Roughly speaking, since, as n_B increases,

1. the estimated regression function $\hat{m}$ becomes closer and closer to the population regression function;
2. the empirical distribution of the residuals $\hat{\varepsilon}_b$ tends to be closer and closer to the distribution of the (population) errors ε_b;

then the distribution of $\tilde{z}_a$ becomes closer and closer to the distribution of z_a. In other words, stochastic imputations based on the local linear regression estimator are asymptotically matching noise free.

Both papers compare these imputation methods by simulation, using data generating models characterized by non-normal distributions as well as nonlinear regression functions. The imputation method based on the local linear regression estimator of the nonparametric regression function seems to be the best choice. Anyway, it is remarkable that distance hot-deck imputations are quite efficient in all the simulation scenarios, without resorting to computationally cumbersome procedures.

References on statistical matching

[1] Aluja-Banet T., Daunis-i-Estadella J., Pellicer D. (2007), *GRAFT, a complete system for data fusion*, Journal of Computational Statistics and Data Analysis, **52**, 635–649.

[2] Ballin M., D'Orazio M., Di Zio M., Scanu M., Torelli N. (2008a), *File concatenation of survey data: a computer intensive approach to sampling weights estimation*, Rivista di Statistica Ufficiale, N. 2–3, pp. 5–12.

[3] Ballin M., Di Zio M., D'Orazio M., Scanu M., Torelli N. (2008b), *Statistical Matching of Two Surveys with a Common Subset*, Working Paper, N. 124, Universitá degli Studi di Trieste.

[4] Chen, J., and Sitter, R.R. (1999), *A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys*, Statistica Sinica, 9, 385–406.

[5] Conti P.L., Marella D., Scanu M. (2008), *Evaluation of matching noise for imputation techniques based on the local linear regression estimator*, Computational Statistics and Data Analysis, **53**, 354–365.

[6] Cohen M.L. (1991), *Statistical matching and microsimulation models*, Improving Information for Social Policy Decisions, the Use of Microsimulation Modeling. Technical Papers, vol. II. National Academy Press.

[7] Conti P. L., Di Zio M., Marella D., Scanu M. (2009), *Uncertainty analysis in statistical matching*, First Italian Conference on Survey Methodology (ITACOSM09), Siena 10–12 June 2009.

[8] D'Orazio M., Di Zio M, Scanu M. (2006), *Statistical Matching: Theory and Practice*. Wiley, Chichester.

[9] D'Orazio M., Di Zio M., Scanu M. (2009), *Uncertainty intervals for nonidentifiable parameters in statistical matching*, 57^{th} Session of the International Statistical Institute, 16–22 August 2009, Durban.

[10] ESSnet ISAD (2008), *Report of WP1. State of the art on statistical methodologies for integration of surveys and administrative data*, http://cenex-isad.istat.it.

[11] Fan J., Gijbels I. (1996), *Local Polynomial Modelling and its Applications*. Chapman & Hall, London.

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

[13] Gilula, Z, McCulloch, R.E., Rossi, P.E. (2006), *A direct approach to data fusion*, Journal of Marketing Research, **43**, 73–83.

[14] Goel P. K., Ramalingham T. (1989), *The Matching Methodology: Some Statistical Properties*. Springer Verlag, New York.

[15] Kadane J. B. (1978), *Some Statistical Problems in Merging Data Files*, in Compendium of Tax Research, Department of Treasury, U.S. Government Printing Office, Washington D.C., 159–179 (Reprinted in 2001, Journal of Official Statistics, **17**, 423–433).

[16] Marella D., Scanu M., Conti P.L. (2008), *On the matching noise of some nonparametric imputation procedures*, Statistics and Probability Letters, **78**, 1593–1600.

[17] Okner B.A (1972), *Constructing a new data base from existing microdata sets: the 1966 merge file*, Annals of Economic and Social Measurements, **1**, 343–345.

[18] Paass G. (1985), *Statistical record linkage methodology, state of the art and future prospects*, Bulletin of the International Statistical Institute, Proceedings of the 45th Session, LI, Book 2.

[19] Raessler, S., Kiesl, H. (2009), *How useful are uncertainty bounds? Some recent theory with an application to Rubin's causal model*, 57th Session of the International Statistical Institute, Durban (South Africa), 16–22 August 2009.

[20] Rao, J.N.K, Wu, C. (2008), *Empirical Likelihood Methods*, In D. Pfeffermann and C.R. Rao (eds.) Handbook of Statistics, Vol. 29. Sample Surveys: Theory, Methods and Inference, pp. 189–207.

[21] Renssen, R. H. (1998), *Use of Statistical Matching Techniques in Calibration Estimation*, Survey Methodology, 24, 171–183.

[22] Rubin, D. B. (1986), *Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations*, Journal of Business and Economic Statistics, 4, 87–94.

[23] Singh A. C., Mantel H., Kinack M. and Rowe G. (1990), *On methods of statistical matching with and without auxiliary information*, Technical Report, DDMD-90-016, Statistics Canada.

[24] Wu, C. (2004), *Combining information from multiple surveys through the empirical likelihood method*, The Canadian Journal of Statistics, 32, 112.

[25] Wu, C. (2005), *Algorithms and R codes for the pseudo empirical likelihood method in survey sampling*, Survey Methodology, 31, 239–243.

Chapter 3

Connections between ecological inference and statistical matching

Summary: Ecological inference and statistical matching are problems with many similarities. This section compares the two problems, describes the models and methods used in ecological inference, highlighting the differences with those used in statistical matching.

Keywords: ecological regression, tomography line

3.1 Introduction

Marco Di Zio

Istituto nazionale di statistica – Istat, Italy

Ecological inference is the process of using aggregate or macrolevel data to draw conclusions at the individual level, where no individual level data are available (King, 1997; Hudson et al. 2010). The usual framework consists of a set of 2x2 tables related to two binary variables, in which only the margins are observed (macrolevel data) and the goal is to examine the association between the two variables (microlevel data). A motivating example is the following (King, 1997).

Example: Let the following data

1. the proportion of voting-age population who are black (p_i),
2. the proportion of voting-age population Turning out to vote (q_i),
3. the number of people of voting-age (N_i)

be known for all the precints i (electorate districts) of a certain county. The goal is to estimate the proportion of voting-age blacks who vote (β_{bi}) and the proportion of voting-age whites who vote (β_{wi}). This problem is usually represented through a table, for each precint i.

Race of voting-age person	Voting decision		
	Vote	No Vote	
Black	β_{bi}	$1-\beta_{bi}$	p_i
White	β_{wi}	$1-\beta_{wi}$	$1-p_i$
	q_i	$1-q_i$	

Hence, the goal is to infer the cells of the tables (microlevel) through knowledge of the marginals for each precinct i, *i.e.* to infer the proportion of voting-age blacks who vote β_b and the proportion of voting-age whites who vote β_w referred to all the population.

It is apparent that there is a close connection with the statistical matching problem when the objective is macro integration.

Following the former example, let the random variable X (assuming 64 categories) denote the precincts, and let the dichotomous variables Y be the ”Race of voting-age person” and Z be the ”Voting Decision”. With this formalization we are exactly in the statistical matching framework, with a slight reparametrization. Instead of making inference on (generally speaking) the distribution of $(Y,\ Z)$ given X (or the joint probability distribution of Y and Z) by means of the knowledge of the conditional distributions $Y|X{=}i$, $Z|X{=}i$, the objective is to make inference on the conditional distributions of Z given (X, Y) (for this reason, the probabilities inside the table sum to 1 by row).

The fundamental problem is that many different relationships at the individual level can generate the same observation at the aggregate level: ecological inference experts refer to this problem as the *ecological fallacy*. As a matter of fact, ecological fallacy corresponds to the *uncertainty* in statistical matching (see D'Orazio, 2006). This is entwined with the problem of inferring a joint distribution through the knowledge of only the marginal (conditional) distributions. As in statistical matching, different models are introduced to reach a single point estimate of the unknown parameters, but also in ecological inference this can be achieved only by introducing further information, for instance in the form of hypotheses, on the relationships of the variables Y and Z that fill the gap of knowledge. In the following some models used for ecological inference and their connection with those used in statistical matching are shown.

3.2 Statistical models in ecological inference

Mauro Scanu

Istituto nazionale di statistica – Istat, Italy

Ecological regression is essentially a problem of inference with partial information. Missing information is that on the relationship between the target variables. The first approaches to ecological inference focused on the definition of those models that makes the problem identifiable, *i.e.* estimable, for the data at hand. A fundamental equation in this setting, following notation of the previous paragraph, is synthesized by the ecological regression equation, known also as tomography line:

$$q_i = \beta_{bi} p_i + \beta_{wi}(1 - p_i).$$

The *Goodman ecological regression model* (Goodman 1953) is the first approach in this sense. It consists in assuming that Z and X are independent given Y. In other terms, the probabilities $\beta_{bi} = \beta_b$ and $\beta_{wi} = \beta_w$ do not change in the different precincts. As a matter of fact, this model is a different conditional independence assumption than the one usually used in statistical matching problems. Anyhow, this model can be easily estimated in an ecological regression model through the tomography line, when Y and Z are dichotomous.

The traditional conditional independence assumption of Y and Z given X is assumed in Freedman *et al* (1991). As in statistical matching, this assumption does not need any restriction on the number of states for the variables Y and Z.

3.3 Uncertainty in ecological regression

Mauro Scanu

Istituto nazionale di statistica – Istat, Italy

The identifiable models described in the previous section have been largely criticized. This problem can be overcome by analyzing all the models compatible with the data at hand. Chambers and Steel (2001) propose to describe discrepancies from the conditional independence model by means of the following relationship:

$$\beta_{wi} = \gamma q_i, \quad \text{for all the precincts} \quad i.$$

Again, this approach can be usefully applied only when Z and Y are dichotomous, so the relationship between the not jointly observed variables is explained by only one parameter. Fréchet bounds can determine lower and upper bounds for γ in each precinct. Hence γ should be between the maximum of the lower bounds and the minimum of the upper bounds for all the precincts.

This interval would correspond to the uncertainty for the problem at hand. Chambers and Steel go further, suggesting to use the midpoint of the interval determined before as an estimate for γ. The idea is essentially Bayesian: the authors assume that all the admissible models given the available data are equally probable, corresponding to a uniform distribution on γ in the interval. The midpoint corresponds to the average γ with respect to this uniform distribution.

Note – As in every non identifiable model, the prior distribution for γ is not updated in a posterior by the available data (see Rubin 1974). Furthermore this approach would not be Bayesian in statistical matching. In fact, the marginal distributions for Y and Z in a statistical matching problem are usually determined by samples. Hence, the state space for the prior distribution on γ would be data dependent.

King (1997) focuses the estimation of the parameters β_{bi} and β_{wi} on the tomography line. In order to get an estimate of β_{bi} and β_{wi}, King:

1. draws values of the pair (β_{bi}, β_{wi}) from a truncated bivariate normal distribution;

2. determines the intersection between the truncated bivariate contour line corresponding to the drawn value and the tomography line;

3. estimates (β_{bi}, β_{wi}) averaging the values obtained in step 2.

As in Chambers and Steel, there is an exploration of the admissible values for the parameters of interest (β_{bi}, β_{wi}) and a final averaging of these admissible values in order to get a unique estimate. King (1997) suggests also modifications that allow a nonparametric exploration of the parameter space (instead of using the truncated bivariate normal distribution).

Bayesian approaches have been reviewed and proposed in Wakefield (2004).

References on ecological inference

[1] Chambers, R.L., Steel, D. G. (2001), *Simple methods for ecological inference in 2x2 tables*, Journal of the Royal Statistical Society, A, Volume 164, 175–192.

[2] Freedman, D., Klein, S., Sacks, J., Smyth, C.A., Everett, C.G. (1991), *Ecological regression and voting rights*, Evaln. Rev., Volume 15, 673–711.

[3] Goodman, L. (1953), *Ecological regression and behavior of individuals*, American Sociological Review, Volume 18, 663–666.

[4] Hudson, I.L., Moore, L., Beh, E.J., Steel D.G. (2010), *Ecological inference techniques: an empirical evaluation using data describing gender and voter turnout at New Zealand elections*, 1893–1919, Journal of the Royal Statistical Society, A, Volume 173, 185–213.

[5] King, G. (1997), *A solution to Ecological Inference Problem*, Princeton: Princeton University Press.

[6] Rubin, D.B. (1974), *Characterizing the estimation of parameters in incomplete-data problems*, Journal of the American Statistical Association, Volume 69, 467–474.

[7] Wakefield, J. (2004), *Ecological inference for 2x2 tables*, Journal of the Royal Statistical Society, A, Volume 167, 385–445.

Chapter 4

Literature review update on data integration methods in statistical disclosure control

Daniela Ichim

Istituto nazionale di statistica – Istat, Italy

Summary: In this section two main links between data integration settings and statistical disclosure control are briefly introduced. First, the relationship between uncertainty and disclosure risk for contingency tables dissemination is described. It is supposed that the contingency tables are derived from the cross-classification of some categorical variables observed on the entire population. Second, the usage of record linkage methodologies for microdata dissemination is illustrated.

Keywords: disclosure control, microdata dissemination, contingency tables, marginal distributions

4.1 Contingency table dissemination – Statistical disclosure control and statistical matching

Among the statistical information disseminated by National Statistical Institutes, tabular data have been the oldest and most well-known. Given the regulations pertaining to the privacy of respondents, the general aim is to release as much information as possible.

The safest way to protect the confidentiality of respondents is to release no data at all. This option is generally discarded due to its completely absent utility. One of the approaches followed by the National Statistical Institutes is to constrain some data utility measure and then to evaluate the risk of disclosure; consequently, the decision to disseminate is taken or not.

Suppose now that a National Statistical Institute would like to disseminate information about a k-way contingency table, i.e. frequency counts (non-negative integers) derived from the cross-classification of k categorical variables. Additionally, suppose that the release of the full k-way contingency table could not be considered a valid alternative, due to the high risk of disclosure. The release of marginal tables could be a solution since the frequency counts of the initial k-way table might not be exactly known. The idea is that this uncertainty on the frequency counts might reduce the disclosure risk. When releasing contingency tables, the risk of disclosure is not related to the frequency counts themselves, but rather from the sensitivity of (some) categories of the cross-classifying variables. Generally speaking, if there is no uncertainty on the frequency values, information about some respondents/units might become public knowledge, in contrast with statistical confidentiality laws. Since it was assumed that the categorical variables were observed on the entire population, both low and high frequencies might favour a confidentiality breach. On the contrary, if the observed units were a sample of the population, the sampling fraction could itself improve protection of the confidentiality of respondents. This latter case is not further addressed in this section.

Contingency tables are generally used to study associations between variables; log-linear models are a common tool to perform such analyses. For log-linear models, it is well-known, see Agresti (2002), that some possibly multivariate marginals are minimal sufficient statistics[1]. Consequently, the release of relevant marginal tables could be as useful[2] as the release of the full k-way contingency table.

It follows that, when data utility is measured **only** in terms of log-linear models, the release of relevant marginal tables could be sufficient. From the risk of disclosure point of view, for each cell entry in the initial table, the uncertainty induced by the release of marginal tables should be evaluated. The constraints given by the released marginal tables induce upper and lower bounds on the interior cells of the initial table. These bounds

[1] Minimal sufficient statistics are helpful for deriving maximum likelihood estimations.

[2] Depending on the log-linear model.

(or feasibility intervals) could be obtained by solving the corresponding NP-hard linear programming problems. If we consider that these NP-hard linear programming problem should be an Integer Programming one (admitting no fractional solutions), the situation is even more complicated. Due to the high computational complexity and burden, other solutions should be used in practical applications.

The similarity between the uncertainty problem in statistical matching and disclosure risk evaluation should be obvious. In statistical matching, one has the marginal distributions (Y, X) and (X, Z) and wants to make inferences (i.e. characterize better) the distribution (Y, X, Z). Using the statistical disclosure control terminology, the same problem may be stated as: evaluate how much information on the distribution (Y, X, Z) may be derived from its marginals (Y, X) and (X, Z).

In statistical disclosure control field, the uncertainty problem was approached by answering the following questions:

a) How many tables are compatible with the given fixed marginal distributions (Y, X) and(X, Z)?

b) Given the fixed marginal tables, how to compute feasible bounds on the cell entries in the initial (full) k-way contingency table? How to compute sharp[3] bounds on those cell entries?

a) For categorical variables, answers to the first question were found by investigating the space of tables with given marginals. Since the initial contingency table and its (given) marginal tables are linked by means of linear relationships, the space of tables with given fixed marginal tables is a polytope. The number of tables in the polytope is strictly related to the disclosure risk. If the number of tables with given marginal tables is extremely reduced, there is a high probability that the possible intruder might very accurately "guess" the initial confidential contingency table. The number of tables belonging to a given polytope depends on the number of categories of the cross-classifying variables and on the values of the marginal tables as well. Due to the mathematical and computational complexity, the problem was approached only for 2-way dimensional tables. For example, in Good (1976) or Gail and Mantel (1977), both the exact enumeration and a normal approximation were proposed. Although some interesting generalisations are illustrated, a sufficiently accurate approximation proposed in Gail and Mantel (1977) is for the number of $r \times 2$ tables:

[3] The tightest possibile.

$$N\left(c, m_1, \ldots, m_r\right) = \left[\prod_i^r \left(m_i + 1\right)\right] \left(2\pi\sigma^2\right)^{-1/2} \exp\left(-\frac{\left(c-\mu\right)^2}{2\sigma^2}\right),$$

where $\mu = \sum m_i/2$ and $\sigma^2 = \sum m_i \left(m_i + 2\right)/12$, c is the column total and $m_1, \ldots, m_r$are the row totals. Some recently proposed approximation algorithms and asymptotic estimations may be found in Barvinok (2009) and Barvinok *et al.* (2010). A software tool for counting the number of lattice points inside convex polytopes may be found at http://www.math.ucdavis.e\discretionary{-}{}{}du/~{}latte/.

b) Besides the fact that generalisations to k-way ($k > 2$) contingency tables would be needed, one important consequence of the previous approach is that it does not consider the issue of overlapping marginal tables. Consequently, its applicability to statistical disclosure control is limited. Using graphical models theory, Dobra and Fienberg developed a framework for computing sharp bounds in many cases. Fréchet sharp bounds for the cell entries in an $I \times J$contingency table with entries $\{n_{ij}\}$, row margins $\{n_{i+}\}$, column margins $\{n_{+j}\}$ and n_{++} grand total, are given by: $\min\{n_{i+}, n_{+j}\} \geq n_{ij} \geq \max\{0, n_{i+} + n_{+j} - n_{++}\}$, see Fréchet (1940). Note that this formula is only a particular case of the formula given in section 4.2 In Fienberg (1999) it was proposed to use $\{n_{i+}\}$ and $\{n_{+j}\}$ to simultaneously measure the risk of disclosure and data utility. First, the risk of disclosure is evaluated by means of the previously presented Fréchet bounds. Second, since the same marginals $\{n_{i+}\}$ and $\{n_{+j}\}$ are minimal sufficient statistics for log-linear models, the data utility constraint a-priori defined by the National Statistical Institute is satisfied. The original Fréchet sharp bounds hold only for two (sets) of minimal sufficient statistics. By induction on the number of sufficient statistics, Dobra and Fienberg (2000, 2001, 2003) generalized the Fréchet bounds formula for decomposable log-linear models: *the upper bounds for the cell entries in the initial table are the minimum of relevant margins, while the lower bounds are the maximum of zero, or sum of the relevant margins minus the separators*. These bounds are sharp in the sense that they are the tightest possible bounds given the marginals.

When the log-linear model associated with the released set of marginals is not decomposable, the same strategy, i.e. decomposition of graphs by means of complete separators, was employed to reduce the computational effort needed to determine the tightest bounds. An independence graph that admits a proper decomposition but is not necessarily decomposable is said to be reducible and a reducible log-linear model is one for which the corresponding minimal sufficient statistcs are marginals that characterize the components of a reducible independence graph. In Dobra and Fienberg (2000, 2001, 2003), it was proved that *when the released set of marginals is the set of minimal sufficient statistics of a reducible log-linear model, then the upper bounds for the cell entries in the initial table are the minimum of upper bounds of relevant components, while the lower bounds are the maximum of zero, or sum of the lower bounds of relevant components minus the separators.*

When the independence graph corresponding to a set of released marginals is not reducible, the Fréchet bounds are not sharp and an iterative procedure should be applied. An example of such iterative procedure is the Generalized Shuttle Algorithm (GSA) developed by Dobra and Fienberg. Let Tbe the set containing all cells in the initial table, formed by collapsing the cells in the initial table in every possible way. The blocks to be joined have to be composed from the same categories in $k-1$dimensions and they are also required not to share any categories in the remaining dimension. Noting that the upper and lower bounds are interlinked, i.e. bounds for some cells induce bounds for some other cells, Dobra and Fienberg equivalently stated the bounds problem: "Find the upper and lower bounds for the cells in T given that the upper and lower bounds for some cells in $T_0 \subset T$ are known". Here we give a very brief description of the GSA; more details may be found in Dobra and Fienberg (2000, 2001, 2003). Let t_1 and $t_2$in Tsuch that their join (or joint table) t_{12}still belongs to T. Then the upper and lower bounds for the cells t_1, $t_2$and t_{12} are related by: $t_1^L + t_2^L \leq t_{12} = t_1 + t_2 \leq t_1^U + t_2^U$ or $t_{12}^L - t_2^U \leq t_1 = t_{12} - t_2 \leq t_{12}^U - t_2^L$(and similar inequalities). At each iteration of the algorithm, such cell dependencies are used to improve[4] the current cell bounds. All the joins forming the current cell are checked as well as the joins to which the current cell belongs to. If the bounds of the current cell cannot be improved, the cell is "moved" into T_0. The algorithm iterates until no further improvement is possible or an inconsistency is found.

Unfortunately, the final bounds found by the GSA are not necessarily sharp, except in the decomposable log-linear model case and in the case of a dichotomous k-way table with all $(k-1)$-way marginals fixed. However, in Dobra and Fienberg (2000, 2001, 2003) a branch-and-bound method was proposed to sequentially improve the found bounds until they become sharp. The main idea is the following: if it is possible to find a feasible table for which the (current) upper bound $U$is attained and if there does not exist another feasible integer table having a count associated with the (current) cell t_1 strictly larger than U, then $U$is the sharpest integer upper bound for t_1.

[4] Decrease of upper bounds or increase of lower bounds.

Obviously, the same statement holds for the lower bounds. The algorithm developed by Dobra and Fienberg, sequentially fixes every cell at integer values between its current bounds and uses GSA to update the bounds for the remaining cells. The authors claim that this sequential improvement of the bounds avoids an exhaustive enumeration of all the combinations of possible values of the cells in Tthat would lead to a very low computational efficiency. Some practical aspects related to the implementation of the algorithm may be found in Dobra *et al.* (2003); a C++ code may be found at http://www.stat.washington.edu/adobra/software/gsa/.

As previously mentioned, the GSA was developed to deal with contingency tables derived by cross-tabulating categorical variables observed on the entire population. Issues related to samples of units and to continuous variables should still be investigated. Another key point is the link with the log-linear models. The release of partial information, i.e. the marginal tables, produces no information loss when log-linear models are used to analyse the associations between variables. Anyway, the sole release of marginal tables might have a significant effect on other types of analyses, e.g. Mantel-Haenszel tests or some logistic regression models, as discussed in Lee and Slavkovic (2008). The updated GSA version illustrated in Gibilisco (2009) exhaustively enumerates all feasible tables consistent with a set of linear constraints, e.g. marginals, bounds, and structural and sampling zeros. It should be noted that the presence of zero counts has a great impact on both data utility[5] and disclosure risk[6]. In statistical disclosure control, the GSA may be used as a theoretical background for perturbation methods, e.g. controlled rounding or generation of distributions over the corresponding space of tables using Markov bases. Such distributions may be used to evaluate the probability mass associated to each feasible[7] table or to generate synthetic contingency tables.

4.2 Microdata dissemination – Statistical disclosure control and record linkage

According to the privacy laws, a disclosure occurs when a unit is identified and/or confidential information about a unit may be retrieved. In sta-

[5] Being related to the non-existence of the maximum likelihood estimates, see Dobra et al. (2008).

[6] Zero counts might tighten the bounds of other non-negative cells, thus increasing the disclosure risk.

[7] Belonging to the polytope induced by the given marginals.

tistical disclosure control, record linkage methodologies have been used to derive several measures of disclosure risk based on external registers scenarios, i.e. making assumptions on how an intruder could identify units in a released microdata file. Generally speaking, the disclosure scenario describes/defines/models the uncertainty on intruder's information (data, tools and knowledge).

Several assumptions defining the external register scenario are: a) the intruder (the person who illegally wants to retrieve confidential information) has access to an external register covering the whole population, b) the external register and the microdata file share a set of key variables measured without error and c) the intruder would use record linkage methods to match a unit in the released microdata file to one in the external register using only the key variables. A detailed description of this external register disclosure scenario may be found in Polettini (2003). Based on these assumptions, several risk measures have been proposed; for example, the number of ”linked” units, which is a global[8] risk measure. At individual level, the probability of correct identification, i.e. the probability of disclosure, is seen as the probability of correct linkage, see Elamir and Skinner (2006) and references therein.

In the statistical disclosure control literature, to measure the risk of disclosure, several methodologies have been developed, taking into account different record linkage variants. First, for continuous variables, distance-based record linkages have been set up; each record in the microdata file is linked to its nearest record in the external register. The disclosure risk has been obviously expressed in terms of the number of units correctly identified. Different distance functions and different data structures have been considered in Domingo-Ferrer and Torra (2002, 2003), for example. More recently, the classical probabilistic record linkage setting has been used to measure the risk in presence of categorical key variables, relaxing also the assumption on measurement errors, see for example Skinner (2008) and Shlomo (2009).

Two common open questions are: how to choose the key (or comparison) variables? how an intruder could use other information about the disseminated microdata file (known population characteristics, known sampling design information, etc.) to improve the record linkage performance? Disclosure risk measures could greatly benefit from accounting for such auxiliary information in the record linkage process.

[8] At file level.

References on the use of data integration methods in statistical disclosure control

[1] Agresti A. (2002), *Categorical Data Analysis*, John Wiley & Sons, Inc., Hoboken, New Jersey.

[2] Barvinok A. (2009), *Asymptotic estimates for the number of contingency tables, integer flows, and volumes of transportation polytopes*, Int Math Res Notices, **2**, 348–385.

[3] Barvinok A., Luria Z., Samorodnitsky A., Yong A. (2010), *An approximation algorithm for counting contingency tables*, Random Structures and Algorithms, 2010, to appear.

[4] Dobra A., Fienberg S. E. (2000), *Bounds for cell entries in contingency tables given marginal totals and decomposable graphs*, Proceedings of the National Academy of Sciences, **97**, 11885–11892.

[5] Dobra A., Fienberg, S. E. (2001), *Bounds for cell entries in contingency tables induced by fixed marginal totals*, Statistical Journal of the United Nations ECE, **18**, 363–371.

[6] Dobra A., Fienberg S.E. (2003), *Bounding entries in multi-way contingency tables given a set of marginal totals*, In: Y. Haitovsky, H. Lerche, and Y. Ritov, editors, Foundations of Statistical Inference: Proceedings of the Shoresh Conference 2000, 3–16, Berlin, 2003. Springer-Verlag.

[7] Dobra, A., Karr, A., Sanil A. (2003), *Preserving confidentiality of high-dimensional tabulated data: statistical and computational issues*, Statistics and Computing, **13**, 363–370.

[8] Dobra, A., Fienberg, S., Rinaldo, A., Slavković, A., Zhou, Y. (2008), *Algebraic statistics and contingency table problems: Log-linear models, likelihood estimation and disclosure limitation*, In: Putinar, M. and Sullivant, S., eds, IMA Volumes in Mathematics and its Applications: Emerging Applications of Algebraic Geometry, vol. 149, pages 63–88. Springer Science+Business Media, Inc.

[9] Domingo-Ferrer J., Torra V. (2002), *Validating distance-based record linkage with probabilistic record linkage*, Lecture Notes in Computer Science. Vol. 2504 *(Topics in Artificial Intelligence)*, 207–215.

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

[11] Elamir E. A.H., Skinner C. (2006), *Record level measures of disclosure risk for survey microdata*, Journal of Official Statistics, **22**(3), 525–539.

[12] Fienberg (1999), *Fréchet and Bonferroni bounds for multi-way tables of counts with applications to disclosure limitation*, Statistical Data Protection, (SDP'98) Proceedings, 115–129. Eurostat, Luxembourg.

[13] Fréchet, M. (1940), *Les probabilitiés associées a un systčme d'evénments compatibles et dépendants*, vol. Premiere Partie. Hermann & Cie. Paris.

[14] Gail M., Mantel N. (1977), *Counting the number of $r \times c$ contingency tables with fixed margins*, Journal of the American Statistical Association, **72**, 859–862.

[15] Gibilisco P., Riccomagno E., Rogantin M. P., Wynn H.P. (2009), *Algebraic and geometric methods in statistics*, Cambridge University Press, New York.

[16] Good I.J. (1976), *On the application of symmetric Dirichlet distributions and their mixtures to contingency tables*, Ann. Statist. **4**, 1159–1189.

[17] Lee, J., Slavkovic, A. (2008), *Synthetic tabular data preserving observed conditional Probabilities*, In: Domingo-Ferrer, J. and Saygin, Y., editors, CD Proceedings, *Privacy in Statistical Databases.*

[18] Polettini, S. (2003), *Some remarks on the individual risk methodology*, Proceedings of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, 299–311.

[19] Shlomo N., (2009), *Assessing disclosure risk under misclassification for microdata*, UNECE/EUROSTAT Worksession on Statistical Confidentiality, Bilbao. Available at http://www.unece.org/stats/documents/2009.12.confidentiality.htm.

[20] Skinner, C. (2008), *Assessing disclosure risk for record linkage*, In, Domingo-Ferrer, Josep and Saygin, Yücel (eds.) Privacy in Statistical Databases. Lecture Notes in Computer Science 5262 Springer, Berlin, Germany, 166–176.

Chapter 5

Micro-Integration: State of the art

Bart F.M. Bakker

Centraal Bureau voor de Statistiek – CBS, The Netherlands

Summary: Data from administrative sources and surveys have measurement and representation errors. We present a theoretical framework for these errors. Micro-integration is the method that aims at improving the data quality by searching and correcting for these errors. We distinguish between completion, harmonization, and correction for the remaining measurement errors. We define the different errors, give examples of these errors from the daily practice (from the Social Statistical Database and the Virtual Census) and propose operating procedures to correct for these errors. If one combines register data with sample survey data consistent repeated weighting can be used for consistent estimation. Finally, the position of micro-integration in the total statistical process is described.

Keywords: Micro-integration, data linkage, data quality, data processing, consistent repeated weighting

5.1 Introduction

Traditionally, censuses and surveys are used to collect information needed for the production of official statistics. Nowadays, register data have become increasingly popular. The use of these data has many advantages: a much smaller response burden, the possibility of large sample sizes for the production of small domain statistics, and comparatively low costs. However, the wider use of register data has also revealed more and more quality issues (Grünewald & Körner, 2005).

One of the limitations of register data is that they usually have a small number of variables. It is not possible to produce the desired crosstables, if the two or more required variables are not in the same register. Data linkage techniques should be used to combine data from different registers and surveys. This report focuses on an important aspect of the statistical process after the linkage of different sources: the integration of administrative registers and household sample surveys at the micro-level in order to create integrated micro-data files of e.g. persons, families, households, jobs, benefits and dwellings.

We use the term "administrative register" if we mean the administrative data collected by the register keeper. We use the term "statistical register" for a statistical information system that is used to produce statistical outcomes. As statistical information systems should provide accurate, relevant and authoritative information, the transformation of social statistics from a wide variety of largely isolated statistics into an integrated statistical system is the logical consequence of these prerequisites. Authoritative outcomes are supported by consistent statistical outcomes.

The method of micro-integration is developed in the last two decades, in particular in the countries in which administrative register information is widely used to produce statistics. However, authoritative literature is absent. The existing literature (e.g. Statistics Denmark, 1995; Al en Bakker, 2000; Schulte Nordholt, Hartgers en Gircour, 2004; Statistics Finland, 2004; CBS, 2006; Wallgren en Wallgren, 2007) are more or less descriptions of best practices and not based on a theoretical basis. To speak of "State of the art" is perhaps premature. The exception to the rule is the method of consistent repeated weighting that is well described in articles in peer reviewed journals.

We start in section 5.2 with the definition of micro-integration and give the differences with related fields such as macro-integration and editing and imputation. As micro-integration aims at improving data quality by correcting for errors, it is necessarry to give an overview of possible errors in research in which different sources are combined. That is the contents of the third section. In the fourth section we give a review of the methods that are used in micro-integration: completion, harmonization, correction for other measurement errors and consistent repeated weighting. In this section, which is "the heart" of the report, we will give examples from the micro-integration processes used in the Social Statistical Database (SSD) and the Virtual Census (VC) of Statistics Netherlands. In section 5.5 we will discuss the use of micro-integration techniques in the statistical process. We conclude with some remarks on the applicability of the method and a bibliography.

5.2 Definition of micro-integration

Combining information from different sources can improve data quality. Data from single administrative sources and surveys have measurement and representation errors (Bakker, 2009b, see also section 5.3). Micro-integration is the method that aims at improving the data quality in combined sources by searching and correcting for the errors on unit level, in such a way that:

- the validity and reliability of the statistical outcomes are optimized,
- only one figure on one phenomenon is published,
- variables from different sources can be combined and as such, source and theme exceeding outcomes can be published, and
- accurate longitudinal outcomes can be published.

The term "error" in the defenition should be understood in a broad sense. It also covers the differences in concepts and operationalization of these concepts in the integrated sources. We shall elaborate on these errors in the next section.

In a strict sense, consistent repeated weighting is no micro-integration because it is not on unit level and it is not intended to correct for errors in the data. As this method is used to satisfy the condition that only one figure on one phenomenon is published, we describe the method anyhow.

Up to now, the method is only widely used for register data or other data in which the entire population is described such as censuses. In theory, there should be no difference between the micro-integration of two administrative registers and a register and a sample survey. Because of this practice, We give only examples from the combinations of register information.

Micro-integration diverges from macro-integration in that the data are corrected on the unit level. After the micro-integration process, all statistical output that is produced from the micro-integrated files is consistent. If one uses macro-integration techniques each new table has to be made consistent again on a meso- or macro-level.

Micro-integration is also related to editing and imputation. Micro-integration uses editing and imputation techniques to make the data of integrated microdata files more consistent. Editing and imputation is primarily used for the datacleaning of one source. An external source may be used to facilitate the detection and correction of errors in the primary source, but will not be edited itself. Because micro-integration is applied to different linked sources, you are able to improve the quality of the data much more than if you have limited information from only one source. Moreover, correction for coverage errors and harmonization, two techniques of micro-integration, are usually not incorporated in editing and imputing processes.

5.3 A framework for errors in statistics based on combined sources

5.3.1 To a framework for register based statistics

Possible errors in traditional survey processes are very well documented. Despite the increasing use and methodological developments, no framework has been created yet to classify the errors in register-based research or combined register and survey data. In this section we present such a framework.

We depart from the idea that the various errors that may occur in surveys are also applicable to administrative registers. In addition to these errors, there are also specific errors when administrative registers are used for the production of statistics, like administrative delay (an event is registered with a certain time lag) and linking errors if several registers are linked.

Most of the registers are constructed with the aid of some survey technique: face-to-face, paper and pencil, telephone or web-based. Let me give an example. Once a year most Dutch citizens fill out a tax form for the income taxes. It is possible to do that electronically or on paper. Filling out the tax form electronically is a kind of Computer Assisted Web Interviewing (CAWI), while filling out the paper tax form is a kind of Paper And Pencil Interviewing (PAPI). The tax authorities pay a lot of attention to the design of the electronic and paper forms, in order to avoid misinterpretation of the questions. Moreover, the electronic help function and the booklet sent with the paper tax form contain more information to achieve that goal. This is much like the design of a questionnaire and explanatory texts for survey research. The explanatory texts can also be used for the instruction of the interviewers.

Groves et al. (2004) published one of the leading publications on errors in survey research. They describe the ‘total survey error’ and distinguish between different components. Based on the life cycle of a survey (figure 5.1) they distinguish between ‘measurement’ and ‘representation’ errors. As the survey outcome is the crosstable of two variables, the errors on the measurement side have a shadow for the second variable. The representation errors are the same for both variables, but the size can differ.

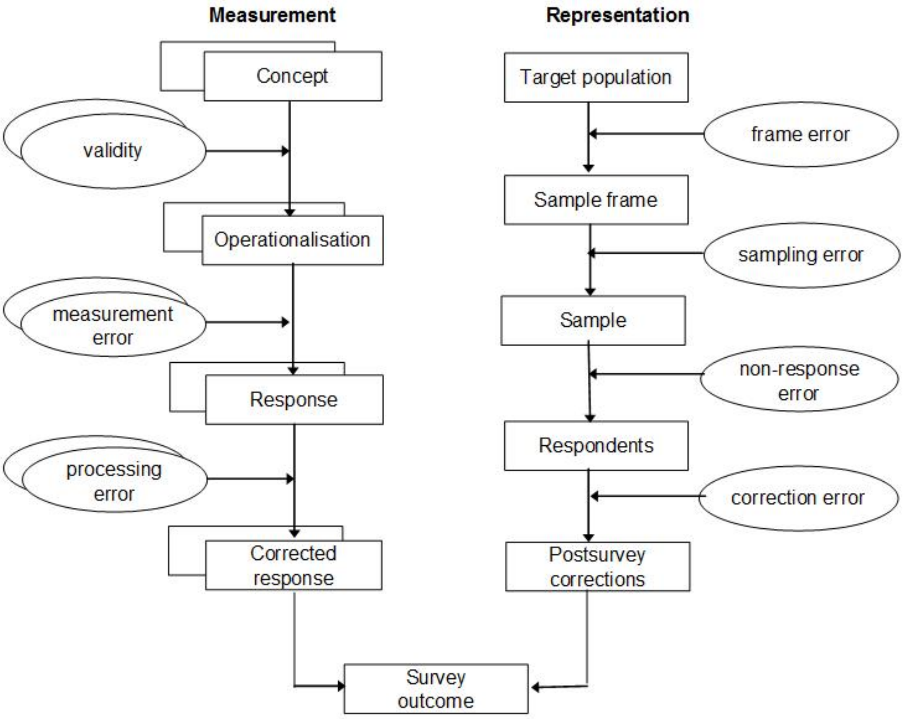


Figure 5.1. The 'life cycle' of and errors in a survey (Groves et al., 2004)

Based on the general idea that it is likely that the errors that normally emerge in surveys will also occur in registers, as most of the register data are collected with the aid of a survey technique, figure 1 can easily be adapted to the most common life cycle of registers. It is possible to distinguish between one register used on its own, and several registers used in combination with each other. As we have limited space, we present only a figure for combined register use (Figure 5.2). The columns "Measurement" and "Representation" errors refer to all the sources used to produce a statistical outcome.

The possible errors that are common in surveys will also occur in administrative register data. However, that does not mean that the errors are identical. They differ in size according to the interest of the register keeper and registered person, they differ to the extent that the results can be influenced by the researcher, and there are errors that are unique for register data.

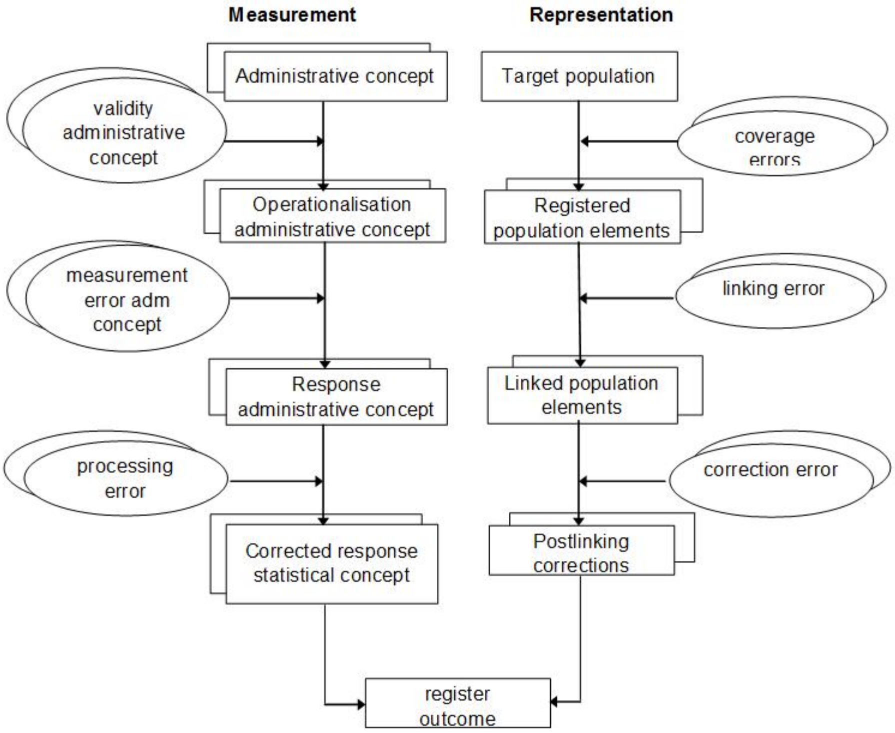

Measurement
Representation
Administrative concept
Target population
validity administrative concept
coverage errors
Operationalisation administrative concept
Registered population elements
measurement error adm concept
linking error
Response administrative concept
Linked population elements
processing error
correction error
Corrected response statistical concept
Postlinking corrections
register outcome

Figure 5.2. The 'life cycle' and errors in a combined register situation

5.3.2 Measurement

The measurement in a register starts with the definition of the theoretical concept the register keeper wants to measure for administrative purposes. This is called the administrative concept.

Operationalisation is more concrete than a theoretical concept: it is a way to collect information about the theoretical concept. The critical task for measurement is to design questions and protocols that perfectly measure the theoretical concepts. Only meticulous questioning can prevent bias in the answers (Groves et al, 2004; Czaja & Blair, 2005 pp. 59–84). Questions can be communicated face-to-face, by telephone, on paper or electronically. In the phrasing of the question to measure the theoretical concept, errors may occur that leads to invalid answers. As long as the theoretical concepts are easy to understand for both register keepers and registered units (persons, companies), no significant problems occur. If the theoretical concept becomes more difficult to understand, this will easily lead to biased measurement. Register keepers invest substantially in the wording of their questionnaires.

The size of errors in administrative registers depends on the control processes that the register keeper executes. Of course all kinds of checks by the register keeper in the administrative process can correct for errors in the preceding interview. An employment officer may demand to see documents (e.g. payslips, diplomas or certificates) to verify the information that the job-seeker has given to prevent possible errors. Or a tax employee can link other register data in order to search for inconsistencies in the data which give rise to suspecting the quality of the data. Any remaining irregularities are mostly corrected in the register in consultation with the reporting instance or person concerned. In some cases the recorded data are audited by accountants or other inspectors. These administrative protocols are formulated in order to maximize the quality of the measurement of the variables that are important for the purpose of the register keeper. Therefore, we may assume that the quality of these data is better than that of variables that are considered less important. An example of such a variable is the end date of jobs in the income taxes. As the tax sum in the Netherlands does not depend on this information, the tax authorities do not pay much attention to it and it is therefore at risk of low quality.

The size of errors also depends on the interest of the registered persons. That is to say, if it is in the interest of the registered person to be registered falsely in a specific way, the probability that this misregistration will occur will be larger. If you are interested in the data quality of a specific registration, it is important to specify the interests of both register keeper and registered. This information can be useful to formulate hypotheses for bias in the data.

After the phrasing of the questions, the interviews are carried out. Many measurement errors are encountered in this step of the life course of a register. We mention only memory effects like false recollection and telescoping (e.g. Sikkel, 1988; Auriat, 1991, 1993; Smith & Duncan, 2003; Schroots, Van Dijkum & Assink, 2004), interviewer effects (Pannekoek, 1988; Brick et al., 1995; Ganninger, Häder & Gabler, 2007), and deliberate misreporting (e.g. Belson, 1986; Groves et al., 2004).

One measurement error is unique to registers. When using registers for the production of statistics, one of the errors that must be taken into account is the so-called administrative delay. This delay is caused by events being recorded some time after they actually occur, and it is an important source of error. Of course, if a survey collects information on past events, this is also a sort of delay, but the information on the past is always available at the time the outcomes are published. Registers that contain administrative delay are used at a moment in time that not all the events have yet been recorded.

This may lead to substantial bias in the estimation of events in a certain period. For instance, marriages contracted in immigrants' country of origin are sometimes recorded two or three years after the event. This can lead to a certain bias in the register outcomes. The direction of the bias depends on fluctuations in the number of events and in the size of the administrative delay. A decrease of the administrative delay will lead to overestimation of the events, an increase will lead to the opposite.

The interviews lead to a response and a new entry in the register. The response is corrected by a set of decision rules. Implausible values are deleted or sometimes imputed, missing information is imputed with the use of a model, new variables are derived by combining the information from several variables, and alphanumerical information is coded. In all these processing steps, it is possible that wrong decision rules are applied.

5.3.3 Representation

On the representation side, the first step is defining the target population. Under-coverage will result if the target population of the register is not completely covered by the entries in the register. For instance, the target population of the Population Register is all inhabitants living in the Netherlands for at least four months. However, the register does not include the 'illegal' population even though it is part of the target population (Van der Heijden *et al.*, 2006). This results in under-coverage of the target population. Administrative delay in registrations can lead to under-coverage (e.g. birth and immigration) and over-coverage (e.g. death and emigration).

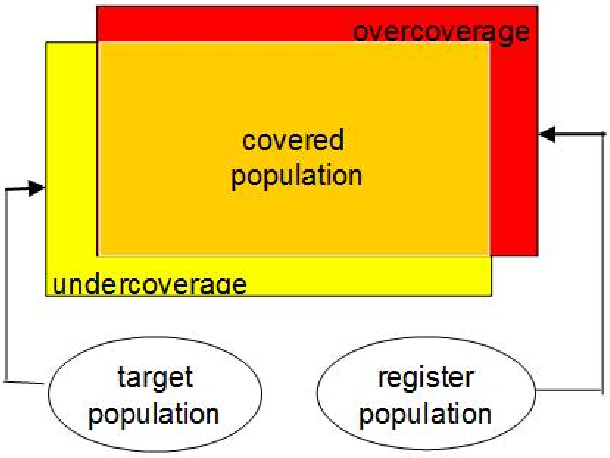


Figure 5.3. Coverage error

Administrative records from different registers should be combined by linking. In most cases the linking key is a personal identification number (PIN), or - if such a number is absent in the register - a combination of variables, for example birth date, sex and address. Two types of linking errors can occur: missed links and mislinks (Fellegi & Sunter, 1969; Arts, Bakker & Van Lith, 2000). Missed links are cases where the record cannot be identified; they correspond with errors caused by non-response in surveys: if missed links are non-random, they will lead to biased outcomes. Mislinks occur if records of two different elements are combined.

5.3.4 Consistent estimates if register and survey data are combined

One of the aims of micro-integration is that only one figure on one phenomenon is published. If a survey is enriched with register data the population totals of the register variables are estimated by assigning weights to each record of the combined dataset. These weights are determined in such a way that the distribution of a set of margins and crossings of register variables are reproduced. This results in estimates that are consistent for all the variables that are used in the weighting procedure. However, the estimates are not consistent with the register variables not used to determine the weights. It is impossible to take all possible variables into the weighting procedure to guarantee consistency. There will be too little degrees of freedom. If two or more surveys are linked to a set of linked registers, using one set of weights per survey, many estimates will be numerically inconsistent across surveys.

To produce consistent estimates the method of ”consistent repeated weighting” has been developed by Statistics Netherlands (Kroese, Renssen & Trijssenaar, 2000; Houbiers, 2004; Gouweleeuw & Hartgers, 2004). Conventional weighting procedures assign weights to records in a (combined) data file and estimate all tables with the use of those weights. Consistent repeated weighting procedures assign weights to records in a (combined) data file for each table in such a way that later estimates are consistent with previous estimates. This technique is discussed in section 5.4.4.

5.4 Micro-integration techniques

5.4.1 Introduction

Representation errors exist if the target population is incompletely described by the data. We distinguish between over-coverage and under-coverage. By means of *completion* we detect and correct for these errors. Measurement errors may also occur when the determining the elements of the population. For the correction of these measurement errors other micro-integration techniques are applied.

Measurement errors exist if characteristics of the population elements are not correctly described. These errors may have different causes. By using information from different sources, these errors can be detected and corrected. For the correction on a conceptual level we use *harmonization.* For the correction on data level we use *correction for measurement errors.*

For the detection of representation and measurement errors we search for inconsistencies in the data. Only if it is of interest for the publication of statistical outcomes, inconsistencies should be solved. Micro-integration can be applied in a situation that different administrative sources are available for the same subject or the same administrative source for different periods. These administrative data are more or less integrally, that means that the population of the administration is covered completely. However, it is not necessary to restrict the application of micro-integration to integral administrative sources. It is also possible to apply micro-integration techniques to a linked sample survey and integral registers and even two sample surveys if the overlap is large enough. The experience with micro-integration of survey and register data is limited. We restrict our examples to the practice of register data.

Consistency is an aspect of quality that has its own merit apart from validity and reliability. Statistical outcomes are of better quality simply and solely because they are consistent with other statistical outcomes as this makes it possible to make comprehensive descriptions of subjects.

5.4.2 Completion

5.4.2.1 Detecting representation errors

In the preparation of research, one of the first steps is defining the target population. The target population is the population on which the research data are collected and outcomes are presented. The difference between the target population and the observed population is called the total representation error. The total representation error can consist of:

- Under-coverage or over-coverage because the population of the integral register (or the sample survey) differs from the target population of the research.

- Under-coverage because elements of the target population are missing in the administrative data. One important cause is administrative delay. If persons are the statistical entity, administrative delay in the registration of birth and immigration will lead to under-coverage.
- Over-coverage because elements that do not belong to the target population are (still) in the administrative data. One important cause is administrative delay. If persons are the statistical entity, administrative delay in the registration of death and emigration will lead to over-coverage.
- Under-coverage because elements that belong to the target population can not be linked (”missed links”).
- Over-coverage because elements that do not belong to the population are wrongfully linked (”mislinks”).

Representation errors ideally are detected by comparing to a reference dataset that contain all population elements. If the data under study include elements that are not in the reference data file, that will be a matter of over-coverage. If the data under study do not include all elements from the reference data file, that will be a matter of under-coverage. In most cases such a reference dataset is not available and has to be created during the process of micro-integration by combining all sources that contain elements of the population. Some examples can illustrate this.

Example 1. The target population of a research is: ”the students in higher education that belong to the Netherlands population on October 1^{st} 2009”. The best fitting source to define this target population is the so-called Central Register for Enrolment in Higher Education in combination with the Population Register. The first source contains yearly information on students in higher education in the Netherlands from study year 1985/’86. There are two important errors in the representation of the statistical target population: the register covers only higher education in the Netherlands that is publicly financed. This means that students who live in the Netherlands but study in Belgium or Germany and students who take courses at private colleges and universities are not covered (Bakker, Linder & Van Roon, 2008). The Population Register has over- and undercoverage problems like e.g. the temporary workers from abroad, illegal population and already emigrated persons who are still registered in the Population register (Bakker, 2009b; Van der Heijden et al., 2006).

Example 2. The target population is the Netherlands' population on January 1^{st} 2009 who are suspected of a criminal offence. The most suitable administrative data source is the combination of the Population Register and the so-called Suspect Identification System (SIS) of the police. Apart from the coverage problems in the Population Register mentioned in example 1, coverage errors also are a result of linking errors. Two types of linking errors can occur: missed links and mislinks (Fellegi & Sunter, 1969; Arts, Bakker & Van Lith, 2000). Missed links are cases where the record cannot be identified; they correspond with errors caused by non-response in surveys: if missed links are non-random, they will lead to biased outcomes. Mislinks occur if records of two different elements are combined. If the different elements both belong to the target population, there is no coverage problem. Mislinks then usually lead to underestimation of the correlation between variables. These errors should be treated as measurement errors, because if one or two variables are measured with certain unreliability, the correlation is usually underestimated. If one of the elements do not belong to the target population and the other does, this will lead to overcoverage. From the SIS records only 89% can be identified in the Population Register, approximately 6% has a foreign address and do not make part of the target population. This means that around 5% are missed links. These missed links are highly selective and therefore will lead to selection bias (Blom et al., 2005).

5.4.2.2 Correction for under-coverage

There are different methods to correct for under-coverage. We distinguish between:

- Combine different sources to create a complete list of the elements of the target population.
- Assign weights to the population elements that are observed in order to represent the target population.
- A form of unit imputation in order to represent the target population.

The second and third methods are well described in the literature on the correction for non response in surveys (e.g. Groves et al., 2004; Stoop, 2005) and therefore we will not elaborate on this. Note that you need a frame to weight or impute. If you miss such a frame, you need to create one as we describe below.

The correction for under-coverage starts with the precise definition of the target population. The statistical target population should be operationalised by using several administrative registers covering different populations, taking into account that these administrative registers themselves are also under or over-covered. Linking records from different sources should provide a check on the completeness of the administrative registers.

The target population of the research of the first example in section 5.4.2.1 is: students in higher education who belong to the Dutch population on October 1^{st} 2009. The under-coverage in the Central Register for Enrolment in Higher Education consists of students who live in the Netherlands but study in Belgium or Germany and students who take courses at private colleges and universities are not covered (Bakker, Linder & Van Roon, 2008). There is one source that contains individual information on these missed students: the Study Financing Law Register, the law covering study grants in the Netherlands. From 1995 onwards all students who receive a study grant from the Dutch government are included in this register. The target population is well covered. All students who received a study grant are registered, also students who study in Belgium or Germany and on part of the private schools. Linking these two administrative registers will cover almost the entire population.

In the case of criminal suspects (Example 2 in section 5.4.2.1) another method is used to reduce under-coverage. It is known that moving is one of the main reasons for missed links in general. On top of that, suspects have interest in misleading the police officers in giving false name and address information. This leads to a situation that their data can not be linked. However, this situation is only temporal, as in many cases after a while the correct personal details are registered. In the SIS the personal details that identify suspects are updated permanently. By making use of the most recent information, more and more of the suspects can be identified.

It is not always possible to correct for under-coverage. If you lack reference data or the administrative data that can be used in combination for that purpose, none of the methods is entirely appropriate. However, it is possible to estimate the size of the under-coverage by using survey information. Of course the sample of this survey should not be restricted in the manner as register data are. But if you have a survey that covers the target population and you can link this survey to the register perfectly, than the under-coverage can be estimated by the weighted total of the records that can not be linked.

5.4.2.3 Correction for over-coverage

Over-coverage of the target population should be corrected by means of deleting the elements that do not belong to the target population. To execute this, those elements have to be identified as such. An exact operationalization is necessary for this purpose. An example of an exact operationalization for e.g. the category of job seekers on October 1^{st} 2008: "the persons who are registered in files of the employment agency 2008, version of April 1^{st} 2010 (which is corrected for administrative delay up to January 1^{st} 2010) and on October 1^{st} 2008 score "yes" on the variable job search.

We can distinguish between the following situations:

The definition of the target population can be operationalized within the year volumes of one data source. In this situation the correction for over-coverage is relatively simple. In the above mentioned definition, the correction for the administrative delay has been executed because all the events that take place afterwards are already processed in the data in the version of April 1^{st} 2010.

The definition of the target population can not be operationalized within the year volumes of one data source, but other sources are required to identify the elements that do not belong to the target population. The over-coverage that is caused by the administrative delay in the employment agencies register, can be corrected by linking the register of the employment agency to the employment registers. The starting date of a job can be considered as the transition date from job seeker to employee.

It is not always possible to correct for over-coverage, e.g. in the case of over-coverage caused by mismatching. The number of mismatches can be estimated (e.g. Arts, Bakker & Van Lith, 2000). However, it is not known to how much over-coverage this will lead, because part of the mismatched records can belong to the target population. In these situation, information on two different elements are linked. These errors are similar to the measurement errors in surveys. Up and above it is not possible to identify the elements that cause the over-coverage. Therefore they can not be deleted.

5.4.3 Harmonization and correction for other measurement errors

5.4.3.1 Detection of measurement errors

Measurement errors occur if characteristics of the elements of the population are described wrongly. Administrative registers and surveys comprise all kinds of different measurement errors. We can classify the measurement errors in surveys into three categories: errors made in the conceptualisation of the variables, errors made in the collection of the data and errors made in the processing of the data (see section 5.3). Inconsistencies in the data are an indication of possible errors. We distinguish between the following situations:

- If different sources contain the same variable, outcomes could be inconsistent, e.g. a person is registered in one register as a male and in another as a female.
- If a logical relationship between variables exists that is violated by the data, e.g. the wages earned in a year unequals the sum of the 12 monthly wages.
- If the state and transition figures are inconsistent, e.g. the population on January 1^{st} 2009 plus the number of birth and immigrants during 2009 minus the number of death and emigrants during 2009, does not count to the population on January 1^{st} 2010.
- If there is an impossible transition from one situation to the other, e.g. a transition from "married" to "never married".
- If there is an implausible combination of situations, e.g. someone has two fulltime jobs at the same time, or a fulltime job and a complete unemployment benefit.
- If data are inconsistent with some reference data. This can be checked very simple by setting range limits for a variable using information from an external source, but also by more complex methods based on relations between two or more variables, and even on outlier detection in regression analysis.

A particular case for inconsistency is longitudinal inconsistency. By that, we mean that the information on a certain period is not correct to estimate the transitions and therefore changes in (sub)populations. Longitudinal inconsistency is mainly caused by administrative delay and changing rules and regulations which leads to other measures of variables. For example, marriages of migrants who marry a bride or a groom from their native country are registered sometimes with a delay of more than two years. This will lead to biased estimates depending on the fluctuations in the administrative delay of these events. If these events are linked to other events registered without any delay, the relationship will be estimated biased.

5.4.3.2 Harmonization

Statistical research starts with the question what should be measured. In the first step this is defined conceptually. Two examples of conceptually definitions: "an employee" is "a person who holds a job and is employed by an employer", and the conceptual definition of a job is "a set of tasks and duties performed, or meant to be performed, by one person..." (International Labour Organization, 2007)

After the conceptual definition of the variable, the concept should be measured. In a survey this is done by transposing the conceptual definition into a questionnaire. In the questionnaire the exact criteria are given to measure the concept. In administrative registers the measurement of the concept is done by deriving the variable from register information. In administrative registers, the degrees of freedom for deriving the conceptually defined variable correctly is limited as the variables in registers are measured for administrative purposes. It is sometimes difficult to derive the correct statistical variable from the administrative information if the information in the administrative variables is not detailed enough, or simply measures something else (Wallgren & Wallgren, 2007, pp. 92–93). In some cases it is impossible to quantify the concept using the administrative data. In the situation that you have only one variable at your disposal in the combined registers and the administrative concept differs from the statistical concept you want to measure, it is almost impossible to validly measure the variable. The transposition of the information of different registers or surveys to one concept is called harmonization.

Harmonization consists for the greater part of the formulation of decision rules, in which the measurement of a concept is determined as precisely as possible, given the existing information in the data sources. To do this correctly, it is necessary to use knowledge on the academic and public meaning of the concept and knowledge on the information in the sources that can be used for measuring the concept.

5.4.3.3 Correction for other measurement errors

After harmonization has been executed to diminish inconsistencies in the data, the remaining inconsistencies are solved by chosing the best source for each variable. In chosing the best source it is important to know which variables are crucial for the register keeper to carry out his administrative duties. If a variable is not important to the register keeper, it will be at greater risk to have a low quality, as the register keeper shall pay little attention to its quality and spend not much auditing time.

The quality of a variable in a source can be strong at one point, but weak on another. For example, the yearly wages in source A can be of very good quality for government employees, but of fairly poor quality for employees in other economic sectors. If source B is fairly good for all employees, the yearly wages of government employees are derived from source A and of the other employees from source B.

If the details of the quality of the sources is unknown, sometimes the new variable is derived from two or more sources by taking the mean. It is also possible to formulate a decision rule in which the data are adjusted in such a way that a relationship between two or more variables is correct. It depends on the quality of the data which information is adjusted.

5.4.4 Consistent repeated weighting

Let us assume that we want to produce consistent estimates from a dataset in which one register and one survey are linked (Figure 5.4). The register is produced by linking several registrations and comprises the x-variables $x_{1,\ldots,}x_n$. By applying micro-integration techniques like completion, harmonization and correcting for other measurement errors, the register records in this dataset are consistent. The data block of the survey comprises the variables $y_{1,\ldots,}y_n$. and does not contain variables already available in the register block. Furthermore, the survey records have been assigned design based weights $d_{.i}$ which are to be calibrated with the use of a weighting model to correct for non-response which results in weights w_i The x-variables that are used to calibrate the design based weights are part of the register part of the dataset. For all variables that are used in the weighting model, consistent estimates are guaranteed, whether you count from the register data block or estimate the variables from the enriched survey data block of the dataset. However, the estimates for other x-variables could be inconsistent between the whole register and the enriched survey data block of the dataset.

Several possible solutions to this problem have been proposed: massive imputation and extending the weighting model by more variables (e.g. Kroese, Renssen & Trijssenaar, 2000). Both techniques have similar limitations. If the number of estimates you want to produce are small, then it is possible to design an imputation or weighting model that produces consistent estimates. If you want to estimate all your statistical output from such a dataset, and if you link all register information into one dataset and combine this with all your survey data this should be the case (Bakker, 2002; Houbiers, 2004), then there are not enough degrees of freedom to get a sufficiently rich imputation or weighting model (Kroese, Renssen & Trijssenaar, 2000). Therefore, an alternative to usual weighting and imputation procedures was developed to be able to produce a consistent set of tables using available registers and surveys: consistent repeated weighting (CRW).

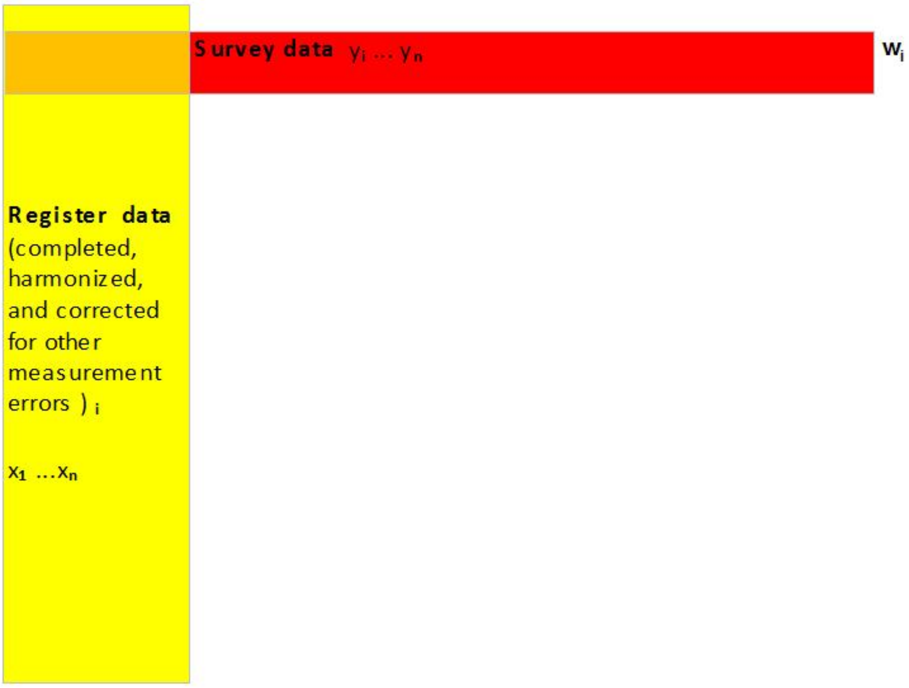

Survey data $y_i \dots y_n$
w_i
Register data (completed, harmonized, and corrected for other measurement errors) $_i$
$x_1 \dots x_n$

Figure 5.4. Example of linked registers and a survey

To estimate a fully consistent set of tables the following procedure is adopted (Kroese, Renssen & Trijssenaar, 2000; Renssen et al., 2001; Houbiers, 2004):

Each cross-tables T_k (k=1,...,K) will be based on the most suitable data block. In most cases this is the data block in which the statistician has the most confidence. As micro-integration already has maximized the validity of the variables measured, normally we have most confidence in the largest data block . Tables form larger data blocks are estimated before tables from smaller data blocks.

If a cross-table T_k has a margin T_m that can be estimated from a larger data block, this margin should be estimated first. In particular this will be the case for the variables x which are used to enrich the survey data. The variables x are measured for the entire population, but only a small part of the units are in the enriched survey data block (red and orange). They should be estimated by counting in the register block (yellow and orange).

All cross-tables that can be estimated consistently with the block weights w_i should be estimated before the tables that cannot be estimated consistently. This should be applicable for all tables with only variables from the survey and variables used in the weighting model.

If a cross-table T_k cannot be estimated consistently with the block weights of the most suitable data block, the table must be estimated by consistent repeated weighting. That is, the block weights w_i should be adjusted in such a way that the margins and cross-tables estimated in the steps mentioned before are reproduced. The block weights w_i are adjusted slightly to estimate the table in question.

Consistent repeated weighting is based on the repeated application of the well-known regression estimator and generates a new set of weights for each table that is estimated. Let y be a scalar variable of which the population parameter -either total or average- ought to be obtained for a table through a set of explanatory variables x from a register. The regression estimator of the population average for y is defined by

$$\hat{\bar{Y}}_{REG} = \hat{\bar{Y}}_d + b_s'\left(\bar{X}_p - \hat{\bar{X}}_d\right)$$

$$b_s = \left(X_s' D_s X_s\right)^{-1} X_s' D_s y_s; \quad D_s = diag(d_1, ..., d_n),$$

where $\bar{X}_p$ and $\bar{Y}_p$ are the population means of x and y, respectively while $\hat{\bar{X}}_d$ and $\hat{\bar{Y}}_d$ are their estimates based on the design weights d_i and b_s is the estimated vector of regression coefficients. X_s is the matrix of sample observations on the x-variables and y_s the vector of observations on the variable y. Instead of these traditional regression estimators, the repeated weighting procedure uses a set of coefficients of the form

$$b_w = \left(Z_s' W_s Z_s\right)^{-1} Z_s' W_s y_s; \quad W_s = diag(w_1, ..., w_n), \tag{5.1}$$

where Z_s is the matrix of sample observations on the variables in the margins of the table with variable y. The averages of the marginal variables z have been estimated already in an earlier table or are known from a register. Denoting these estimates or register counts by $\hat{\bar{Z}}_{RW}$, the repeated weighting estimator of $\bar{Y}$ is defined by

$$\hat{\bar{Y}}_{RW} = \hat{\bar{Y}}_{REG} + b_w'\left(\hat{\bar{Z}}_{RW} - \hat{\bar{Z}}_{REG}\right). \tag{5.2}$$

Substituting (5.1) into (5.2), it can be shown that the weights thus obtained for the records in the micro-dataset are adapted in such a way that the new table estimate is consistent with all earlier table estimates; see Knottnerus and Van Duin (2006).

If more surveys are linked to the linked register data block and they have variables in common, a separate rectangular data block consisting of records from the union of these surveys can be created. Cross-tables concerning these common variables can be estimated more accurately from the union of these survey data. Following the steps mentioned before, this can be achieved by applying CRW. However, it requires that the definition and the measurement of the variable is the same in both or all surveys and preferably the sampling frames should be the same too (Houbiers, 2004).

A point of attention should be the order in which the tables are estimated. Even when the cross-tables are estimated according to the steps mentioned before, there is no unique estimate for tables that are estimated by repeated weighting. Because the adjusted weights for each table may differ since they depend on the weighting model used. The weighting model, in turn, depends on the tables that have already been estimated. In order to tackle this problem, a fixed order can be used in addition to the rule that cross-tables from larger data blocks are estimated before cross-tables from smaller data blocks. It is called the splitting up procedure. Let us assume that we are interested in a three way cross-table of x, y and z. Firstly, the one-way tables for x, y and z are estimated. Secondly all two-way tables (x by y, x by z, y by z) are estimated under the restriction that the one-way tables of x, y and z are reproduced. Finally the three-way table x by y by z is estimated, taking the two-way tables into account.

Another point of attention is related to the occurrence of empty cells in the survey: sampling zeros. If the interior of a cross-table has to be calibrated on some counted or estimated population total but in the data block from which the table must be estimated there are no records satisfying the conditions, it will then be impossible to find a solution for the repeated weighting estimator. This problem arises in particular when a survey data block has a large and selective non-response. One way to deal with this problem is to combine several categories in the variables where the problem occurs. As a consequence all estimates of a higher order that were executed before should be repeated. Another way of dealing with this problem can be found in the use of synthetic estimators. One replaces the sampling zeros by a very small value to avoid the estimation problems analogously to the application of log linear models (see Bishop, Fienberg & Holland, 1975; Houbiers, 2004).

If one uses CRW to estimate consistent tables one should take edit rules into account. Edit rules are used in the micro-integration process in particular but not exclusively to correct for "other measurement errors". The CRW could lead to cross-tables that violate the edit rules. To avoid this, one should include the variables used in the edit rules in the CRW weighting model (Renssen et al., 2001). Consider a register containing the categorical variable age and a sample containing the categorical variable driving license ownership. Suppose that the frequency of age as a classification variable has already been estimated and that we define an edit rule: if "age < 18 then license = no". Let P (license) denote the population fraction of license ownership and $P(\geq 18)$ the population fraction of persons older than seventeen, then we have

$$P(\text{license}) = P(\text{license}| \geq 18)[P(\geq 18)] + P(\text{license} < 18)[1 - P(\geq 18)]$$

Utilizing the edit rule we determine that $P(\text{license}| < 18) = 0$, from which it follows that $P(\text{license}) = P(\text{license}| \geq 18)[P(\geq 18)]$, where $[P(\geq 18)]$ is already estimated from the register. It is rather easy to formulate a reweighting scheme for this particular example by taking the minimal re-weighting scheme from crossing between age and driver's license, we obtain post-stratification with the age classes as post-strata.

If one uses different aggregations of one variable than these different aggregation levels should be hierarchically nested. Otherwise, the number of categories that the estimation should be consistent with will be too large and it also leads to empty or almost empty cells.

Knottnerus and Van Duin (2006) give the variance formulae for the CRW estimator, and test CRW estimators under various conditions. Several simulation studies, e.g. Boonstra (2004) and Van Duin and Snijders (2003) show that the method of consistent repeated weighting leads to estimates with lower variances than usual estimation methods, due to a better use of auxiliary information.

5.5 The position of micro-integration in the statistical process

Micro-integration includes the processes that are executed to repair the errors in the preceding administrative processes, i.e. the process from the response of the administrative concept of a set of registered population elements to the outcomes of the statistical concepts and the statistical population (Figure

5.5). The starting point of micro-integration is twofold: the inconsistencies in the linked dataset and the knowledge of the errors in the original sources.

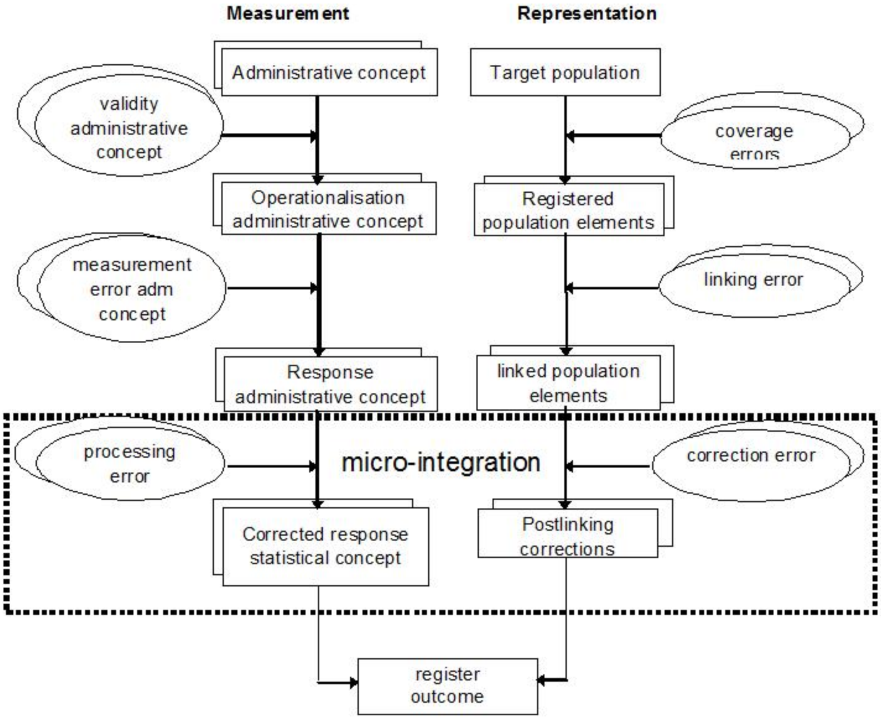

Measurement
Representation
Administrative concept
Target population
validity administrative concept
coverage errors
Operationalisation administrative concept
Registered population elements
measurement error adm concept
linking error
Response administrative concept
linked population elements
processing error
micro-integration
correction error
Corrected response statistical concept
Postlinking corrections
register outcome

Figure 5.5. The position of micro-integration in the statistical process

Van der Laan (2000) provided the first model for a micro-integration process:

- *harmonization of units*: are the statistical units defined uniformly in all sources? (special reference to comparability in space and time);
- *harmonization of reference periods*: do all data refer to the same period or the same point in time?
- *completion of populations (coverage)*: do all sources cover the same target population?
- *harmonization of variables*: are corresponding variables defined in the same way? (special reference to comparability in space and time);
- *harmonization of classifications*: are corresponding variables classified in the same way? (special reference to comparability in space and time);

- *adjusting for measurement errors (accuracy)*: after harmonising definitions, do the corresponding variables have the same value?
- *adjusting for missing data (item non-response)*: do all the variables possess a value?
- *derivation of variables*: are all variables derived using the combined information from different sources?
- *checking overall consistency*: do the data meet the requirements imposed by identity relations?

This model was developed in the nineties in a pilot of the so-called Social Statistical dataBase (SSB) in the Netherlands and was very valuable at the time. By now, it is an idealized image of daily practice and no longer complete. In this section we give comments on the model with the aim to improve and update it. The first comment is that some steps can be formulated more generically. The second comment is that the steps in the model flow in practice more together. The third comment is that some important steps are not in the model, e.g. consistent and repeated weighting.

In the preparation of statistical research, the research questions are formulated. To make this more tangible, the target population and the concepts you want to measure of this target population are defined.

In the first stage in the micro-integration process the target population is operationalized in all necessary sources you combine. This is done by identifying all the population elements by assigning a linking key. For disclosure purposes, this could be a meaningless number. It is not only necessary to uniformly define and measure the population elements, but define and measure it according to a conceptual definition that was developed beforehand.

In a number of cases it is necessary to derive one or more variables that are needed to define whether an element belongs to the population. An example can illustrate this. If the target population is: ”the jobs on ultimo March 2010”, and a job is defined as a contract between a ”company” and a ”person” in which is agreed that the employee executes particular activities for which the employer pays a loan in return, it is necessary to define and measure ”company” and ”person” and harmonize all the information on these variables. For ”person” this will not be problematic, but for ”company” this certainly is difficult.

After that, you have to decide whether the combined dataset contains double population elements. Not entirely harmonized information on ”company” for instance will lead to missed links and overcoverage of the population of jobs: double population elements are not recognized as such. In other words, you first harmonize the elements that are used to define the units, than you derive the units according to a standard definition, and you delete the double population elements. It is possible that each source contains unique jobs which are not covered by other sources.

Particular attention should be given to the reference period and space. As far as the reference period concerns, the information on the end dates of jobs can differ between sources. Because we know that the end dates of jobs are of relatively low quality in all of the registers, we confront the information on jobs with the information on social benefits and other jobs later in time. This information leads to a number of corrections in the end dates of jobs.

Space is important because you have to decide whether jobs in companies established in a foreign country belong to the population or not. In addition you have to answert the question whether jobs of persons living in a foreign country and working in a Dutch company belong to the population or not.

After the definition and measurement of the population elements in all sources, the linking of the elements and the deletion of the double elements, attention should be paid to the number of missed links in each of the sources. If you combine for instance three sources and the linking effectiveness is 98%, you hope that a missed link in the one source is a link in one of the other sources. Assuming that the union of the sources contain all population elements, you completed the population. Of course, you will never be sure just because you can not fully identify the information.

To test whether the population elements cover the target population, ideally you should compare the result of the above processes with a list of population elements. However, if such a list would have existed you surely would have used it in the process of operationalization of the target population. So in most cases you have to test the completeness in another way. One of the possibilities is to estimate the number of population elements in a survey. In our example the number of jobs can be estimated from the Labour Force Survey (LFS). The response of the LFS is weighted to the total population and you can compare the results. Of course you have to use the same concept of job and the same reference period and reference space. If the results differ it is likely that there is over-coverage or under-coverage. It is also possible that this is the case even if the estimates do not differ, namely if under- and over-coverage are of equal size. Linking the LFS to the registers should shed more light on the under- and over-coverage.

After the determination of the list of population elements, the variables of the population elements should be harmonised in all the sources. This has already been executed for the variables used for the determination of the target population. This process step starts with the conceptual definition of the variables that are necessary for answering the research questions. The information in the combined data file is used to measure these concepts as good as possible. If the information in the original sources differ, the information from all the sources is transposed to the same concept. This is called harmonization of variables. A specific part of this is the use of standard classifications. If sources contain different classifications of the variables, these should be converted to the standard classification. It is also possible to derive variables from variables from two or more different sources. The number of missing values in the original variables should be low. A high number of missing values in different variables in the original sources will lead to very high number of missing values in the derived variable, because each missing value in any of the sources will automatically lead to a missing value in the derived variable unless the missing values are replaced by imputation into a ”real” value.

In the next step you should correct for other measurement errors as is described in 5.4.3.3 and check the overall consistency in the same way. The crux of this technique is to define the right edit rules. In the last step, and only if register and survey data are combined, consistent and repeated weighting could be applied.

5.6 Some concluding remarks

Micro-integration is a technique for clearing data from combined sources, in particular for administrative registers. However, there is also an opinion that the micro-integration of register data mask part of the measurement errors in the data while there is no guarantee that the data quality improves substantially (Van der Velden, 2003). An alternative way to correct at least partly for measurement error is to apply linear structural equation models with a measurement model part. In psychology this is a frequently applied methodology (e.g. Jöreskog & Sörbom, 1996; Kline, 2005). It is based on the classical test theory in which repetition of measurement or multiple indicator measurement is used to model the error structure of the data. This is a valid methodology for testing hypothesis with pathmodels, but is less applicable if one want to publish cross-tables. Because publishing cross-tables is the core business of national statistical institutes this is not a realistic alternative to micro-integration. However, it is a promissing method for research into the measurement errors of register data (Kaptein & Ypma, 2007; Bakker, 2009b).

Denk and Hackl (2003) emphasize that a previous analysis of differences between sources could prevent problems that may arise when linking and data-integration is actually executed. Ideally all the relevent information should be in the meta-information of the sources, but in practice, this is normally not the case. One of the important reasons why register data error is not known, is that register keepers are not interested in the quality of part of the register data or they have interest in preventing that those errors become known. Therefore effort should be put into research to the data quality of the different sources. The "life cycle of register based research" as is discussed in section 5.3 can be used to formulate a research plan.

Consistent repeated weighting is a technique to get estimates from linked register data (or census data) and sample surveys. Of course it is possible to use other techniques to produce those estimates. Haslett et al. (2010) describe three alternatives: small area estimation and in particular the ELL-method (Elbers, Lanjouw and Lanjouw, 2003), mass imputation (Kovar & Whitridge, 1995; De Waal, 2000) and (spatial) microsimulation (O'Donoghue, 2001). These techniques have in common that they all produce a dataset which is rectangular without missing values created by substitution of missing information using an implicit or explicit statistical model. However, Haslett et al. (2010, p. 59) make clear that whatever technique is used:

- There are major benefits in the use of an explicit rather than an implicit statistical model.
- The structure of the underlying statistical model (e.g. linear or nonlinear, with or without random effects) needs to be determined on strong theoretical grounds.
- The model needs to be fitted and tested, and should explain a substantial part of the variation in most target variables.

It is not in the scope of this paper to discuss these methods, how they are related to each other and which method should be used under which conditions. All techniques show strong structural similarities with statistical matching which is the subject of another State of the Art paper.

References

[1] Al, P. & B.F.M. Bakker, (2000), *Re-engineering Social Statistics by micro-integration of different sources: An introduction*, In: P. Al & B.F.M. Bakker (eds.), Re-engeneering Social Statistics by micro-integration of different sources. (Themanummer) Netherlands Official Statistics, Vol. 15, nr. summer, pp.. 4–6.

[2] Arts, K., B.F.M. Bakker & E. van Lith, (2000), *Linking administrative registers and household surveys*, In: P. Al & B.F.M. Bakker (red.), Re-engineering Social Statistics by micro-integration of different sources. Themanummer Netherlands Official Statistics, jrg. 15, nr. summer, blz. 16–22.

[3] Auriat, N., (1991), *Who forgets? An analysis of memory effects in a retrospective survey on migration history*, European Journal of Population, vol. 7, nr. 4, pp. 311–342.

[4] Auriat, N., (1993), *My wife knows best. A comparison of event dating accuracy between the wife, the husband, the couple and the Belgium population register*, Public Opinion Quarterly, vol. 57, nr. 2, pp. 165–177.

[5] Bakker, B.F.M., (2002), *Statistics Netherlands' Approach to Social Statistics: The Social Statistical Dataset*, OECD Statistics Newsletter, vol. 2002, nr. 11, blz. 4–6.

[6] Bakker, Bart F.M., (2009a), *Micro-integratie. Statistische Methoden (09001)* (Den Haag/Heerlen: Statistics Netherlands) (In Dutch).

[7] Bakker, Bart F.M., (2009b), *Trek alle registers open!* (Amsterdam: Free University Press) (In Dutch).

[8] Bakker, Bart F.M., Frank Linder en Dominique van Roon (2008), *Could that be true? Methodological issues when deriving educational attainment from administrative datasources and surveys* (Shanghai: Paper prepared for the IAOS Conference on Reshaping Official Statistics, 14–16 October 2008).

[9] Belson, W.A., (1986), *Validity in survey research* (Brookfield: Gower).

[10] Bishop, Y.M.M., S.E Fienberg & P.W. Holland, (1975), *Discrete multivariate analysis: theory and practice* (Cambridge Massachusetts: MIT Press).

[11] Blom, M., J. Oudhof, R.V. Bijl & B.F.M. Bakker (red.), (2005), *Verdacht van criminaliteit. Allochtonen en autochtonen nader bekeken.* WODC-cahier 2005-2 (Den Haag/Voorburg: WODC/CBS) (in Dutch).

[12] Boonstra, H.J., (2004), *A simulation study of repeated weighting estimation. Discussion paper 04003*, Statistics Netherlands, Voorburg / Heerlen.

[13] Brick, J.M., R. McGuinness, S.J. Lapham, M. Cahalan, D. Owens & L. Gray, (1995), *Interviewer variance in two telephone surveys*, American Statistical Association: Proceedings of the Section on Survey Research Methods, blz. 447–452.

[14] Czaja, R. & J. Blair, (2005), *Designing surveys. A guide to decisions and procedures* (Three Oakes/ London/ New Dehli: SAGE).

[15] Denk, M. & P. Hackl, (2003), *Data integration and record matching: an Austrian contribution to research in official statistics*, Austrian Journal of Statistics, vol. 32, nr. 4, pp. 305–321.

[16] De Waal, T., (2000), *A brief overview of imputation methods applied at Statistics Netherlands*, Netherlands Official Statistics, vol. 15, autumn, pp. 23–27.

[17] Elbers, C., J. Lanjouw & P. Lanjouw, (2003), *Micro-level estimation of poverty and inequality*, Econometrica, vol. 71, nr. X, pp. 355–364.

[18] Fellegi & Sunter, (1969), *A theory of record linkage*, Journal of the American Statistical Association, jrg. 64, blz. 1183–1210.

[19] Ganninger, M., S. Häder & S. Gabler, (2007), *Design effects and interviewer effects in the European Social Survey: Where are we now and*

where do we want to go tomorrow? (Mannheim: Centre for Survey Research and Methodology).

[20] Gouweleeuw, J., & M. Hartgers, (2004), *The method of repeated weighting in the 2001 Census*, In: E. Schulte Nordholt, M. Hartgers & R. Gircour (eds.), The Dutch Virtual Census of 2001 (Voorburg/Heerlen: Statistics Netherlands), pp. 261–276.

[21] Groves, R.M., F.J. Fowler jr., M.P. Couper, J.M. Lepkowski, E. Singer, & R. Tourangeau, (2004), Survey Methodology (New York: Wiley Interscience).

[22] Grünewald, W. & T. Körner, (2005), *Quality on its way to maturity: results of the European conference on Quality and methodology in Official Statistics (Q2004)*, Journal of Official Statistics, vol. 21, nr. 4, pp. 747–759.

[23] Haslett, S., G. Jones, A. Noble & D. Ballas, (2010), *More for Less. Using statistical modelling to combine existing data sources to produce sounder, more detailed and less expensive Official Statistics*, Official Statistics Research Series, 6. Available from www.statisphere.govt.nz/osresearch.

[24] Houbiers, M., P. Knottnerus, A.H. Kroese, R.H. Renssen and V. Snijders, (2003), *Estimating consistent table sets: position paper on repeated weighting. Discussion paper 03005*, (Voorburg / Heerlen: Statistics Netherlands) Available from www.cbs.nl/en/publications/articles/general/discussion-papers/discussion-paper-03005.pdf.

[25] Houbiers, M., (2004), *Towards a Social Statistical Database and unified estimates at Statistics Netherlands*, Journal of Official Statistics, vol. 20, nr. 1, pp. 55–75.

[26] International Labour Organization, (2007), *Resolution concerning updating the International Standard Classification of Occupations* (Geneva: ILO).

[27] Jöreskog, K. & D. Sörbom, (1996), *LISREL 8. User's reference guide* (Chicago: Scientific Software International).

[28] Kapteyn, A. & J.Y. Ypma, (2007), *Measurement error and misclassification: a comparison of survey and administrative data*, Journal of Labor Economics, vol. 25, nr. 3, pp. 513–551.

[29] Kline, R.B., (2005), *Psycological testing:a practical approach to design and evaluation* (New York: SAGE).

[30] Knottnerus, P. and C. van Duin, (2006), *Variances in repeated weighting with an application to the Dutch Labour Force Survey*, Journal of Official Statistics, vol. 22, nr. 3, pp. 565–584.

[31] Kovar, J.G., & P.J. Whitridge, (1995), *Imputation of business survey data*, In: B. Cox, D.A. Binder, B. Nanjamma Chinnappa, A. Christianson, M.J. College & P.S. Kott (eds.), Business survey methods (New York: John WEiley and Sons).

[32] Kroese, A.H., Renssen R.H. en M. Trijssenaar, (2000), *Weighting or imputation: constructing a consistent set of estimates based on data from different sources*, In: P. Al en B.F.M. Bakker (red.), Re-engeneering social statistics by micro-integration of different sources, Netherlands Official Statistics, vol. 15, nr. Summer, pp. 23–31.

[33] O'Donoghue, (2001), *Dynamic microsimulation: a methodological survey*, Brazilian Electronic Journal of Economics, vol. 4, nr. 2, available from: www.beje.decon.ufpe.br/v4n2/cathal.htm.

[34] Pannekoek, J., (1988), *Interviewer variance in a telephone survey*, Journal of Official Statistics, jrg. 4, nr. 4, blz. 375–384.

[35] Renssen, R.R., A.H. Kroese & A. Willeboordse, (2001), *Aligning estimates by repeated weighting*, Research paper 491-01-TMO (Voorburg/Heerlen: Statistics Netherlands).

[36] Schroots, J.J.F., C. van Dijkum, M.H.J. Assink, (2004), *Autobiographical memory form a life span perspective*, The International Journal of Aging and Human Development, vol. 58, nr. 1, pp.99–115.

[37] Schulte Nordholt, E., M. Hartgers & R. Gircour (Eds.), (2004), *The Dutch Virtual Census of 2001*, Analysis and Methodology (Voorburg / Heerlen: Statistics Netherlands) available from www.cbs.nl/en-GB/menu/themas/dossiers/volkstellingen/publicaties/2001-b57-epub.

[38] Schulte Nordholt, Eric & Frank Linder, (2007), *Record matching for Census purposes in the Netherlands*, Statistical Journal of the IAOS, vol. 24, pp. 163 –171.

[39] Smith, J.P., & Th. Duncan, (2003), *Remembrances of things past: test-retest reliability of retrospective migration histories*, Journal of the Royal Statistical Society, vol. 166, nr. 1, pp. 23–49.

[40] Sikkel, D., (1988), *Quality aspects of statistical data collection* (Amsterdam: Sociometric Research Foundation).

[41] Statistics Denmark, (1995), *Statistics on Persons in Denmark – A register-based statistical system* (Luxembourg: Eurostat).

[42] Statistics Finland, (2004), *Register based statistics. Best Practices* (Helsinki: SF).

[43] Stoop, I.A.L., (2005), *The hunt for the last respondent. Nonresponse in sample surveys* (Den Haag: SCP).

[44] Heijden, P.G.M., G. van Gils, M. Cruijff en Dave Hessen, (2006), *Een schatting van het aantal in Nederland verblijvende illegale vreemdelingen in 2005* (Utrecht: IOPS Universiteit Utrecht) (In Dutch).

[45] Van der Laan, P., (2000), *Integrating administrative registers and household surveys*, In: P. Al en B.F.M. Bakker (eds.), Re-engeneering social statistics by micro-integration of different sources, Netherlands Official Statistics, vol. 15, nr. Summer, pp.7–15.

[46] Van der Velden, R., (2003), *SSB: wensdroom of nachtmerrie*, In: B. Bakker & L. Putman, De virtuele Volkstelling en het Sociaal Statistisch Bestand (Amsterdam: SISWO) [in Dutch].

[47] Van Duin, C. & V. Snijders, (2003), *Simulation studies of repeated weighting*, Discussion paper 03008, Statistics Netherlands, Voorburg / Heerlen. Available from www.cbs.nl/en/publications/articles/general/discussion-papers/discussion-paper-03008.pdf.

[48] Wallgren, A., & B. Wallgren, (2007), *Register-based Statistics: Administrative Data for Statistical Purposes*, Wiley Series in Survey Methodology (New York: Wiley).

Bibliography on Record Linkage

[1] Abowd J.M. and Woodcock S.D. (2001) *Disclosure limitation in longitudinal linked data*, In: Doyle P., Lane J.I., Theeuwes J.J., and Zayatz L.V. (eds.) Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies (Chapter 10): 215–278. North-Holland, Amsterdam, Netherlands.

[2] Adams M.M. et al. (1997) *Constructing reproductive histories by linking vital records*, American Journal of Epidemiology 1997;145: 339–48.

[3] Agrawal R., Bayardo R., Faloutsos C., Kiernan J., Rantzau R. and Srikant R. (2004) *Auditing Compliance with a Hippocratic Database*, Proceedings of the 30th International Conference on Very Large Databases (VLDB 2004), Toronto, Canada.

[4] Agrawal R., Evfimievski A.V., Kiernan, J. and Velu R. (2007) *Auditing Disclosure by Relevance Ranking*, In:26th ACM SIGMOD International Conference on Management of Data. Session on Database privacy and security:79-90. Beijing, China. doi: 10.1145/1247480.1247491

[5] Aizawa A., Oyama K. (2005) *A fast linkage detection scheme for multi-source information integration*, Web Information Retrieval and Integration (WIRI'05):30-39. IEEE CS,Tokio. doi: 10.1109/WIRI.2005.2. ISBN: 0-7695-2414-1.

[6] Alleva G., Fortini M., Tancredi A. (2007) *The Control of Non-Sampling Errors on Linked Data: an Application on Population Census*, SIS 2007 - Proceedings of the 2007 intermediate conference. Risk and Prediction. Venice. ISBN:88-7178-791-9. [Link #2].

Papers can be accessed via the links in the electronic version of the document. Additional references, comments, corrections or requests can be sent to essnet.di@istat.it , scanu@istat.it or mguigo@ine.es *(last updated: 5-12-2011)*

[7] Alvey, W. and Jamerson, B. (eds.) (1997) *Record Linkage Techniques – 1997*, Proceedings of an International Record Linkage Workshop and Exposition on March 20-21, 1997, Arlington, Virginia. Federal Committee on Statistical Methodology, Washington, DC.

[8] Arellano, M.G. (1992) *Comment to Newcombe, H.B., Fair, M.E. and Lalonde P. The use of names for linking personal records*, Journal of the American Statistical Association, Volume 87:1204-1206. Reprinted in Record Linkage Techniques – 1997. Proceedings of an International Record Linkage Workshop and Exposition (Alvey, W. and Jamerson, B. (eds)):345-347. [See Newcombe, H.B., Fair, M.E. and Lalonde P. (1992)].

[9] Armstrong J. and Mayda J.E. (1992) *Estimation of record linkage models using dependent data*, Proceedings of the Survey Research Methods Section, American Statistical Association, 1992: 853–858.

[10] Armstrong J. and Mayda J.E. (1993) *Model-based estimation of record linkage error rates*, Survey Methodology, Volume 19: 137–147.

[11] Armstrong J. and Saleh M. (2000) *Weight estimation for large scale record linkage applications*, Proceedings of the section on Survey Research Methods, American Statistical Association, 2000: 1–10.

[12] Armstrong J., Block C. and Saleh M. (1999) *Record Linkage for Electoral Administration*, SSC Annual Meeting, June 1999 - Proceedings of the Survey Methods Section: 57–64. Statistical Society of Canada.

[13] Bartlett S., Krewski D., Wang Y. and Zielinski, J.M. (1993) *Evaluation of error rates in large scale computerized record linkage studies*, Survey Methodology, Volume 19: 3–12.

[14] Batini C. and Scannapieco, M. (2006) *Data quality: concept, methods and techniques*, Springer, Berlin.

[15] Baxter R., Christen P. and Churches T. (2003) *A Comparison of fast blocking methods for record linkage*, Proceedings of 9th ACM SIGKDD Workshop on Data Cleaning, Record Linkage, and Object Consolidation, August 2003:25-27, Washington, DC, USA.

[16] Beebe G.W. (1985) *Why are epidemiologists interested in matching algorithms?*, Record Linkage Techniques – 1985. Proceedings of the Workshop on Exact Matching Methodologies (Kills B. and Alvey W. (eds)): 139–143.

[17] Belin T.R. (1989) *A proposed improvement in computer matching techniques*, Proceedings of the Section on Survey Research Methods, American Statistical Association, 1989: 784–789.

[18] Belin T.R. (1993) *Evaluation of sources of variation in record linkage through a factorial experiment*, Survey Methodology, Volume 19: 13–29.

[19] Belin T.R. and Rubin D.B. (1990) *Calibration of errors in computer matching for Census undercount (with discussion)*, Proceedings of the Government Statistics Section of the American Statistical Association, 1990: 124–131.

[20] Belin T.R. and Rubin D.B. (1995) *A method for calibrating false-match rates in record linkage*, Journal of the American Statistical Association, Volume 90: 694–707.

[21] Belin T.R., Ishwaran H., Duan N., Berry S.H. and Kanouse D.E. (2005) *Identifying likely duplicates by record linkage in a survey of prostitutes*, In: Gelman A., Meng X.-L. (Eds.) Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives. Wiley Series in Probability and Statistics. Wiley, NY. doi: 10.1002/0470090456.

[22] Bhattacharya I., Getoor L. (2004) *Iterative Record Linkage for Cleaning and Integration*, ACM SIGMOD Workshop on Data Mining and Knowledge Discovery 2004, Paris, France.

[23] Bhattacharya I., Getoor L. (2006) *A Latent Dirichlet Model for Unsupervised Entity Resolution*, The 6th SIAM Conference on Data Mining (SIAM SDM), Bethesda, Maryland, April 2006.

[24] Bigelow W., Karlson T. and Beutel P. (1999) *Using Probabilistic Linkage to Merge Multiple Data Sources for Monitoring Population Health*, Association for Health Services Research Meeting 1999; 16: 4–5. Center for Health Systems Research and Analysis, Madison, Wisconsin, USA.

[25] Bilenko M. (2006) *Learnable Similarity Functions and Their Application to Record Linkage and Clustering*, PhD Thesis, Department of Computer Sciences, University of Texas at Austin, Austin, TX, August.

[26] Bilenko M., Kamath B., Mooney R.J. (2006) *Adaptive blocking: Learnig to scale up record linkage*, Proceedings of the Sixth IEEE International

Conference on Data Mining (ICDM-06), Hong Kong: 87–96. ACM, Edimbourgh.

[27] Bilenko M. and Mooney R.J. (2002) *Learning to Combine Trained Distance Metrics for Duplicates Detection in Databases*, Technical Report AI-02-296, University of Texas at Austin, Feb 2002.

[28] Bilenko M., Mooney, R., Cohen W., Ravikumar P. and Fienberg S. (2003) *Adaptive Name Matching in Information Integration*, IEEE Intelligent Systems 18(5): 16–23 (2003). IEEE Computer Society. doi:10.1109/MIS.2003.1234765.

[29] Bilke, A. and Naumann, F. (2005) *Schema Matching using Duplicates*, Proceedings of International Conference on Data Engineering - 05, Tokyo, Japan, 69–80.

[30] Blakely T. and Salmon C. (2002) *Probabilistic record linkage and a method to calculate the positive predictive value*, International journal of epidemiology. 31(6), pp. 1246–1252. International Epidemiological Association, Oxford. ISSN 0300-5771.

[31] Brenner H., Schmidtmann I., Stegmaier, C. (1997) *Effects of record linkage errors on registry-based follow-up studies*, Statistics in Medicine, Vol 16, No 23:2633-2643. Wiley. doi:10.1002/(SICI)1097-0258(19971215)16:23<2633::AID-SIM702>3.0.CO; 2-1.

[32] Brenner H., Schmidtmann I. (1998) *Effects of record linkage errors on disease registration*, Methods of information in medicine. 1998, vol. 37(1): 69–74. Schattauer, Stuttgart, Germany. ISSN 0026-1270.

[33] Broadbent K. and Iwig W. (1999) *Record Linkage at NASS using AutoMatch*, 1999 FCSM Research Conference.

[34] Campbell K.M. (2005) *Rule Your Data with The Link King (a SAS/AF application for record linkage and unduplication)*, Proceedings of the SAS Users Group International SUGI 30: 020–030. SAS, Philadelphia.

[35] Campbell K.M., Deck D. and Krupski A. (2008) *Record Linkage Software in the Public Domain: A Comparison of Link Plus, The Link King, and a "Basic" Deterministic Algorithm*, Health Informatics Journal, Vol 14(1): 5–15. Sage Publications Ltd. doi:10.1177/1460458208088855.

[36] Cella P., Cibella N., Tuoto T. (2006) *Evaluating Matching Errors: an Application to the Coverage Rate Estimation of the Italian Agricultural Census*, Proceedings of the European Conference on Quality and Methodology in Official Statistics, Q2004 (cd-rom), Mainz, Germany, 24–26 May 2004.

[37] Chaudhuri, S., Gamjam, K., Ganti, V., and Motwani, R. (2003) *Robust and Efficient Match for On-Line Data Cleaning*, ACM SIGMOD '03, 313–324.

[38] Chernoff H. (1980) *The identification of an element of a large population in the presence of noise*, Annals of Statistics, Volume 8(6): 1179–1197.

[39] Chesher A. and Nesheim L. (2006) *Review of the Literature on the Statistical Properties of Linked Datasets*, DTI Occasional Paper No. 3. Department for Business, Enterprise and Regulatory Reform, London.

[40] Christen P. (2007a) *Improving data linkage and deduplication quality through nearest-neighbour based blocking*, Submitted to the thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'07).

[41] Christen P. (2007b) *A two-step classification approach to unsupervised record linkage*, AusDM'07, Conferences in Research and Practice in Information Technology (CRPIT), Vol. 70: 111–119. Gold Coast, Australia.

[42] Christen P. (2007c) *Towards Parameter-free Blocking for Scalable Record Linkage*, Joint Computer Science Technical Report Series TR-CS-07-03. The Australia National University, Canberra, Australia. doi:10.1.1.73.3454.

[43] Christen P. (2008a) *Automatic training example selection for scalable unsupervised record linkage*, PAKDD'08, Springer LNAI 5012:511–518, Osaka, Japan, 2008.

[44] Christen P. (2008b) *Automatic record linkage using seeded nearest neighbor and support vector machine classification*, ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD'08): 151–159. Las Vegas, Nevada, USA. doi:10.1.1.133.1107.

[45] Christen P. (2008c) *Febrl - a freely available record linkage system with a graphical user interface*, HDKM'08, Conferences in Research and

Practice in Information Technology (CRPIT), Vol. 80. Wollongong, Australia.

[46] Christen P. (2008d) *Febrl - an open source data cleaning, deduplication and record linkage system with a graphical user interface*, ACM International Conference on Knowledge Discovery and Data Mining 2008 (SIGKDD'08):1065-1068. Las Vegas, Nevada, USA.

[47] Christen P. (2009) *Development and User Experiences of an Open Source Data Cleaning, Deduplication and Record Linkage System*, ACM SIGKDD Explorations Newsletter, Vol.11(1): 39–48.

[48] Christen P. and Churches T. (2004) *Blind Data Linkage using n-gram Similarity Comparisons*, Proceedings of the 8th PAKDD'04 (Pacific-Asia Conference on Knowledge Discovery and Data Mining), Sydney. Springer Lecture Notes in Artificial Intelligence, (3056).

[49] Christen P. and Churches T. (2005a) *A Probabilistic Deduplication, Record Linkage and Geocoding System*, Proceedings of the ARC Health Data Mining workshop, University of South Australia, April 2005.

[50] Christen P. and Churches T. (2005b) *Febrl: Freely extensible biomedical record linkage Manual. release 0.3 edition*, Technical Report Computer Science Technical Reports no.TR-CS-02-05, Department of Computer Science, FEIT, Australian National University, Canberra.

[51] Christen P., Churches T. and Hegland M. (2004) *A Parallel Open Source Data Linkage System*, Proc of The Eighth Pacific-Asia Conference on Knowledge Discovery and Data Mining, Sydney.

[52] Christen P., Churches T., and Zhu J.X. (2002) *Probabilistic Name and Address Cleaning and Standardization*, Presented at the Australasian Data Mining Workshop, Canberra.

[53] Christen P., Churches T., Lim K. and Zhu J.X (2002) *Preparation of name and address data for record linkage using hidden Markov models*, at BioMed Central Medical Informatics and Decision Making.

[54] Christen P. et al. (2002a) *Parallel Computing Techniques for High-Performance Probabilistic Record Linkage*, Proceedings of the Symposium on Health Data Linkage, Sydney.

[55] Christen P. et al. (2002b) *High-Performance Computing Techniques for Record Linkage*, Proceedings of the Australian Health Outcomes Conference (AHOC-2002), Canberra.

[56] Christen P., Gayler R. (2008) *Towards scalable real-time entity resolution using a similarity-aware inverted index approach*, AusDM'08, CRPIT vol. 87. Glenelg, Australia.

[57] Christen P., Gayler R., Hawking D. (2009) *Similarity-aware indexing for real-time entity resolution*, Technical Report TR-CS-09-01. School of Computer Science, The Australian National University, Canberra, Australia. doi:10.1145/1645953.1646173.

[58] Christen P., Goiser K. (2007) *Quality and complexity measures for data linkage and deduplication*, In: F. Guillet and H. Hamilton (eds.) Quality Measures in Data Mining, volume 43 of Studies in Computational Intelligence. Springer.

[59] Cibella N., Fernández G.-L., Fortini M., Guigó M., Hernández F., Scannapieco M., Tosco L., Tuoto T. (2009) *Sharing Solutions for Record Linkage: the RELAIS Software and the Italian and Spanish Experiences*, Proc. Of the NTTS (New Techniques and Technologies for Statistics) Conference, Bruxelles, Belgium, 2009.

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

[61] Cohen W. and Richman J. (2002) *Learning to Match and Cluster Large High-Dimensional Data Sets for Data Integration*, Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD).

[62] Committee on State Voter Registration Databases. National Research Council (2010) *Improving State Voter Registration Databases Final Report*, National Academies Press, Washington DC.

[63] Conigliani C., Tancredi A. (2006) *Comparing parametric and semi-parametric approaches for Bayesian cost-effectiveness analyses*, Departmental Working Papers of Economics 0064. Department of Economics - University Roma Tre.

[64] Copas J. R. and Hilton F. J. (1990) *Record linkage: statistical models for matching computer records*, Journal of the Royal Statistical Society, A, Volume 153: 287–320. doi:10.2307/2982975

[65] Day C. (1997) *A checklist for evaluating record linkage software*, In:Alvey, W. and Jamerson, B. (eds.) (1997) Record Linkage Techniques, Washington, DC: Federal Committee on Statistical Methodology.

[66] Dempster A.P., Laird N.M. and Rubin D.B. (1977) *Maximum Likelihood From Incomplete Data via the EM Algorithm*, Journal of the Royal Statistical Society Series B, 39(1): 1–38. doi:10.1.1.133.4884

[67] Denk M. and Hackl P. (2003) *Data integration and record matching: an Austrian contribution to research in official statistics*, Austrian Journal of Statistics, Volume 32, 305–321.

[68] DuVall S.L., Kerber R.A., Thomas A. (2010) *Extending the Fellegi-Sunter probabilistic record linkage method for approximate field comparators*, Journal of Biomedical Informatics, Volume 43, Issue 1, February 2010: 24–30, ISSN 1532-0464. doi:10.1016/j.jbi.2009.08.004.

[69] Domingo-Ferrer J. and Torra V. (2002) *Validating distance-based record linkage with probabilistic record linkage*, Lecture Notes in Computer Science. Vol. 2504 (Topics in Artificial Intelligence): 207–215. Springer, Berlin / Heidelberg. ISBN: 978-3-540-00011-2. ISSN: 0302-9743. doi: 10.1007/3-540-36079-4

[70] Domingo-Ferrer J. and Torra V. (2003a) *On the connections between statistical disclosure control for microdata and some artificial intelligence tools*, Information Sciences, Vol. 151: 153–170. ISSN: 0020-0255.

[71] Domingo-Ferrer J. and Torra V. (2003b) *Disclosure risk assessment in statistical microdata protection via advanced record linkage*, Statistics and Computing, Vol. 13(4): 343–354. Kluwer Academic, Hingham, MA, USA. ISSN: 0960-3174. doi:10.1023/A:1025666923033.

[72] Domingo-Ferrer J. and Torra V. (2004) *Disclosure risk assessment in statistical data protection*, Journal of Computational and Applied Mathematics, Vol. 164: 285–293. ISSN: 0377-0427.

[73] Elamir E. A.H., Skinner C. (2006) *Record level measures of disclosure risk for survey microdata*, Journal of Official Statistics, 22(3): 525–539.

[74] Elfeky M.G., Verykios V.S. and Elmagarmid A.K. (2002) *TAILOR: A Record Linkage Toolbox*, Proceedings of the 18th International Conference on Data Engineering (ICDE'02), pp.0017. IEEE.

[75] Elfeky M.G., Verykios V.S., Elmagarmid A.K., Ghanem T.M. and Huwait A.R. (2003) *Record Linkage: A Machine Learning Approach, a Toolbox, and a Digital Government Web Service*, Department of Computer Sciences, Purdue University, Technical Report CSD-TR 03-024.

[76] Elmagarmid A.K., Ipeirotis P.G., Verykios V.S. (2006) *Duplicate Record Detection: A survey*, Technical report CeDER-06-05, Stern School of Business, New York University.

[77] Elmagarmid A.K., Ipeirotis P.G., Verykios V.S. (2007) *Duplicate Record Detection: A survey*, IEEE Transactions on Knowledge and Data Engineering (TKDE), 19(1): 1–16. doi:10.1109/TKDE.2007.9.

[78] Hundepool A. et al. (2010) *Handbook on Statistical Disclosure Control*, Version 1.2. ESSnet SDC.

[79] Hundepool A. et al. (2010b1) *Illustrative record linkage example*, ESSnet SDC.

[80] Eurostat (ed.) (2009) *Insights on Data Integration Methodologies*, Proceedings of ESSnet-ISAD Workshop, Vienna 2008. Methodologies and Working Papers. Eurostat, Luxembourg. ISBN 978-92-79-12306-1, ISSN 1977-0375. doi:10.2785/20079.

[81] ESSnet ISAD (2008) *Report of WP1: State of the art on statistical methodologies for integration of surveys and administrative data*, Istat – CBS – CzSO – INE – STAT.

[82] Fair M. (2004) *Generalized Record Linkage System (GRLS) – Statistics Canada's Record Linkage Software*, Austrian Journal of Statistics, Volume 33, Number 1&2: 37–53.

[83] Fellegi I.P. and Sunter A.B. (1969) *A theory for record linkage*, Journal of the American Statistical Association, Volume 64, 1183–1210.

[84] Fellegi I.P. (1997) *Record linkage and public policy - A dynamic evolution*, Record Linkage Techniques – 1997. Proceedings of an International Record Linkage Workshop and Exposition (Alvey, W. and Jamerson, B. (eds)): 3–12.

[85] Fienberg S.E., Manrique-Vallier D. (2008) *Integrated methodology for multiple systems estimation and record linkage using a missing data formulation*, Advances in Statistical Analysis, Volume 93, Number 1:49–60. Springer-Verlag, Berlin / Heidelberg. ISSN: 1863-818X. doi:10.1007/s10182-008-0084-z.

[86] Fortini M., Liseo B., Nuccitelli A. and Scanu M. (2000) *Bayesian approaches to record linkage*, Technical report Universitŕ degli Studi di Roma ”La Sapienza”, Dipartimento di studi geoeconomici, statistici, storici per l’analisi regionale, working paper n. 15.

[87] Fortini M., Liseo B., Nuccitelli A. and Scanu M. (2001) *On Bayesian record linkage*, Research in Official Statistics, Volume 4:185-198. Published also in Monographs of Official Statistics, Bayesian Methods (E. George (ed.)), EUROSTAT, pp. 155–164.

[88] Fortini M., Nuccitelli A., Liseo B. and Scanu M. (2002) *Modelling issues in record linkage: a Bayesian perspective*, Proceedings of the Section on Survey Research Methods, American Statistical Association: 1008–1013.

[89] Fortini M., Scannapieco M., Tosco L. and Tuoto T. (2006) *Towards an Open Source Toolkit for Building Record Linkage Workflows*, Proceedings SIGMOD 2006 Workshop on Information Quality in Information Systems (IQIS’06), Chicago, USA.

[90] Gill L. (2001) *Methods for Automatic Record Matching and Linking and their use in National Statistics*, National Statistics Methodological Series No. 25. National Statistics, London.

[91] Goiser K. and Christen P. (2006) *Towards Automated Record Linkage*, Proceedings of the Fifth Australasian Data Mining Conference (AusDM2006), Sydney.

[92] Goldacre M.J., Wotto C.J., David Yeates D., Seagroatt V., Flint J. (2007) *Cancer in people with depression or anxiety: record-linkage study*, Social Psychiatry and Psychiatric Epidemiology, Springer Berlin / Heidelberg, Volume 42, Number 9.

[93] Gomatam S. et al. (2002) *An empirical comparison of record linkage procedures*, Statistics in Medicine, Volume 21 Issue 10: 1485–1496. Wiley, New York. doi:10.1002/sim.1147

[94] Gu L. and Baxter R. (2004) *Adaptive Filtering for Efficient Record Linkage*, SIAM Int. Conf. on Data Mining, April 22–24, Orlando, Florida.

[95] Gu L., Baxter R., Vickers D. and C. Rainsford C. (2003) *Record linkage: Current practice and future directions*, Technical Report 03/83, CSIRO Mathematical and Information Sciences, Canberra. doi:10.1.1.14.8119.

[96] Guerrero J.N. (2009) *Contributions to Record Linkage for Disclosure Risk Assessment*, Ph.D. Thesis. IIIA Series, volume 40. IIIA-CSIC, Bellaterra, Spain. ISBN: 978-84-00-08848-4, 2009.

[97] Guisado-Gamez J., Prat-Perez A., Nin J., Muntes-Mulero V., Larriba-Pey J. (2008) *Parallelizing Record Linkage for Disclosure Risk Assessment*, Privacy in Statistical Databases (PSD), volume 5262 of Lecture Notes in Computer Science:190-202. Springer.

[98] Harville, D. and Moore, R. (1999) *Determining record linkage parameters using an iterative logistic regression approach*, Proceedings of the Section on Survey Research Methods, American Statistical Association.

[99] Haslinger A. (2004) *Data matching for the maintenance of the business register of Statistik Austria*, Austrian Journal of Statistics, Volume 33: 55–67.

[100] Heasman D. (2008) *Record linkage development in the Office for National Statistics*, Office for National Statistics, London.

[101] Hernandez M. and Stolfo S. (1995) *The Merge/Purge Problem for Large Databases*, Proc. of 1995 ACT SIGMOD Conf., pages 127–138.

[102] 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).

[103] Herzog T. (2004) *Playing with Matches: Exploring Data Quality Issues With an Emphasis on Record Linkage Techniques*, SOA 2004 New York Annual Meeting - 18TS, Playing With Matches: Applications of Record Linkage Techniques and Other Data Quality Procedures.

[104] Herzog T.N., Scheuren F.J. and Winkler W.E. (2007) *Data Quality and Record Linkage Techniques*, Springer Science+Business Media, New York.

[105] Herzog T.N., Scheuren F.J. and Winkler W.E. (2010) *Record Linkage*, Wiley Interdisciplinary Reviews Computational Statistics, Volume 2, September /October 2010: 535–543, John Wiley & Sons, New York.

[106] Houbiers, M. (2004) *Towards a social statistical database and unified estimates at Statistics Netherlands*, Journal of Official Statistics, Volume 20, pp 55–75.

[107] Howe G.R. and Lindsay J. (1981) *A generalized iterative record linkage computer system for use in medical follow-up studies*, Computers and Biomedical Research, 14, 327–340.

[108] Jabine T.B. and Scheuren F.J. (1986) *Record linkages for statistical purposes: methodological issues*, Journal of Official Statistics, Volume 2, 255–277.

[109] Jaro M.A. (1972) *UNIMATCH: a computer system for generalized record linkage under conditions of uncertainty*, Proceedings of the 1972 Spring Joint Computer Conference, Session: The computer in government: a tool for change: 523–530. American Federation of Information Processing Societies, New York.

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

[111] Jaro M.A. (1995a) *Probabilistic linkage of large public health data files*, Statistics in Medicine, Volume 14: 491–498.

[112] Jaro M.A. (1995b) *Record linkage research and the calibration of record linkage algorithms*, Bureau of the Census, Statistical Research Division, Statistical Research Report Series, n.RR84/27.

[113] Jurczyk P. (2009) *FRIL: Fine-grained Record Integration and Linkage Tool – Tutorial*, Birth Defects Research (Part A): Clinical and Molecular Teratology 82:822–829. Emory University, Math&CS Department.

[114] Karakasidis A., Verykios V.S. (2009) *Privacy Preserving Record Linkage Using Phonetic Codes*, Proceedings of the 4th Balkan Conference in Informatics:101-106. IEEE Computer Society, Thessaloniki, Greece. ISBN: 978-0-7695-3783-2. doi: 10.1109/BCI.2009.29.

[115] Kelley R.P. (1983) *A preliminary study of the error structure of statistical matching*, Proceedings of the Social Statistics Section, American Statistical Association, pp. 206–208.

[116] Kelley R.P. (1984) *Blocking considerations for record linkage under conditions of uncertainty*, Proceedings of the Social Statistics Section, American Statistical Association, pp. 602-605. (Also available in Statistical Research Division Report Series, SRD Research Report No. RR-84/19. Bureau of the Census, Washington. D.C.).

[117] Kelley R.P. (1985) *Advances in record linkage methodology: a method for determining the best blocking strategy*, Record Linkage Techniques – 1985. Proceedings of the Workshop on Exact Matching Methodologies (Kills, B. and Alvey, W. (eds)): 199–203.

[118] Kelley R.P. (1986) *Robustness of Census Bureau's record linkage system*, Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 620–624.

[119] Kennickell A. (1997) *Multiple Imputation and Disclosure Protection*, Record Linkage Techniques – 1997. Proceedings of an International Record Linkage Workshop and Exposition (Alvey, W. and Jamerson, B. (eds)): 248–267.

[120] Kestenbaum B. (1996) *Probability Linkage Using Social Security Administration Files*, Working Paper form the Social Security Administration. Proceedings of the '96 Annual Research Conference, Concurrent session XI-A.

[121] Kills B. and Alvey W. (Comp. And Eds.) (1985) *Record Linkage Techniques – 1985*, Proceedings of the Workshop on Exact Matching Methodologies, Arlington, Virginia, 9–15 May 1985. Internal Revenue Service, Washington D.C.

[122] Kirkendall N.J. (1985) *Weights in computer matching: applications and an information theoretic point of view*, Record Linkage Techniques – 1985. Proceedings of the Workshop on Exact Matching Methodologies, (Kills, B. and Alvey, W. (eds)): 189–197.

[123] Konold M., L'Assainato S. (2009) *Matching Business Data from Different Sources: The Case of the KombiFiD-Project in Germany*, Conference 'New Techniques and Technologies for Statistics (NTTS 2009)', Brussels.

[124] Lahiri P. and Larsen M.D. (2000) *Model-based analysis of records linked using mixture models*, Proceedings of the section on Survey Research Methods, American Statistical Association, pp. 11–19.

[125] Lahiri P. and Larsen M.D. (2005) *Regression Analysis With Linked Data*, Journal of the American Statistical Association, 100(469):222–230. doi:10.1198/016214504000001277.

[126] Lambert D. (1993) *Measures of Disclosure Risk and Harm*, Journal of Official Statistics, Vol. 9, 313–331.

[127] Larsen M.D. (1997) *Modeling Issues and the Use of Experience in Record Linkage*, Record Linkage Techniques – 1997. Proceedings of an International Record Linkage Workshop and Exposition (Alvey, W. and Jamerson, B. (eds)): 95–105.

[128] Larsen M.D. (1999) *Multiple imputation analysis of records linked using mixture models*, Proceedings of the Survey Methods Section, Statistical Society of Canada, pp. 65–71.

[129] Larsen M.D. (2001a) *Methods for model-based record linkage and analysis of linked files*, Proceedings of the Annual Meeting of the American Statistical Association, Mira Digital publishing, Atlanta.

[130] Larsen M.D. (2001b) *Record linkage using finite mixture models*, Book chapter in Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives, A. Gelman and X-L. Meng, editors. pp. 309–318.

[131] Larsen M.D. (2002) *Comments on hierarchical bayesian record linkage*, Proceedings of the American Statistical Association, Section on Bayesian Statistical Science. Alexandria, US.

[132] Larsen M.D. (2004) *Record Linkage of Administrative Files and Analysis of Linked Files*, IMS-ASA's SRMS Joint Mini-Meeting on Current Trends in Survey Sampling and Official Statistics. The Ffort Radisson, Raichak, West Bengal, India.

[133] Larsen M.D. (2005a) *Hierarchical Bayesian Record Linkage Theory*, Iowa State University, Department of Statistics, Preprint #05-03, August 31.

[134] Larsen M.D. (2005b) *Advances in Record Linkage Theory: Hierarchical Bayesian Record Linkage Theory*, 2005 Proceedings of the American Statistical Association, Survey Research Methods Section [CD-ROM], pp. 3277–3284. Alexandria, VA: American Statistical Association.

[135] Larsen M.D. and Rubin D.B. (2001) *Iterative automated record linkage using mixture models*, Journal of the American Statistical Association, 96, 32–41.

[136] Latanya S. (1997) *Computational Disclosure Control for Medical Microdata: The Datafly System*, Record Linkage Techniques – 1997. Proceedings of an International Record Linkage Workshop and Exposition (Alvey, W. and Jamerson, B. (eds)): 95–105.

[137] Levenshtein V. (1966) *Binary codes capable of correcting deletions, insertions, and reversals*, Soviet Physics Doklady 10(8): 707–710. For an available reference, see Navarro, G. (2001).

[138] Li C., Jin L., Mehrotra, S. (2006) *Supporting Efficient Record Linkage for Large Data Sets Using Mapping Techniques*, World Wide Web, 9(4): 557–584. Kluwer Academic Publishers, Hingham (MA), USA. ISSN:1386-145X.

[139] Liseo B., Tancredi, A. (2004) *Statistical inference for data files that are computer linked*, Proceedings of the International Workshop on Statistical Modelling- Firenze: 224–228. Univ. Press.

[140] Liseo B., Tancredi A. (2009a) *Bayesian estimation of population size via linkage of multivariate Normal data sets*, Technical Report Universitŕ La Sapienza, Dipartimento di Studi Geoeconomici, Linguistici, Statistici e Storici per l'Analisi regionale. Working papers, Vol.66. Roma.

[141] Liseo B., Tancredi A. (2009b) *Model based record linkage: a Bayesian perspective*, Eurostat (ed.) Insights on Data Integration Methodologies. Proceedings of ESSnet-ISAD Workshop, Vienna 2008. Methodologies and Working Papers. Eurostat, Luxembourg.

[142] Maggi F. (2008) *A Survey of Probabilistic Record Matching Models, Techniques and Tools*, Cycle XXII Scientific Report TR-2008-22, Advanced Topics in Information Systems B: 1–20. DEI, Politecnico di Milano, Doctoral Program in Information Technology.

[143] Malin B., Sweeney L. and Newton E. (2003) *Trail Re-Identification: Learning Who You Are From Where You Have Been*, Workshop on Privacy in Data, Carnegie-Mellon University, March 2003.

[144] McCallum A., Nigam K. and Ungar L. (2000a) *Efficient clustering of high-dimensional data sets with application to reference matching*, Proc. of the sixth ACM SIGKDD Int. Conf. on KDD, pages 169–178.

[145] McGlincy M.H. (2004) *A Bayesian Record Linkage Methodology for Multiple Imputation of Missing Links*, '04 Joint Statistical Meeting, American Statistical Association, Alexandria, US.

[146] Meray N., Reitsma J.B., Ravelli A.C., Bonsel G.J. (2007) *Probabilistic record linkage is a valid and transparent tool to combine databases*

without a patient identification number, The Journal of Clinical Epidemiology 60, pp. 883–891.

[147] Michelson M., Knoblock C.A. (2006) *Learning blocking schemes for record linkage*, National Conference on Artificial Intelligence (AAAI-2006). Boston.

[148] Monge A.E. (2000a) *Matching algorithm within a duplicate detection system*, IEEE Data Engineering Bulletin, 23(4), 14–20.

[149] Monge A.E. (2000b) *An Adaptive and Efficient Algorithm for Detecting Approximately Duplicate Database Records*,

[150] Monge A.E. and Elkan C. (1997) *An efficient domain-independent algorithm for detecting approximately duplicate database records*, The proceedings of the SIGMOD 1997 workshop on data mining and knowledge discovery, May 1997.

[151] Moustakides G.V., Verykios V.S. (2009) *Optimal Stopping: A Record-Linkage Approach*, Journal of Data and Information Quality, Volume 1, Issue 2 (September 2009), Article 9: 1–9. ACM, New York, USA. ISSN:1936-1955.

[152] Navarro G. (2001) *A guided tour to approximate string matching*, ACM Computing Surveys. Vol. 33 (1): 31–88. doi: doi=10.1.1.21.3112. other.

[153] Neiling M. (1998) *Data Fusion with Record Linkage*, In: I. Schmitt, C. Turker, E. Hildebrandt, and M. Hoding, editors, 3. Workshop 'Foederierte Datenbanken', Aachen, 1998. Shaker Verlag.

[154] Neiling M. and Lenz H.-J. (2000) *Data Fusion and Object Identification*, SSGRR2000, l'Aquila, Italy.

[155] Neiling M. and Muller R.M. (2001) *The good into the Pot, the bad into the Crop. Preselection of Record Pairs for Database Fusion*, Proc. of the First International Workshop on Database, Documents, and Information Fusion, Magdeburg, Germany.

[156] Neiling M., Jurk S., Lenz H.-J. and Naumann F. (2003) *Object Identification Quality*, International Workshop on Data Quality in Cooperative Information Systems, Siena, Italy.

[157] Neiling M., Schaal M. and Schumann M. (2002) *WrapIt, Automated Integration of Web Databases with Extensional Overlaps*, 2nd International Workshop of the Working Group "Web and Databases" of the

German Informatics Society (GI) (Workshop WebDB 2002), Erfurt, Thuringia, Germany.

[158] NeSmith N.P. (1997) *Record linkage and genealogical files*, Record Linkage Techniques – 1997. Proceedings of an International Record Linkage Workshop and Exposition (Alvey, W. and Jamerson, B. (eds)): 358–361.

[159] Neter, J., Maynes, E.S, and Ramanathan, R. (1965) *The effect of mismatching on the measurement of response errors*, Journal of the American Statistical Association, 60, 1005–1027.

[160] Newcombe H. B. and Kennedy J. M. (1962) *Record linkage, making maximum use of the discriminating power of Identifying information*, Communication of the Association for Computing Machinery, Volume 5(11), pp. 563–566.

[161] Newcombe H. B., Kennedy J.M., Axford S.J. and James A.P. (1959) *Automatic linkage of vital records*, Science, 130: 954–959.

[162] Newcombe H.B. (1988) *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*, Oxford Medical Publications, Oxford.

[163] Newcombe H.B. (1993) *Distinguishing Individual Linkages of Personal Records from Family Linkages*, Methods of Information in Medicine, Vol.32, No.5: 358–364.

[164] Newcombe H.B., Fair M.E. and Lalonde P. (1992) *The use of names for linking personal records (with discussion)*, Journal of the American Statistical Association, Volume 87:1193-1208. Reprinted in Record Linkage Techniques – 1997. Proceedings of an International Record Linkage Workshop and Exposition (Alvey, W. and Jamerson, B. (eds)): 335–349.

[165] Nin J., Herranz J., Torra V. (2007a) *On the Disclosure Risk of Multivariate Mi-croaggregation*, Data and Knowledge Engineering (DKE), Elsevier, volume 67, issue 3, pages 399-412. SCI index (2007): 1.144.

[166] Nin J., Torra V. (2007b) *Analysis of the Univariate Microaggregation Disclosure Risk*, New Generation Computing, Ohmsha-Springer, volume 27, pages 177- 194. SCI index (2007):0.854.

[167] Nygaard L. (1992) *Name standardization in record linkage: an improved algorithmic strategy*, History and Computing, 4 (1992): 63–74.

[168] On B.W., Lee D., Kang J., Mitra P. (2005) *Comparative Study of Name Disambiguation Problem using a Scalable Blocking-based Framework*, ACM/IEEE Joint Conf. on Digital Libraries (JCDL). Jun. 2005.

[169] Oropallo F. (2004) *Enterprise Microsimulation Models and Data Challenges: Preliminary Results from the Diecofis Project*, In: P.D. Falorsi, A. Pallara and A. Russo (Eds) 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 ammnistrative, Franco Angeli

[170] Oropallo F., Inglese F. (2003) *The Development of an Integrated and Systematized Information System for Economic and Policy Impact Analysis*, Austrian Journal of Statistics, Volume 33, pp 211–236.

[171] Polettini S. (2003) *Some remarks on the individual risk methodology*, Proceedings of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, 299–311.

[172] Porter E.H. and Winkler W.E. (1997a) *Approximate string comparison and its effect on advanced record linkage system*, Bureau of the Census, Statistical Research Division, Statistical Research Report Series, n. RR97/02.

[173] Porter E.H. and Winkler W.E. (1997b) *Approximate string comparison and its effect on advanced record linkage system*, Bureau of the Census, Statistical Research Division, Statistical Research Report Series, n. RR97/02.

[174] Ravikumar P. and Cohen W. (2004) *A Hierarchical Graphical Model for Record Linkage*, UAI 2004.

[175] Robinson-Cox J.F. (1998) *A record-linkage approach to imputation of missing data: Analyzing tag retention in a tag-recapture experiment*, Journal of Agricultural, Biological, and Environmental Statistics, Volume 3(1):48-61. International Biometric Society, Washington, DC.

[176] Scannapieco M., Cibella N., Tosco L., Tuoto T., Valentino L., Fortini M. (eds.) *RELAIS Versión 2.0 – User's Guide.*

[177] Scheuren F. and Alvey W. (1974) *Selected bibliography on the matching of person records from different sources*, Proceedings of the Social Statistics Section, American Statistical Association, pp. 151–154.

[178] Scheuren F. and Oh H.L. (1975) *Fiddling around with nonmatches and mismatches*, Proceedings of the Social Statistics Section, American Statistical Association, pp. 627–633.

[179] Scheuren F. and Winkler W.E. (1993) *Regression analysis of data files that are computer matched – Part I*, Survey Methodology, Volume 19, pp. 39–58.

[180] Scheuren F. and Winkler W.E. (1996a) *Recursive analysis of linked data files*, U.S. Bureau of the Census, Statistical Research Division Report Series, n.1996/08.

[181] Scheuren F. and Winkler W.E. (1996b) *Recursive Merging and Analysis of Administrative Lists and Data*, Proceedings of the Section of Government Statistics, American Statistical Association, 123–128.

[182] Scheuren F. and Winkler W.E. (1997) *Regression analysis of data files that are computer matched – Part II*, Survey Methodology, Volume 23, pp. 157-165.

[183] Schnell R., Bachteler T., Reiher J. (2009) *Privacy-preserving record linkage using Bloom filters*, BMC Medical Informatics and Decision Making 2009, 9:41. BioMed Central. doi:10.1186/1472-6947-9-41. [Link #2]

[184] Shlomo N. (2009) *Assessing Disclosure Risk under Misclassification for Microdata*, Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, WP43. Bilbao, Spain.

[185] Shlomo N., Skinner C. (2009) *Assessing The Disclosure Protection Provided by Misclassification for Survey Microdata*, Working Paper M09/14. Southampton Statistical Sciences Research Institute.

[186] Shkolnikov V. M., Jasilionis D., Andreev E. M., Jdanov D. A., Stankuniene V, Ambrozaitiene D. (2007) *Linked versus unlinked estimates of mortality and length of life by education and marital status: evidence from the first record linkage study in Lithuania*, Social Science and Medicine, Vol. 64, 1392–1406.

[187] Singla P. and Domingos P. (2004) *Multi-Relational Record Linkage*, Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (KDD'04): Workshop on Multi-Relational Data Mining:31-48. ACM Press, Seattle.

[188] Skinner C. (2008) *Assessing disclosure risk for record linkage*, In: Domingo-Ferrer J. and Saygin Y. (eds.) Privacy in Statistical Databases. Lecture Notes in Computer Science 5262/2008:166-176. Springer, Berlin/Heidelberg. ISBN: 978-3-540-87470-6. doi: 10.1007/978-3-540-87471-3_14.

[189] Skinner C., Shlomo N. (2008) *Assessing Identification Risk in Survey Microdata Using Log-Linear Models*, Journal of the American Statistical Association, Volume 103: 989–1001.

[190] Smith M.E. and Newcombe H.B. (1975) *Methods for computer linkage of hospital admission-separation records into cumulative health histories*, Methods of Information in Medicine, Volume 14 (July), pp. 118–125.

[191] Smith M.E. and Silins J. (1981) *Generalized iterative record linkage system*, Proceedings of the Social Statistics Section, American Statistical Association, pp. 128–137.

[192] Statistics New Zealand (2006) *Data integration manual*, Statistics New Zealand, Wellington.

[193] Steel P. and Konschnik C. (1997) *Post-Matching Administrative Record Linkage Between Sole Proprietorship Tax Returns and the Standard Statistical Establishment List*, Record Linkage Techniques 1997, Washington, DC: National Academy Press, 179–189.

[194] Suchindran C. M., Leiss J.K. and Salama I. (2001) *Alternative Methods for Record Linkage Application to Linking Vital Records*, Proceedings of the Annual Meeting of the American Statistical Association, August 5–9, 2001.

[195] Sweeney L. (1997) *Computational Disclosure Control for Medical Microdata*, Record Linkage Techniques – 1997. Proceedings of an International Record Linkage Workshop and Exposition (Alvey, W. and Jamerson, B. (eds)).

[196] Tancredi A. (2002) *Accounting for heavy tails in stochastic frontier models*, Working paper n. 2002.16. Dipartimento di Scienze Statistiche, Universitŕ di Padova.

[197] Tancredi A., Liseo B. (2005) *Bayesian inference for linked data*, Proceedings of the conference SCO 2006: 203–208.

[198] Tancredi A., Liseo B. (2011) *A hierarchical Bayesian approach to record linkage and population size problems*, Ann. Appl. Stat., Vol. 5, pp. 1553–1585.

[199] Tepping B.J. (1968) *A model for optimum linkage of records*, Journal of the American Statistical Association, Volume 63, pp. 1321–1332.

[200] Thibaudeau Y. (1989) *Fitting log-linear models when some dichotomous variables are unobservable*, Proceedings of the Section on statistical computing, American Statistical Association, pp. 283-288.

[201] Thibaudeau Y. (1992) *Identifying discriminatory models in record linkage*, Proceedings of the SurveyResearch Methods Section, American Statistical Association, pp. 835–840.

[202] Thibaudeau Y. (1993) *The discrimination power of dependency structures in record linkage*, Survey Methodology, Volume 19, pp. 31–38.

[203] Thomas B. (1999) *Probabilistic Record Linkage Software: A Statistics Canada Evaluation of GRLS and Automatch*, 1999 SSC Annual Meeting, Proceedings of the Survey methods Section. Statistical Society of Canada, Ottawa.

[204] Trepetin S. (2008) *Privacy-Preserving String Comparisons in Record Linkage Systems: A Review*, Information Security Journal: A Global Perspective, 1939-3547, Volume 17(5):2008: 253–266. doi:10.1080/19393550802492503.

[205] Tromp M., Méray N., Ravelli A.C. J., Johannes B.R., Gouke J.B. (2008) *Ignoring dependency between Linking Variables an Its Impact on the Outcome of Probabilistic Record Linkage Studies*, Journal of the American Medical Informatics Association 2008 15: 654–660.

[206] Tromp M., Ravelli A.C.J., Méray N., Reitsma J.B., Bonsel G.J. (2008) *An Efficient Validation Method of Probabilistic Record Linkage Including Readmissions and Twins*, Methods of Information in Medicine, 47 (4): 356-363. doi: 10.3414/ME0489.

[207] Verykios V.S. (2000) *A Decision Model for Cost Optimal Record Matching*, National Institute of Statistical Sciences Affiliates Workshop on Data Quality, Morristown, New Jersey.

[208] Verykios V.S., Elfeky M.G., Elmagarmid A.K., Cochinwala M. and Dalal S. (2000) *On The Accuracy And Completeness Of The Record*

Matching Process, Sloan School of Management, editor, Proceedings of Information Quality Conference, MIT, Cambridge, MA.

[209] Verykios V.S., Karakasidis A., Mitrogiannis V.K. (2009) *Privacy preserving record linkage approaches*, International Journal of Data Mining, Modelling and Management, Volume 1, Number 2/2009: 206–221. doi: 10.1504/IJDMMM.2009.026076

[210] White, D. (1997) *A review of the statistics of record linkage for genealogical research*, Record Linkage Techniques – 1997. Proceedings of an International Workshop and Exposition (Alvey, W. and Jamerson, B. (eds)): 362–373.

[211] Winkler W.E. (1984) *Exact Matching Using Elementary Techniques*, Energy Information Administration Technical Report:237-241. U. S. Dept. of Energy, Washington DC.

[212] Winkler W.E. (1985a) *Exact matching lists of businesses: blocking, subfield identification and information theory*, Proceedings of the Section on Survey Research Methods, American Statistical Association: 438–443.

[213] Winkler W.E. (1985b) *Exact matching lists of businesses: blocking, subfield identification and information theory*, Record Linkage Techniques – 1985. Proceedings of the Workshop on Exact Matching Methodologies (Kills B. and Alvey W. (eds)): 227–241.

[214] Winkler W.E. (1985c) *Preprocessing of Lists and String Comparison*, In: Kills B. and Alvey W. (Comp. And Eds.) (1985) Record Linkage Techniques – 1985. Proceedings of the Workshop on Exact Matching Methodologies, Arlington, Virginia, 9-15 May 1985. Internal Revenue Service, Washington D.C.

[215] Winkler W.E. (1988) *Using the EM algorithm for weight computation in the Fellegi-Sunter Model of record linkage*, Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 667–671. [See Winkler W.E. (2000b)].

[216] Winkler W.E. (1989a) *Frequency-based matching in the Fellegi-Sunter model of record linkage*, Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 778–783. [See Winkler W.E.)2000c)].

[217] Winkler W.E. (1989b) *Near automatic weight computation in the Fellegi-Sunter model of record linkage*, Proceedings of the Annual Research Conference, Washington D.C., U.S. Bureau of the Census, pp. 145–155.

[218] Winkler W.E. (1989c) *Methods for adjusting for lack of independence in the Fellegi-Sunter model of record linkage*, Survey Methodology, Volume 15, pp. 101–117.

[219] Winkler W.E. (1990) *String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage*, Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 354–359.

[220] Winkler W.E. (1991) *Error model for analysis of computer linked files*, Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 472–477.

[221] Winkler W.E. (1992) *Comparative analysis of record linkage decision rules*, Proceedings of the Section on Survey Research Methods, American Statistical Association,pp. 829–834.

[222] Winkler W.E. (1993a) *Improved decision rules in the Fellegi-Sunter model of record linkage*, Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 274–279.

[223] Winkler W.E. (1993b) *Improved decision rules in the Fellegi-Sunter model of record linkage*, U.S. Bureau of the Census, Statistical Research Report Series, No. RR93/12. U.S. Bureau of the Census, Washington, D.C.

[224] Winkler W.E. (1994a) *Advanced methods for record linkage*, Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 467–472.

[225] Winkler W.E. (1994b) *Advanced methods for record linkage*, U.S. Bureau of the Census, Statistical Research Report Series, No. RR93/12. U.S. Bureau of the Census, Washington, D.C.

[226] Winkler W.E. (1995a) *Matching and Record Linkage*, Business Survey Methods (Cox B.G., Binder D.A., Chinnappa B.N., Christianson A., Colledge M., Kott P.S. (eds)), pp. 355–384. Wiley, New York.

[227] Winkler W.E. (1995b) *Matching and Record Linkage*, Record Linkage Techniques – 1997. Proceedings of an International Record Linkage Workshop and Exposition (Alvey, W. and Jamerson, B. (eds)): 3–12.

[228] Winkler W.E. (1997) *Producing public-use microdata that are analytically valid and confidential*, Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 41–50.

[229] Winkler W.E. (1998) *Re-identification methods for evaluating the confidentiality of analytically valid microdata*, Research in Official Statistics, Volume 2, pp. 87-104 - (2005). U.S. Bureau of the Census, Statistical Research Report Series, No. RR2005/09. U.S. Bureau of the Census, Washington, D.C.

[230] Winkler W.E. (1999a) *The state of record linkage and current research problems*, Proceedings of the Section on Survey Methods, SSC Annual Meeting.

[231] Winkler W.E. (1999b) *The state of record linkage and current research problems*, U.S. Bureau of the Census, Statistical Research Report Series, No. RR1999/04. U.S. Bureau of the Census, Washington, D.C.

[232] Winkler W.E. (2000a) *Machine learning, information retrieval and record linkage*, Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 20–29.

[233] Winkler W.E. (2000b) *Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage*, U.S. Bureau of the Census, Statistical Research Report Series, No. RR2000/05. U.S. Bureau of the Census, Washington, D.C.

[234] Winkler W.E. (2000c) *Frequency-Based Matching in Fellegi-Sunter Model of Record Linkage*, U.S. Bureau of the Census, Statistical Research Report Series, No. RR2000/06. U.S. Bureau of the Census, Washington, D.C.

[235] Winkler W.E. (2001) *Record Linkage Software and Methods for Merging Administrative Lists*, U.S. Bureau of the Census, Statistical Research Report Series, No. RR2001/03. U.S. Bureau of the Census, Washington, D.C.

[236] Winkler W.E. (2002) *Methods for Record Linkage and Bayesian Networks*, U.S. Bureau of the Census, Statistical Research Report Series, No. RRS2002/05, US Bureau of the Census.

[237] Winkler W.E. (2003a) *Record Linkage Software and Methods for Merging Administrative Lists*, Statistical Research Report Series No. RR2001/03. Statistical Research Division, U.S. Bureau of the Census, Washington D.C.

[238] Winkler W.E. (2003b) *Data Cleaning Methods*, Proceedings of the ACM Workshop on Data Cleaning, RecordLinkage and Object Identification, Washington, DC.

[239] Winkler W.E. (2004a) *Re-identification methods for masked microdata*, U.S. Bureau of the Census, Statistical Research Division Report Series, n.2004/3.

[240] Winkler W.E. (2004b) *Methods for Evaluating and Creating Data Quality*, ICDT Workshop on Data Quality in Cooperative Information Systems, Siena, Italy, January 2003.

[241] Winkler W.E. (2004c) *Methods for Evaluating and Creating Data Quality*, Information Systems (2004), 29 (7): 531–550. doi:10.1.1.84.6762.

[242] Winkler W.E. (2004d) *Masking and Re-Identification Methods for Public-Use Microdata: Overview and Research Problems*, U.S. Bureau of the Census, Statistical Research Division Report Series, n.2004/6.

[243] Winkler W.E. (2004e) *Approximate String Comparator Search Strategies for Very Large Administrative Lists*, Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 4595–4602.

[244] Winkler W.E. (2005a) *Approximate String Comparator Search Strategies for Very Large Administrative Lists*, U.S. Bureau of the Census, Statistical Research Division Report Series, n.2005/2.

[245] Winkler W.E. (2005b) *Re-Identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata*, U.S. Bureau of the Census, Statistical Research Division Report Series, n.2005/9.

[246] Winkler W.E. (2006a) *Overview of Record Linkage and Current Research Directions*, U.S. Bureau of the Census, Statistical Research Division Report Series, n.2006/2.

[247] Winkler W.E. (2006b) *Data Quality: Automated Edit-Imputation and Record Linkage*, U.S. Bureau of the Census, Statistical Research Division Report Series, n.2006/7.

[248] Winkler W.E. (2007) *Automatically Estimating Record Linkage False Match Rates*, U.S. Bureau of the Census, Statistical Research Division Report Series, n.2007/5.

[249] Winkler W.E. (2008a) *General Methods and Algorithms for Modeling and Imputing Discrete Data under a Variety of Constraints*, U.S. Bureau of the Census, Statistical Research Division Report Series, n.2008/8.

[250] Winkler W.E. (2008b) *General Discrete-data Modeling Methods for Producing Synthetic Data with Reduced Re-identification Risk that Preserve Analytic Properties*, IAB Workshop on Confidentiality and Disclosure, Nuremberg, Germany, November 20–21, 2008.

[251] Winkler W.E. (2010) *General Discrete-data Modeling Methods for Producing Synthetic Data with Reduced Re-identification Risk that Preserve Analytic Properties*, Statistics #2010-02 Research Report Series, U.S. Census Bureau, Washington, D.C. USA.

[252] Winkler W.E. (2010) *General Discrete-data Modeling Methods for Producing Synthetic Data with Reduced Re-identification Risk that Preserve Analytic Properties*, U.S. Bureau of the Census, Statistical Research Division Report Series, n.2010/2.

[253] Winkler W.E. and Scheuren F. (1991) *How Matching Error Effects Regression Analysis: Exploratory and Confirmatory Results*, Unpublished report, Washington DC: Statistical Research Division Technical Report, U.S.

[254] Winkler W.E. and Scheuren F. (1996) *Recursive analysis of linked data files*, Proceedings of the 1996 Census Bureau Annual Research Conference, pp. 920–935.

[255] Winkler W.E. and Thibaudeau Y. (1991) *An application of the Fellegi-Sunter model of record linkage to the 1990 U.S. decennial Census*, Bureau of the Census, Statistical Research Division,Statistical Research Report Series, n. RR91/09. [Link #2]

[256] Winkler W.E., Yancey W.E. and Porter E.H. (2010) *Fast Record Linkage of Very Large Files in Support of Decennial and Administrative Records Projects*, Proceedings of the Section on Survey Research Methods, American Statistical Association.

[257] Yakout M., Atallah M.J., Elmagarmid A. (2009) *Efficient Private Record Linkage*, Proceedings of the 2009 IEEE International Conference on Data Engineering: 1283–1286. IEEE Computer Society. ISSN: 978-0-7695-3545-6. doi:10.1109/ICDE.2009.221.

[258] Yan S., Lee D., Kany M.-Y., Giles C.L. (2007) *Adaptive Sorted Neighborhood Methods for Efficient Record Linkage*, Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries: 185–194. Vancouver, Canada. ISBN: 978-1-59593-644-8.

[259] Yancey W.E. (2000) *Frequency-dependent probability measures for record linkage*, Statistical Research Report Series, n. RR2000/07. U.S. Bureau of the Census, Washington, DC.

[260] Yancey W.E. (2002) *BigMatch: A Program for Extracting Probable Matches from a Large File for Record Linkage*, Technical Report RRC 2002-01. Statistical Research Division, U.S. Bureau of the Census, Washington, DC.

[261] Yancey W.E. (2004a) *Improving EM algorithm estimates for record linkage parameters*, Statistical Research Division Report Series, n. 2004/01. U.S. Bureau of the Census, Washington, DC.

[262] Yancey W.E. (2004b) *An Adaptive String Comparator for Record Linkage*, Statistical Research Division Report Series, n. 2004/02. U.S. Bureau of the Census, Washington, DC.

[263] Yancey W.E. (2004c) *BigMatch: A Program for Large-Scale Record Linkage*, Proceedings of the Section on Survey Research Methods, American Statistical Association.

[264] Yancey W.E. (2005) *Evaluating String Comparator Performance for record linkage*, Statistical Research Division Report Series, n. 2005/05. U.S. Bureau of the Census, Washington, DC.

[265] Yancey W.E. (2007) *BigMatch: A Program for Extracting Probable Matches from a Large File for Record Linkage*, Technical Report RR2007/01. U.S. Bureau of the Census, Washington, DC.

[266] Yancey W.E., Winkler W.E., and Creecy R.H. (2002) *Disclosure Risk Assessment in Perturbative Microdata Protection*, In: Domingo-Ferrer, J. (Ed.) (2002) Inference Control in Statistical Databases: 49–60. Springer Berlin / Heidelberg. ISSN: 1611-3349. doi:10.1007/3-540-47804-3_11.

[267] Yin X., Han J. and Yu P.S. (2006) *LinkClus: Efficient Clustering via Heterogeneous Semantic Links*, VLDB ‘06, September, 2006, Seoul, Korea ACM.

[268] Zardetto D., Scannapieco M., Catarci T. (2010) *Effective Automated Object Matching*, 26th IEEE International Conference on Data Engineering, Long Beach, USA.

[269] Zhu V.J., Overhage M.J., Egg J., Downs S.M., Grannis S.J. (2009) *An Empiric Modification to the Probabilistic Record Linkage Algorithm Using Frequency-Based Weight Scaling*, Journal of the American Medical Informatics Association, 2009, 16: 738–745. doi:10.1197/jamia.M3186.

Bibliography on Statistical Matching

[1] Abello, R. and Phillips, B. (2004) *Statistical matching of the HES and NHS: an exploration of issues in the use of unconstrained and constrained approaches in creating a basefile for a microsimulation model of the pharmaceutical benefits scheme*, Technical report, Australian Bureau of Statistics. Methodology Advisory Committee Paper, June.

[2] Adamek, J.C. (1994) *Fusion: Combining data from separate sources*, Marketing Research: A Magazine of Management and Applications 6, 48–50.

[3] Anagnoson, J.T. (2000) *Microsimulation of public policy*, In: G.D. Garson (ed.), The Handbook of Public Information Systems. New York: Marcel Dekker.

[4] Anderson, T.W. (1957) *Maximum likelihood estimates for a multivariate normal distribution when some observations are missing*, Journal of the American Statistical Association 52, 200–203.

[5] Antoine, J. (1987) *A case study illustrating the objectives and perspectives of fusion techniques*, In: H. Henry (ed.), Readership Research: Theory and Practice, pp. 336–351. Amsterdam: Elsevier Science.

[6] Antoine, J. and Santini, G. (1987) *Fusion techniques: alternative to single source methods*, European Research 15, 178–187.

[7] Baker, K. (1990) *The BARB/TGI fusion*, Technical report, Ken Baker Associates, Ickenham, UK.

Papers can be accessed via the links in the electronic version of the document. Additional references, comments, corrections or requests can be sent to essnet.di@istat.it, scanu@istat.it or mguigo@ine.es *(last updated: 30-11-2011)*

[8] Baker, K., Harris, P. and O'Brien, J. (1989) *Data fusion: an appraisal and experimental evaluation*, Journal of the Market Research Society 31, 152–212.

[9] Ballin M., Di Zio, M, D'Orazio, M, Scanu, M, Torelli, N. (2008) *File Concatenation of Survey Data: a Computer Intensive Approach to Sampling Weights Estimation*, Rivista di Statistica Ufficiale.

[10] Barr, R.S. and Turner, J.S. (1981) *Microdata file merging through large-scale network technology*, Mathematical Programming Study, Volume 15, pp. 1–22.

[11] Barr, R.S. and Turner, J.T. (1990) *Quality issues and evidence in statistical file merging*, In: G.E. Liepins and V.R.R. Uppuluri (eds), Data Quality Control: Theory and Pragmatics, pp. 245–313. New York: Marcel Dekker. The technical report is available here http://faculty.smu.edu/barr/pubs/90liepBarrTurn-QualityIssues.pdf.

[12] Barry, J.T. (1988) *An investigation of statistical matching*, Journal of Applied Statistics 15, 275–283.

[13] Bordt, M., Cameron, G.J., Gibble, S.F., Murphy, B.B., Rowe, G.T. And Wolfson, M.C. (1990) *The social policy simulation database and model: an integrated tool for tax/transfer policy analysis*, Canadian Tax Journal 38, 48–65.

[14] Cassel, C.M. (1983) *Statistical matching-statistical prediction. What is the difference? An evaluation of statistical matching and a special type of prediction using data from a survey on living conditions*, Statistisk Tidskrift 5, 55–63.

[15] Citoni, G., Di Nicola, F., Lugaresi, S. and Proto, G. (1991) *Statistical matching for tax benefit microsimulation modelling: a project for Italy*, Technical Report 'Gruppo di Lavoro sulle Analisi delle Politiche Redistributive', Istituto di Studi per la Programmazione Economica, November.

[16] Cohen, M.L. (1991) Statistical matching and microsimulation models. In: C.F. Citro and E.A. Hanushek (eds), *Improving Information for Social Policy Decisions: The Uses of Microsimulation Modeling*, Vol. II: Technical Papers.Washington, DC: National Academy.

[17] Coli, A., Tartamella, F., Sacco, G., Faiella, I., Scanu, M., D'Orazio, M., Di Zio, M., Siciliani, I., Colombini, S. and Masi, A. (2005) *La*

costruzione di un archivio di microdati sulle famiglie italiane ottenuto integrando l'indagine ISTAT sui consumi del.

[18] Conti P. L., Di Zio M., Marella D., Scanu M. (2009) *Uncertainty analysis in statistical matching*, First Italian Conference on Survey Methodology (ITACOSM09), Siena 10-12 June 2009.

[19] Conti P.L., Marella D., Scanu M (2006) *Nonparametric evaluation of matching noise*, Proceedings of the IASC conference "Compstat 2006", Roma, 28 August - 1 September 2006, Physica-Verlag/Springer, pp. 453–460.

[20] Conti P.L., Marella D., Scanu M. (2008) *Evaluation of matching noise for imputation techniques based on the local linear regression estimator*, Computational Statistics and Data Analysis, 53, 354–365.

[21] Conti, P.L. and Scanu, M. (2005) *On the evaluation of matching noise produced by nonparametric imputation techniques*, Rivista di Statistica Ufficiale, 1/2006, 43–56.

[22] D'Orazio M., Di Zio M., Scanu M. (2006) *Statistical Matching, Theory and Practice*, Wiley, Chichester.

[23] D'Orazio, M, Di Zio, M., Scanu, M. (2009) *Uncertainty intervals for nonidentifiable parameters in statistical matching*, 57th Session of the International Statistical Institute, Durban (South Africa), 16-22 August 2009.

[24] D'Orazio, M., Di Zio, M. and Scanu, M. (2002) *Statistical matching and official statistics*, Rivista di Statistica Ufficiale 2002/1, 5–24.

[25] D'Orazio, M., Di Zio, M. and Scanu, M. (2005) *A comparison among different estimators of regression parameters on statistically matched files through an extensive simulation study*, Technical Report, Contributi 2005/10, Istituto Nazionale di Statistica, Rome.

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

[27] DeGroot, M.H. (1987) *Record linkage and matching systems*, In: S. Kotz and N.L. Johnson (eds) Encyclopedia of Statistical Sciences, Vol. 7, pp. 649–654. Wiley, New York.

[28] DeGroot, M.H. and Goel, P.K. (1976) *The matching problem for multivariate normal data*, Sankhya, B 38, 14–29.

[29] DeGroot, M.H., Feder, P.I. and Goel, P.K. (1971) *Matchmaking*, Annals of Mathematical Statistics 42, 578–593.

[30] Denk, M. and Hackl, P. (2003) *Data integration and record matching: an Austrian contribution to research in official statistics*, Austrian Journal of Statistics 32, 305–321.

[31] Dobra, A. and Fienberg, S. E. (2000) *Bounds for cell entries in contingency tables given marginal totals and decomposable graphs*, Proceedings of the National Academy of Sciences, 97, No. 22, 11885–11892.

[32] Dobra, A. and Fienberg, S. E. (2001) *Bounds for cell entries in contingency tables induced by fixed marginal totals*, UNECE Statistical Journal, 18, 363–371.

[33] Dobra, A. and Fienberg, S. E. (2003) *Bounding entries in multi-way contingency tables given a set of marginal totals*, In: Y. Haitovsky, H. R. Lerche and Y. Ritov (eds.) Foundations of Statistical Inference, Proceedings of the Shoresh Conference 2000, 3–16.

[34] Dobra, A. and Fienberg, S.E. (2010) *The generalized shuttle algorithm*, In: P. Gibilisco, E. Riccomagno, M.P. Rogantin and H.P. Wynn (eds.) Algebraic and geometric methods in statistics, Volume dedicated to Professor Giovanni Pistone, Cambridge University.

[35] Dobra, A., Erosheva, E. A. and Fienberg, S. E. (2002) *Disclosure limitation methods based on bounds for large contingency tables with application to disability data*, In: H. Bozdogan (ed.) Statistical Data Mining and Knowledge Discovery, CRC Press.

[36] Dobra, A., Fienberg, S. E. and Trottini, M. (2003) *Assessing the risk of disclosure of confidential categorical data*, In: J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West (eds.) Bayesian Statistics 7, Oxfo.

[37] Dobra, A., Fienberg, S.E., Rinaldo, A., Slavkovic, A. and Zhou, Y. (2009) *Algebraic statistics and contingency table problems: estimation and disclosure limitation*, In: S. Sullivant and M. Putinar (eds.) IMA Volume 149 on Emerging applications of algebrai.

[38] Dobra, A., Karr, A. and Sanil, A. (2003) *Preserving confidentiality of high-dimensional tabulated data: statistical and computational issues*, Statistics and Computing, 13, 363–370.

[39] Dobra, A., Tebaldi, C. and West, M. (2006) *Data augmentation in multi-way contingency tables with fixed marginal totals*, Journal of Statistical Planning and Inference, 136, 355–372.

[40] D'Orazio M., Di Zio M., Scanu M., Torelli N., Ballin M. (2009) *Statistical matching of two surveys with a common subset*, First Italian Conference on Survey Methodology (ITACOSM09), Siena 10-12 June 2009.

[41] Erosheva, E. A., Fienberg, S. E., and Junker, B. W. (2002) *Alternative statistical models and representations for large sparse multi-dimensional contingency tables*, Ann. Fac. Sci. Toulouse Math. (6) 11, no. 4, 485–505.

[42] Fienberg, S. E. and Slavkovic, A. B. (2005) *Preserving the Confidentiality of Categorical Data Bases When Releasing Information for Association Rules*, Data Mining and Knowledge Discovery, 11, 155–180.

[43] Filippello, R., Guarnera, U. and Jona Lasinio, G. (2004) *Use of auxiliary information in statistical matching*, Proceedings of the XLII Conference of the Italian Statistical Society, pp. 37–40. Bari (Italy), 9-11 June 2004. Padua: CLEUP.

[44] Fisseler, J., Fehér I (2006) *A probabilistic approach to data fusion*, Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, Seattle, Washington, USA, 139–146.

[45] Flores G.A., Albacea E.A. (2007) *A Genetic Algorithm for Constrained Statistical Matching*, 10th National Convention on Statistics (NCS).

[46] Gavin N.I. (1985) *An application of statistical matching with the survey of income and education and the 1976 health interview survey*, Health Survey Research, 20, 183–198.

[47] Gilula, Z, McCulloch, R.E., Rossi, P.E. (2006) *A direct approach to data fusion*, Journal of Marketing Research, 43, 73–83.

[48] Goel, P.K. and Ramalingam, T. (1989) *The Matching Methodology: Some Statistical Properties*, New York: Springer-Verlag.

[49] Jephcott, J. and Bock, T. (1991) *The application and validation of data fusion*, Journal of the Market Research Society 40, 185–205.

[50] Kadane, J.B. (1978) *Some statistical problems in merging data files*, 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.

[51] Kamakura, W.A. and Wedel, M. (1997) *Statistical data fusion*, Journal of Marketing Research 34, 485–498.

[52] Klevmarken, N.A. (1986) *Comment on Paass* (1986) In: G.H. Orcutt, J. Merz and H. Quinke (eds), Microanalytic SimulationModels to Support Social and Financial Policy, pp. 421–422. Amsterdam: Elsevier Science.

[53] Kum H., Masterson T. (2008) *Statistical Matching Using Propensity Scores: Theory and Application to the Levy Institute Measure of Economic Well-Being*, Working Paper No. 535, The Levy Economics Institute of Bard College.

[54] Liu, T.P. and Kovacevic, M.S. (1994) *Statistical matching of survey datafiles: a simulation study*, Proceedings of the Section on Survey Research Methods of the American Statistical Association, pp. 479–484.

[55] Marella D., Scanu M., Conti P.L. (2008) *On the matching noise of some nonparametric imputation procedures*, Statistics and Probability Letters, 78, 1593–1600.

[56] Moriarity C. (2009) *Regression-Based Statistical Matching: Past, Present, and Future*, The 57th session of the International Statistical Institute, Durban (South Africa) 16-22 August 2009.

[57] Moriarity C. (2009) *Statistical Properties of Statistical Matching*, VDM Verlag.

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

[59] Moriarity, C. and 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.

[60] Moriarity, C. and Scheuren, F. (2004) *Regression-based statistical matching: recent developments*, Proceedings of the Section on Survey Research Methods, American Statistical Association.

[61] Noll, P. Alpar, P. (2007) *A Methodology for Statistical Matching with Fuzzy Logic*, Fuzzy Information Processing Society, 2007. NAFIPS '07. Annual Meeting of the North American, pp. 73–78.

[62] O'Brien, S. (1991) *The role of data fusion in actionable media targeting in the 1990's*, Marketing and Research Today 19, 15–22.

[63] Okner, B.A. (1972) *Constructing a new data base from existing microdata sets: the 1966 merge file*, Annals of Economic and Social Measurement 1(3), 325–342.

[64] Okner, B.A. (1974) *Data matching and merging: an overview*, Annals of Economic and Social Measurement 3(2), 347–352.

[65] Paass, G. (1985) *Statistical record linkage methodology: state of the art and future prospects*, Bulletin of the International Statistical Institute, Proceedings of the 45th Session, Vol. LI, Book 2. Voorburg, Netherlands: ISI.

[66] Paass, G. (1986) *Statistical match: evaluation of existing procedures and improvements by using additional information*, In: G.H. Orcutt, J. Merz and H. Quinke (eds) Microanalytic Simulation Models to Support Social and Financial Policy, pp. 401–422. Amsterdam: Elsevier Science.

[67] Pesti CS., Kaposzta J. (2008) *Adaptation of Statistical Matching in Micro-Regional Analysis of Agricultural Production*, Bull. of the Szent István Univ., Gödöllő, pp. 277–284.

[68] *Federal Committee on Statistical Methodology*.

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

[70] Raessler, S. (2003) *A non-iterative Bayesian approach to statistical matching*, Statistica Neerlandica 57(1), 58–74.

[71] Raessler, S. (2004) *Data fusion: identification problems, validity, and multiple imputation*, Austrian Journal of Statistics 33(1-2), 153–171.

[72] Raessler, S. and Fleischer, K. (1999) *An evaluation of data fusion techniques*, Proceedings of the XVI International Methodology Symposium-Statistics Canada, 2-5 May 1999, pp. 129–136. Ottawa: Statistics Canada.

[73] Raessler, S., Kiesl, H. (2009) *How useful are uncertainty bounds? Some recent theory with an application to Rubin's causal model*, 57th Session of the International Statistical Institute, Durban (South Africa), 16-22 August 2009.

[74] Reiter J. (2009) *Bayesian finite population imputation for data fusion*, Duke Population Research Institute, On-line Working Paper Series, PWP-DUKE-2009-008.

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

[76] Roberts, A. (1994) *Media exposure and consumer purchasing: an improved data fusion technique*, Marketing and Research Today 22, 159–172.

[77] Rodgers, W.L. (1984) *An evaluation of statistical matching*, Journal of Business and Economic Statistics 2, 91–102.

[78] Rubin, D.B. (1974) *Characterizing the estimation of parameters in incomplete-data problems*, Journal of the American Statistical Association 69, 467–474.

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

[80] Ruggles, N. (1999) *The development of integrated data bases for social, economic and demographic statistics*, In: N. Ruggles and R. Ruggles (eds) Macro- and Microdata Analyses and Their Integration, pp. 410–478. Cheltenham: Edward Elgar.

[81] Ruggles, N. and Ruggles, R. (1974) *A strategy for merging and matching microdata sets*, Annals of Economic and Social Measurement 1(3), 353–371.

[82] Saporta, G. (2002) *Data fusion and data grafting*, Computational Statistics and Data Analysis, 38, Issue 4, 465–473.

[83] Sims, C.A. (1972) *Comments on Okner (1972)*, Annals of Economic and Social Measurement 1(3), 343–345.

[84] Singh, A.C., Armstrong, J.B. and Lemaitre, G.E. (1988) *Statistical matching using log linear imputation*, Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 672–677.

[85] Singh, A.C., Mantel, H., Kinack, M. and Rowe, G. (1990) *On methods of statistical matching with and without auxiliary information*, Technical Report SSMD-90-016E, Methodology Branch, Statistics Canada.

[86] Singh, A.C., Mantel, H., Kinack, M. and Rowe, G. (1993) *Statistical matching: Use of auxiliary information as an alternative to the conditional independence assumption*, Survey Methodology 19, 59–79.

[87] Slavkovic, A. B. (2009) *Partial Information Releases for Confidential Contingency Table Entries: Present and Future Research Efforts*, Journal of Privacy and Confidentiality, 1(2), Article 9.

[88] Slavkovic, A. B., and Fienberg, S.E. (2004) *Bounds for Cell Entries in Two-Way Tables Given Conditional Frequencies*, Proceedings of Privacy in Statistical Databases 2004. Lecture Notes in Computer Science No.3050, 30–43.

[89] Slavkovic, A. B., and Fienberg, S.E. (2004) *Making the Release of Confidential Data from Multi-Way Tables Count*, Chance, 17, 3, (2004), 5–10.

[90] Slavkovic, A. B., and Fienberg, S.E. (2008) *A Survey of Statistical Approaches to Preserving Confidentiality of Contingency Table Entries*, Privacy-Preserving Data Mining: Models and Algorithms, Vol. 34, 291–312.

[91] Slavkovic, A. B., Smucker, B. (2008) *Cell Bounds in Two-Way Contingency Tables Based on Conditional Frequencies*, Privacy in Statistical Databases, Lecture Notes in Computer Science No.5262, 64–77.

[92] Sutherland, H., Taylor, R. and Gomulka, J. (2001) *Combining household income and expenditure data in policy simulations*, Technical Report MU0101, Department of Applied Economics, University of Cambridge.

[93] Tereshchenko G. (2008) *Improving of Household Sample Surveys Data Quality on Base of Statistical Matching Approaches*, Institute for Demography and Social Studies of the National Academy of Sciences of Ukraine.

[94] Van der Putten, P. W. H. (2010) *On Data Mining in Context: Cases, Fusion and Evaluation*, PhD Theses, University of Leiden.

[95] Vantaggi B. (2008) *Statistical matching of multiple sources: A look through coherence*, International Journal of Approximate Reasoning, 49, Issue 3, pp. 701–711.

[96] Vantaggi, B. (2005) *The role of coherence for the integration of different sources*, In: F.G. Cozman, R. Nau and T. Seidenfeld (eds), Proceedings: 4th International Symposium on Imprecise Probabilities and Their Applications, pp. 269–378. Pittsburgh: Bright.

[97] Wiegand, J. (1986) *Combining different media surveys: The German partnership model and fusion experiments*, Journal of the Market Research Society 28, 189–208.

[98] Wilks, S.S. (1932) *Moments and distributions of estimates of population parameters from fragmentary samples*, Annals of Mathematical Statistics 3, 163–194.

[99] Wolfson, M., Gribble, S., Bordt, M., Murphy, B. and Rowe, G. (1987) *The social policy simulation database: an example of survey and administrative data integration*, In: J.W. Coombs and M.P. Singh (eds), Proceedings: Symposium on Statistical Uses of Administrative Data, pp. 201–229. Ottawa: Statistics Canada.

[100] Wolfson, M., Gribble, S., Bordt, M., Murphy, B.B. Rowe, G.T., Scheuren, F. (1989) *The social policy simulation database and model: an example of survey and administrative data integration*, Survey of Current Business 69, 36–41.

[101] Wu, C. (2004) *Combining information from multiple surveys through the empirical likelihood method*, The Canadian Journal of Statistics, 32, 15–26.

[102] Yoshizoe, Y. and Araki, M. (1999) *Statistical matching of household survey files*, Technical Report 10, ITME (Information Technology and the Market Economy) Project of the Japan Society for the Promotion of Science.