1. Purpose of the method

A data set (recipient) is imputed of values from a second data set (donor) on a variable observed in the donor but not in the recipient, by means of variables commonly observed in both the donor and recipient data sets.

2. The related scenarios

2.1. The method refers to a statistical matching problem of two data sets that refer to the same population, whose observed sets of units is not overlapping and where the common variables are a strict subset of the available variables in the two data sets. It mainly refers to data configuration 3 (deliverable 1).

2.2. Statistical usages are mainly Substitution and Supplementation, with the purpose of Direct tabulation for micro simulation models: The resulting data set can be analysed together with the concept of uncertainty for any other statistical analyses.

2.3. Statistical tasks: Creation of joint statistical microdata.

3. Description of the method

The method consists in the application of imputation techniques for imputing a new variable on a recipient data set from a donor data set where that variable is available. A key role is played by the common variables among the ones observed in the recipient and donor files.

The most common approaches are nonparametric and belong to the hot deck imputation methods. For each record in the recipient, a donor is taken from the other file according to these features:

- Random hot deck: the record is taken at random for the ones observed in the donor data set;
- Rank hot deck: the records in the recipient and donor files are ordered according to one common variable and the corresponding cumulative distributions functions are computed; the donor is chosen among those whose value of the cumulative distribution function is the nearest to the corresponding value in the recipient;
- Distance hot-deck: taken a distance functions of the common variables observed in the two files, each record in the recipient is imputed with that record in the donor whose observed common variables are the nearest to the ones observed in the recipient record.

Each of the previous methods can be modified (constrained) in order to:

- Restrict donor records to subsets of those available in the donor file (for instance considering only records whose values of some common variables are fixed);
- Use each donor only once, at maximum.

Among these methods, attention to the case the data sets to fuse are obtained by complex survey designs and each record is complemented with survey weights is considered. The weight-split algorithm (Kovacevic and Liu, 1994) is one of these methods: the objective is to use a hot-deck method in such a way that the imputed variable in the recipient file computes the same distribution as observed in the donor file. Assume one variable X and order the observations in the two files according to X. Rescale the survey weights in the two files so that their sum is equal to 1 in both the recipient and donor files. Compute the cumulative distribution functions of these weights in the two files. If a record in the recipient has a value of the

cumulative distribution function in between the corresponding values of two ordered consecutive observed values in the donor, with the possibility of equality with the higher donor (a tie between recipient and donor *X*), then impute the recipient with the record corresponding to the higher donor. For those records in the donor that do not have a tie, use each one of them as the donor record to that recipient with the lowest cumulative distribution function among those with cumulative distribution function larger than the one observed in the donor record. Consequently the whole number of imputed records are the sum of the donor and recipient file sizes less the number of ties. Assign to each of these records the minimum between the cumulative distribution function values of the donor and recipient, and compute the difference between these values for the consecutive records. These differences are the new survey weights to attach to the records of this new data set.

Parametric methods have also been proposed (e.g. through imputations by means of regression functions, with the possibility to add noise around the function itself), anyway their dependency on the normality assumption is very strong. For this reason imputations by parametric models are also mixed with non-parametric methods, leading to the so called statistical matching mixed methods.

Among the tools used for data fusion at the micro level, it is worthwhile to mention those based on multiple imputation. These methods are important in order to include the assessment of uncertainty on parameters that cannot be estimated on the available samples. A thorough discussion on these methods in Raessler (2002) and enhancements are in Reiter (2009).

4. Examples

Statistics Canada imputes the Survey of Labour and Income Dynamics (recipient) with other three data sets: a sample of anonymized Personal income tax return data, a sample of the Employment Insurance claim histories, and the Survey of household spending. They essentially apply donor based techniques using constrained distance hot deck methods by means of the weight-split algorithm (all the data sets have their own survey weights). Donors are created by constraining the data sets to values of common observed values (e.g. gender, age class and region of residence). The weight split algorithm has been applied to the variable income, so that this variable is taken into account in the imputation process (attenuating the Conditional Independence Assumption between income and the imputed variables in the resulting file). Note that the weight-split algorithm works on the cumulative distribution functions, hence the actual values of the income as observed in the files to match is not considered in the matching process, but only its order with respect to the other values (in other words, poor people are imputed with other poor people, without taking into account the actual income as declared in the two files, but only their relative position with respect to the other observed values).

5. Input data (characteristics, requirements for applicability)

Two (or more) data sets containing microdata, whose observed sets of units are disjoint, and the observed variables admit a strict subset of common variables.

6. Output data (characteristics, requirements)

A recipient data set with additional variables as observed in the donor file.

7. Tools that implement the method

R package *StatMatch*.

8. Appraisal

Accuracy of results should be assessed by means of the so called uncertainty analysis in statistical matching.

Usually, the conditional independence assumption between the specific variables in the two files given the common variables affects the resulting data set. If this assumption can be considered as an approximation of the actual but unknown relationship model between the specific variables in the two files given the common variables, than the resulting data set can be sued for statistical analysis. This assumption has already been considered in different contexts, e.g. for the statistically matching household income and expenditures, by using among the matching variables the ordered households according to their income, form the poorest to the richest (irrespective of the values declared in the two files).

9. References

Donatiello D., D'Orazio M., Frattarola D., Rizzi A., Scanu M., Spaziani M. (2016) The role of the conditional independence assumption in statistically matching income and consumption. Statistical Journal of the IAOS, 32, 667-675, DOI 10.3233/SJI-161000

Hennessy, Sanmartin, Eftekhary, Plager, Jones, Onate, Mc Evoy, Hicks, Deber (2015) Creating a synthetic database for use in microsimulation models to investigate alternative health care financing strategies in Canada. International Journal Of Microsimulation, 8(3) 41-74

Kovacevic M and Liu T (1994) 'Statistical Matching of Survey Datafiles: A Simulation Study', Proceedings of the Section on Survey Research Methods, American Statistical Association, 479-484

Statistics Canada (2014) SPSD/M database creation guide', Social Analysis and Modelling Division, Statistics Canada, Ottawa, Ontario. (http://www.statcan.gc.ca/eng/microsimulation/spsdm/spsdm)

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

Reiter, J. (2009), Using Multiple Imputation to Integrate and Disseminate Confidential Microdata," International Statistical Review, 77, 179–195.

1. Purpose of the method

The purpose of object matching (known commonly as record linkage) is to identify the same real world entity, at micro level, that can be differently represented in data sources, even if unique identifiers are not available or are affected by errors. In statistics, record linkage is needed for several applications, including: enriching the information stored in different data-sets; de-duplicating data-sets; improving the data quality of a source; measuring a population amount by capture-recapture method; checking the confidentiality of public-use micro data. Starting from the earliest contributions, dated back to 1959, there has been a proliferation of different approaches based on statistics, databases, machine learning, knowledge representation.

It's possible to distinguish between weighted matching of object characteristics and unweighted matching of object characteristics (see Memobust 2014): the first, can be formulated as an optimization problem and it is applied to match two data sets with some relevant overlapping units, on common object characteristics. The method (Willenborg and Heerschap, 2012) is able to value the strength of possible (candidate) matches by using matching weights; the latter can be seen as a special case of the first but it is applied in case both dataset don't have good quality object identifiers.

2. The related scenarios

2.1. The matching of object characteristics can be applied to all the usages (both direct and indirect) delineated in Deliverable 1, that involves more than one source and requires the input and output data at micro level when error-free unique unit identifiers are lacking in the sources.

2.2. Creation of joint statistical data, combining data at micro level belonging to the same unit.

2.3. Probabilistic record linkage (Object matching): It is important to distinguish between the statistical matching that involves the integration of different units, e.g. derived from different sample surveys and the record linkage that concerns the integration of sources composed mainly of the same units, partially or completely overlapping.

3. Description of the method

The weighted matching method uses weights to match records on the same object from different data sets. The module draws heavily on Willenborg and Heerschap (2012) to which the interested reader is referred for additional information. The weights are used because not all of the variables are equally reliable, that is, that they do not have reliable scores or the weights can express the degree of similarity or dissimilarity in the different objects corresponding with records that are matching candidates. Or you want to use a probability to show that two objects are probably the same. Then a probability model is needed to quantify differences in scores on the matching key, and the resulting probabilities can be used as matching weights. Weighted matching can be formulated as an optimisation problem, in which the optimal (weighted) sum of matches is calculated, under certain constraints, such as that each record can appear in at most one match. The goal of the method is to find solutions to such problems, exact ones or good approximations.

The unweighted matching can be seen as a particular case of the previous one and it is consist of several steps and applied when the object identifiers are of poor quality) from both datasets. First the potentially matching records in the two data sets are identified. This requires a suitable metric and a cut-off value so that records that are too different are not considered as candidate matches. In the next step from these potential matches, a subset is computed that maximises the number of matches, under suitable constraints.

4. Examples

Some examples of the methods can be derived from Memobust (2014).

5. Input data (characteristics, requirements for applicability)

At least two different data sources at micro level. The unique identifiers could be not available or affected by errors.

6. Output data (characteristics, requirements)

At the end, the matching of the object characteristics creates an integrated data set still composed by unit at micro level overlapping in the two data sources with the joint variables added.

7. Tools that implement the method

No tool specific.

8. Appraisal

The strengths of these methods depends on the quality of the common object that are in the sources; in case of poor quality the unweighted matching may be considered as an option.

In case of different reference period for the sources, heterogeneity in the recorded variables the quality of the matching procedure could be poor.

As stated for the probabilistic record linkage method the results can be evaluated in terms of the false matches and false unmatched rates, performed by clerical reviews or can be assessed based on the inspection of matches of test files. It is a labour intensive job to carry out. The quality indicators are influenced by the way that the weights are calculated.

9. References

Fellegi, I. P. and Sunter, A. B. (1969), A theory for record linkage. Journal of the American Statistical Association 64, 1183–1200.

Lawler, E. L. (1976), Combinatorial Optimization: Networks and Matroids. Holt, Rinehart, Winston.
Nemhauser, G. L. and Wolsey, L. A. (1988), Integer and Combinatorial Optimization. Wiley.

Memobust (2014), Method: Unweighted Matching of Object Characteristics In: *Memobust Handbook on Methodology of Modern Business Statistics,* https://ec.europa.eu/eurostat/cros/content/memobust_en.

Memobust (2014), Method: Weighted Matching of Object Characteristics In: *Memobust Handbook on Methodology of Modern Business Statistics,* https://ec.europa.eu/eurostat/cros/content/memobust_en.

Papadimitriou, C. H. and Steiglitz, K. (1998), Combinatorial Optimization: Algorithms and Complexity. Dover, Mineola (NY).

Willenborg, L. and Heerschap, N. (2012), Matching. Contribution to the Methods Series, Statistics Netherlands, The Hague.

1. Purpose of the method

The record linkage (also referred to as object identification, record matching, entity matching, entity resolution, reference reconciliation) is the set of methods aiming at accurately and quickly identify, at micro level, the same real world entity that can be differently represented in various data sources, even if unique identifiers are not available or are affected by errors. The record linkage can be seen as a complex process involving different knowledge areas and, in case of the National Statistical Institutes (NSIs), the integrated use of statistical and administrative sources is a product of a rationalization of all the available sources needed for several applications: enriching the information stored in different data-sets; de-duplicating data-sets; improving the data quality of a source; measuring a population amount by capture-recapture method; checking the confidentiality of public-use micro data. In research literature, the deterministic linkage is associated with the use of a specific decision rules in order to individuates links, that is a couple is a match if and only if there is a full agreement of unique or common identifiers - the matching variables- or it satisfies some a priori defined rules; the probabilistic approaches, in the other hand, makes an explicit use of probabilities for deciding when a given pair of records is actually a match; compared with the deterministic approach, the probabilistic one can solve problems caused by bad quality data and can be helpful when differently spelled, swapped variables are stored in the two data files; the attention here is only devoted to the probabilistic record linkage approach which allows also to evaluate the linkage errors, calculating the likelihood of the correct match.

It is important to distinguish between the statistical matching that involves the integration of different units, e.g. derived from different sample surveys and the record linkage that concerns the integration of sources composed mainly of the same units, partially or completely overlapping, e.g., in the case of integration of administrative registers and sample surveys (ESSnet ISAD, 2011).

2. The related scenarios

2.1. Usages and Komuso Data Configurations (deliverable 1): Probabilistic record linkage procedures may be applied to all the usages (both direct and indirect) delineated in Deliverable 1, that involves more than one source and requires the input and output data at micro level when error-free unique unit identifiers are lacking in the sources.

Linkage estimation procedures may be required in the first and most basic data configuration 1 (see deliverable 1), called the "split-variable" case, concerning multiple cross-sectional data sources covering the target population where the different data sets contain different target variables and common variables able jointly to identify the units. This linkage step is also needed in the basic data configurations 4 and 5; the former is characterised by multiple cross-sectional data sources covering the target population with overlapping units between the different data sources, the latter is characterised by under coverage of the target population in the different integrated sources.

2.2. Statistical tasks, (deliverable 2): The Probabilistic record linkage can be adopted for the *Creation of joint statistical data, combining data at micro level belonging to the same unit*.

2.3. Possible competing methods or related methods could be quoted in this part: It is important to distinguish between the statistical matching that involves the integration of different units, e.g. derived from different sample surveys and the record linkage that concerns the integration of sources composed mainly of the same units, partially or completely overlapping, e.g., in the case of integration of administrative registers and sample surveys (ESSnet ISAD, 2011).

3. Description of the method

Due to the characteristic of the data, that is missing and uncorrected values, the data integration is realised by the probabilistic record linkage, according to the classical theory of Fellegi and Sunter (1969).

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

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

The probability models for linkage assume that the probability distribution of the comparison pattern comes from a mixture of two probability distributions: the first one comes from the pairs (a,b) that actually are the same unit, called distribution m; the other one comes from the pairs (a,b) that actually represent different units, called distribution u. Starting from the estimations of the two distribution m(γ) and u(γ), it is possible to define the composite matching weight, given by the likelihood ratio:

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

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

Fellegi and Sunter proposed an equation system to achieve the explicit formulas for the estimates of m(γ) and u(γ) when the matching variables are at most three. In more general situations, the conditional distribution estimates can be obtained via the EM algorithm [19], assuming a latent class model, in which the latent variable is just the link status.

According to the Fellegi and Sunter theory, once the composite weight r is estimated, it is possible to classify a pair as a link if the corresponding weight r is above a certain threshold T_m, and as a non-link if the weight lays below the threshold T_u; finally, for the pairs corresponding to weights falling into the range $I=(T_u, T_m)$, no-decision is made and the pair is assigned to a clerical review analysis. According to the Fellegi and Sunter theory, a decision on the threshold levels has to be made in order to properly manage the tradeoff between the need of a small number of expected no-decisions and small misclassified error rates for the pairs.

4. Examples

As an example of the application of the probabilistic record linkage method in T4 has been considered the integration process carried out by ISTAT between the administrative data on road accidents coming from police authorities and the administrative data on deaths by cause of death Registry. The integration process aims at identifying death persons, between road accidents survey and cause of death register, in the Toscana region and the 2008 year (see T4 for details).

5. Input data (characteristics, requirements for applicability)

At least two different data sources at micro level. The unique identifiers could be not available or affected by errors.

6. Output data (characteristics, requirements)

At the end, the record linkage process creates an integrated data set still composed by unit at micro level in order to produce counts or studying relationship between events and their determinants.

7. Tools that implement the method

The main use of the record linkage techniques in official statistics produced many software and tools both in the academic and private sectors, like BigMatch (Yancey, 2007), GRLS (Fair, 2001), Febrl (http://www.sourceforge.net/projects/febrl), Link Plus (http://www.cdc.gov/cancer/npcr/tools/registryplus/lp.htm), Tailor (Elfeky et al., 2002), etc.

In the ESSnet on Integration of Surveys and Administrative data (ISAD) the characteristics of some available software tools explicitly developed for record linkage and based on a probabilistic paradigm were analysed (see WP3, Chapter 1,section 1.1 and 1.3, Cibella et al., 2008c). The probabilistic record linkage tools that have been selected among the most well-known and adopted ones are: AutoMatch, developed at the US Bureau of Census, now under the purview of IBM [Herzog et al., 2007, chap.19]; Febrl - Freely Extensible Biomedical Record Linkage, developed at the Australian National University [FEBRL]; Generalized Record Linkage System (GRLS), developed at Statistics Canada [Herzog et al.,2007, chap.19]; RELAIS, developed at ISTAT [RELAIS]; The Link King, commercial software [LINKKING]; Link Plus, developed at the U.S. Centre for Disease Control and Prevention (CDC), Cancer Division [LINKPLUS].

8. Appraisal

8.1. Discuss the strengths and short-comings (gap analysis)

The record linkage can be seen as a complex process involving different knowledge areas and, in case of the National Statistical Institutes (NSIs), the integrated use of statistical and administrative sources is a product of a rationalization of all the available sources needed for several applications.

The Fellegi-Sunter and Jaro method is recommended when unique identifiers are not available for all the units or when they are affected by errors.

Some manual check are required to be sure to avoid false matches and missing matches. Without the availability of generalized tools for probabilistic record linkage, the knowledge of probabilistic record linkage methodologies should be considered a disadvantage. Fortunately, several tools are available for performing such integration procedure, e.g. Relais is particularly set to official statistics tasks (Relais, 2015).

The Fellegi-Sunter and Jaro approach is heavily dependent on the accuracy of $m(\gamma)$ and $u(\gamma)$ estimates. In case of misspecifications in the model assumptions, or lack of information in the whole record linkage process it could be possible a loss of accuracy in the estimates.

8.2. How to evaluate the results of the method (measures of model fitting, accuracy measure, etc.)

In order to assess the "quality" of the procedure it is necessary to classify records as true link or true non link, minimising, according to the Fellegi and Sunter theory, the two types of possible errors: false matches and false non-matches that refers respectively to the matched records which do not represent the same entity and to the unmatched records not correctly classified. False non-matches of matching cases are the most critical ones because of the difficulty of checking and detecting them.

In general, it's not easy to find automatic procedures to estimate these types of errors so as to evaluate the quality of record linkage procedures. Errors can also be introduced by the choices that are made in the matching process itself. For instance, an incorrect or overly limited matching key may be used, the way in which the weights are calculated may be incorrect, or the cut-off values against which the weights are set off may lead to matching errors. Also the time consumed by software programmes and by the number of records that require manual review could be considered additional performance criteria for the process (see the WP1 of the ESSnet on ISAD, Section 1.7 for details (Cibella et al., 2008a)) or also, as stated in the module "Micro-Fusion – Object Matching (Record Linkage)", all the choices that are made in the matching process itself could have an impact on the record linkage quality (e.g., an incorrect or overly limited matching key).

The final step of the whole record linkage process is devoted to the subsequent studies of the linked data set, taking in mind that this file can contain matching errors and all the derived analysis could be affected by the two types of errors: the percentage of incorrect acceptance of false matches and, on the other hand, the incorrect rejection of true matches. Record linkage procedures must deal with the existing trade-off between these two errors and/or measure the effects on the parameter estimates of the models that are associated to the obtained files.

8.3. Alternative methods if relevant

The standard statistical methods for probabilistic record linkage are by Fellegi and Sunter (1969) and Jaro (1989). They propose to define the record linkage problem as a classification one, where the matching status is unknown and needs to be estimated by means of a decision/classification rule.

In alternative, the record linkage problem can be faced in a Bayesian framework, (Fortini et al, 2001, Tancredi and Liseo, 2011) where prior distribution of the linkage probabilities and the number of matches and posterior distributions are simulated using MCMC. More recently, Steorts et al (2014) propose a different Bayesian approach.

In official statistics, some research activities are devoted to solve recent issues related to the use of record linkage techniques. One of the active research fields is the privacy preserving record linkage (PPRL), related to the use of pseudo-anonymized data, so as to preserve data from privacy issues and be able to link data provided in a more protective framework; see Fienberg, (2010) and Christen (2012) and for the most comprehensive survey Vatsalan et al, (2013).

Another topic is the simultaneous linkage of more than two sources, the so called multiple record linkage (Sadinle et al.). The longitudinal nature of some linkage process, that involves updated version of administrative data has to be taken into account as well. In these cases, the statistical estimation methods are sometimes confused with IT solutions as incremental record linkage Gruenheid et al (2014), and parallel record linkage Christen et al (2002).

9. References

Chambers R. (2009). Regression Analysis Of Probability-Linked Data. Official Statistics Research Series 4.

Christen, P. (2012). Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Heidelberg: Springer, Chapter 8: Privacy Aspects of Data Matching.

Christen P., Hegland M., Roberts S., Nielsen O. M., Churches T., Lim K. (2002). Parallel computing techniques for high-performance probabilistic record linkage available at http://datamining.anu.edu.au/linkage.html.

ESSnet Statistical Methodology Project on Integration of Survey and Administrative Data (ISAD), Report of WP2. Recommendations on the use of methodologies for the integration of surveys and administrative data,

2011, available at https://ec.europa.eu/eurostat/cros/content/deliverables-wp2di-integration-applications-isad-wp2_en.

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

Fienberg S. Hall R. (2010) Privacy-preserving record linkage, in: Privacy in Statistical Databases, Springer Lecture Notes in Computer Science, vol. 6344, Corfu, Greece, 2010, pp. 269–283.

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, 2006.

Fortini M., Liseo B., Nuccitelli A., Scanu M. (2001), On Bayesian record linkage, Research In Official Statistics, 4, 185–191.

Gruenheid A. Dong X. L., Srivastava D. (2014) Incremental Record Linkage. Proc VLDB Endowment 7 (9):697-708

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, 84, 414–420.

Sadinle M., Hall R., and Fienberg S. Approaches to Multiple Record Linkage available at https://www.cs.cmu.edu/~rjhall/ISIpaperfinal.pdf

Shlomo, N., & Skinner, C. (2010). Assessing the protection provided by misclassification-based disclosure limitation methods for survey microdata. Annals of Applied Statistics, 4(3), 1291-1310. DOI: 10.1214/09-AOAS317. Publication link: fbb6d0a6-32fd-41c0-9aac-6db410555bab

Steorts, Hall, Fienberg (2014) SMERED: A Bayesian Approach to Graphical Record Linkage and De-duplication, a new approach in the Bayesian framework, Journal of Machine learning research

Tancredi, Liseo (2011) Hierarchical Bayesian approach to record Linkage and population size problems. The Annals of Applied Statistics, 2011, Vol. 5, No. 2B, 1553—1585

Vatsalan, Christen, Verykios (2013) A taxonomy of privacy-preserving record inkage techniques Information Systems 38 (2013) 946–969.

1. Purpose of the method

When using information from different sources, the composite records may consist of several combinations of sources. The combination may give rise to consistency problems because the information is conflicting in the sense that edit rules that involve variables obtained from the different sources are violated. The purpose of reconciling conflicting microdata is to solve the consistency problems by making small adjustments to some of the variables involved.

2. The related scenarios

2.1. Direct tabulation exploiting multiple administrative sources: Indirect estimation where administrative and statistical data are used on an equal footing. Data configuration 4 there exists overlap of units and measurements between the different data sources. See deliverable 1 and 2.

2.2. Statistical tasks: Data editing and imputation, measurement alignment.

3. Description of the method

In the *minimum adjustment method*, the values in the record with inconsistent microdata are changed, as little as possible, such that the modified record is consistent in the sense that it satisfies all edit rules. It can be described as minimising a chosen distance between the original (inconsistent) record and the adjusted record, subject to the constraint that all edit rules are satisfied by the adjusted record. The optimisation approach resolves inconsistencies in data records with numerical variables that are required to adhere to a set of specified linear edit rules. The numerical variables in a record are denoted by x_i with $i=(1,\ldots,n)$ and can be represented as a vector of variables: $\mathbf{x}=(x_1,x_2,\ldots,x_n)$. The general form of a linear edit rule is as follows:

$$e_{j1}x_1+\cdots+e_{jn}x_n-c_j=0, \qquad (1)$$

for equalities and

$$e_{j1}x_1+\cdots+e_{jn}x_n-c_j\geq 0 \qquad (2)$$

for inequalities. Where $j=(1,\ldots,J)$ numbers the edit rules, e_{ji} are numerical coefficients and c_j are numerical constants.

To describe the minimum adjustment methods it is convenient to express the edit rules in matrix notation. The equalities (1) can be expressed as $\mathbf{Ex}=\mathbf{c}$, with E the $J\times n$ "edit matrix" with elements e_{ji} and c the *J*-vector with elements c_j.

It is necessary to take the distinction between free variables that are allowed to be adjusted and fixed variables that are not. The complete data vector can be partitioned into $\mathbf{x}_{fre}$ for the free variables and $\mathbf{x}_{fix}$ for the fixed ones. A corresponding partitioning of the edit matrix yields, say, $\mathbf{E}_{fre}$ and $\mathbf{E}_{fix}$. Now we can write

$$\mathbf{Ex}=\mathbf{E}_{fre}\mathbf{x}_{fre}+\mathbf{E}_{fix}\mathbf{x}_{fix}=\mathbf{c},$$
$$\text{and so } \mathbf{E}_{fre}\mathbf{x}_{fre}=\mathbf{c}-\mathbf{E}_{fix}\mathbf{x}_{fix}, \text{ which can be expressed as } \mathbf{Ax}_{fre}=\mathbf{b}.$$

The r.h.s. of this last expression contains all constants including the values of fixed variables and the l.h.s. contains the free variables that may be changed. They are the actual variables for the optimisation problem. For ease of notation we will, in the context of the optimisation problem, simply write x for the relevant, not

fixed, variables and suppress the suffix *fre*. Thus we will write $\mathbf{Ax} = \mathbf{b}$ for the constraints on the relevant variables.

In addition to the equality constraints we also often have linear inequality constraints. The simplest case is the non-negativity of most economic variables. The optimisation approach can also handle linear inequality constraints. The constraints can then be formulated as $\mathbf{A}_{eq}\mathbf{x} = \mathbf{b}_{eq}$ and $\mathbf{A}_{ineq}\mathbf{x} \geq \mathbf{b}_{ineq}$, where $\mathbf{A}_{eq}$ contains the rows of A corresponding to the equality constraints and $\mathbf{A}_{ineq}$ the ones corresponding to the inequality constraints. For ease of exposition we shall, without noting otherwise, write these equality/inequality constraints more compactly as $\mathbf{Ax} \geq \mathbf{b}$

With the notation and conventions introduced above we can write the optimisation approach to the problem of finding the smallest possible adjustments compactly as

$$\begin{aligned} &\tilde{\mathbf{x}} = \arg\min_{\mathbf{x}} D(\mathbf{x}, \mathbf{x}_0) \\ &s.t. \quad \mathbf{A}\tilde{\mathbf{x}} \geq \mathbf{b} \end{aligned}, \tag{3}$$

with $\mathbf{x}_0$ the adjustable part of the record *before* adjustment and $\tilde{\mathbf{x}}$ the corresponding sub-record *after* the adjustment and $D(\mathbf{x}, \mathbf{x}_0)$ a function measuring the distance or deviance between $\mathbf{x}$ and $\mathbf{x}_0$. In Memobust (2014b), the optimisation problem is studied under the choice for the distances given by the weighted least square adjustments, and the Kullback-Leibler divergence.

There are other methods to reconcile data, *prorating* and *generalised ratio adjustments.* Prorating is a simple ratio adjustment for balance edits. It solves the possible inconsistencies for each constraint separately. The generalised ratio adjustments method aims to make the adjustments as uniform as possible. For more details see Memobust (2014a) and Memobust (2014c).

4. Examples

In this illustrative example, we suppose the total turnover *(Turnover)*, the number of employees *(Employees)* and total amount of wages paid (*Wages*) are observed in a data source (register) that is considered highly reliable. A sample survey is conducted to obtain the additional details. After linking the sample data to the register, the situation arises that for the key variables, two sources are available for each responding unit in the sample. To be consistent with already published figures on *Turnover* and possibly other key variables, the register values are used for the key variables and the survey values for the other variables.

Let us suppose having an observation like that in Table 1 (in bold the values observed in the Register).

Table 1.

Variable	***Name***	***Survey values***	***Composite***
$\mathbf{x_1}$	Profit	330	**330**
$\mathbf{x_2}$	Employees (Number of employees)	20	**25**
$\mathbf{x_3}$	Turnover main (Turnover main activity)	1000	**1000**
$\mathbf{x_4}$	Turnover other (Turnover other activities)	30	**30**
$\mathbf{x_5}$	Turnover (Total turnover)	1030	**950**
$\mathbf{x_6}$	Wages (Costs of wages and salaries)	500	**550**
$\mathbf{x_7}$	Other costs	200	**200**
$\mathbf{x_8}$	**Total costs**	**700**	**700**

In such a context, there are logical relations (edits) that should be satisfied:

e_1: $x_1 - x_5 + x_8 = 0$ (*Profit = Turnover – Total Costs*)

e_2: $- x_3 + x_5 - x_4 = 0$ (*Turnover = Turnover main + Turnover other*)

e_3: $- x_6 - x_7 + x_8 = 0$ (*Total Costs = Wages + Other costs*).

The composite record fails the edits, so we need to adjust values in order to make data pass the edits.

The edits can be expressed in the form Ex = c with

$$\mathbf{E} = \begin{pmatrix} 1 & 0 & 0 & 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & -1 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & -1 & 1 \end{pmatrix} \text{ and } \mathbf{c} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

The adjustment problem was to adjust the survey values such that the edit rules are satisfied while leaving the administrative values unchanged. Results based on least square (LS) and Kullback-Leibler (KL) adjustments are shown in Table 2

Table 2

Vari-able	***Name***	***Composite record I***		
		Unadj.	LS	**KL**
$\mathbf{x_1}$	Profit	330	260	**249**
$\mathbf{x_2}$	Employees	25	25	**25**
$\mathbf{x_3}$	Turnover main	1000	960	**922**
$\mathbf{x_4}$	Turnover other	30	-10	**28**
$\mathbf{x_5}$	Turnover	950	950	**950**
$\mathbf{x_6}$	Wages	550	550	**550**
$\mathbf{x_7}$	Other costs	200	140	**151**
$\mathbf{x_8}$	**Total costs**	**700**	**690**	**701**

The LS adjustment procedure leads to one negative value for *Turnover other*, which is not allowed for this variable. Therefore the LS-procedure was run again with a non-negativity constraint added for the variable *Turnover other.* This results simply in a zero for that variable and a change in *Turnover main* to ensure that *Turnover = Turnover main + Turnover other.* Without the non-negativity constraint, the LS-results clearly show that for variables that are part of the same constraints (in this case the pairs of variables x_3, x_4 and x_6, x_7. that are both appearing in one constraint only), the adjustments are equal: -40 for x_3, x_4 and -16 for x_6, x_7. The results for the KL solution show that the adjustments are larger, in absolute value, for large values of the survey variables than for smaller ones. 5. Input data (characteristics, requirements for applicability)

At least two different data sources at micro level. The unique identifiers could be not available or affected by errors.

5. Input data (characteristics, requirements for applicability)

Data records with possibly inconsistent values and edit rules.

6. Output data (characteristics, requirements)

The output consists of the same individual records as the input, with values adapted when needed to ensure consistency with the edit rules.

7. Tools that implement the method

For the weighted least square adjustment the R package RSPA.

8. Appraisal

The method should be used after detection and treatment of missing values.

When inconsistencies arise due to large errors in some values, these errors may propagate to other values due to adjustment. Influential errors should therefore be treated before the method is applied.

9. References

Memobust (2014a). Prorating, in *Memobust Handbook on Methodology of Modern Business Statistics.* https://ec.europa.eu/eurostat/cros/content/memobust_en

Memobust (2014b). Minimum adjustments methods, in *Memobust Handbook on Methodology of Modern Business Statistics.* https://ec.europa.eu/eurostat/cros/content/memobust_en

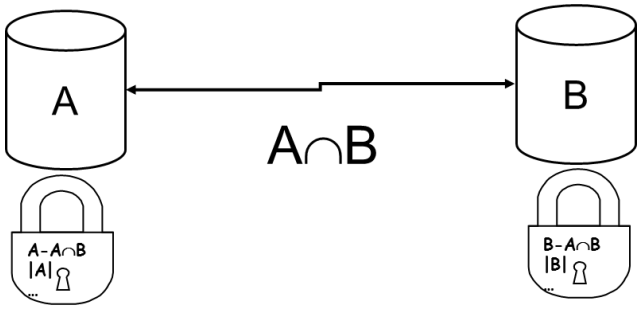
Memobust (2014c). Generalised Ratio Adjustments, in *Memobust Handbook on Methodology of Modern Business Statistics.* https://ec.europa.eu/eurostat/cros/content/memobust_en

Pannekoek J. (2011). Models and algorithms for micro-integration, in Report on WP2: Methodological developments, ESSNET on Data Integration, available at http://www.essnet-portal.eu/di/wp2-development-methods.

1. Purpose of the method

Hashing and anonymization techniques has recently become very popular (Fienberg 2010, Christen 2012), mainly due to the several application in Privacy Preserving Record Linkage (PPRL). Privacy Preserving Record Linkage (PPRL) has the purpose of performing record linkage between two sources, say A and B, such that at the end of the process A will know only a set A ∩ B, consisting of records in A that match with records in B. Similarly B will know only the set A ∩B. Of particular importance is the aim that no information will be revealed to A and B concerning records that do not match each other. Figure 1 depicts a private record linkage scenario, in which the information that each of the two sources wants to keep secret is represented within a padlock (see Batini and Scannapieco,2006).

Figure 1.



Nowadays NSIs live in an environment in which there is the need to share information with other subjects, and to possibly integrate owned information with other sources. These can be considered as the principal applications of PPRL techniques. More specifically, when considering networked information systems, the relationship between record linkage and privacy can be characterized in two ways (Batini and Scannapieco,2006):

- Record linkage in data publishing. A node can publish its own data so that they are available to the whole networked system. If privacy constraints must be enforced on published data, the node must ensure that no record linkage could be done on published data with the purpose of identifying identities of published individuals and entities.
- Record linkage in data integration. Nodes joining a networked system are willing to share information with other nodes. If such information need be privacy protected, record linkage should be enabled while preserving privacy.

In data publishing, a major problem is to assess the risk of privacy violation, once properly disclosed data are published. Typically, anonimyzation does not guarantee zero privacy risk. Indeed, suppose that personal data like `DateOfBirth`, `City` and `MaritalStatus` are published, whereas identifiers like `SSN`, `Name` and `Surname` are removed for the purpose of privacy preservation. By performing record linkage of such data with a public available list, such as an electoral list, it can be easy to identify the individuals with whom the publish data are referenced. Therefore, more sophisticated techniques, like the ones proposed in the this contribution, need to be applied for more properly dealing with privacy assurance.

Data integration becomes more and more important with the availability of new sources (see e.g. (Schnell, R. ,2014). Let's think for instance to Big Data sources: these are data collected or available not for statistical purposes, hence using them poses privacy issues. PPRL techniques could be a valid tool to overcome at least some of the issues related to the combination of Big data sources with survey and/or administrative data. Finally, in domains, like the Health one, that inherently exhibit privacy requirements, hashing and anonymization techniques may allow to share data in well-established privacy constrained between different partners.

2. The related scenarios

2.1. Usages and Komuso Data Configurations (deliverable 1): Hashing and anonymization techniques may be applied to all the usages (both direct and indirect) delineated in Deliverable 1, that involves record linkage procedures between more than one source and requires the input and output data at micro level when privacy has to be preserved. Privacy preserving linkage procedures may be required in the first and most basic data configuration 1 (see deliverable 1), called the "split-variable" case, concerning multiple cross-sectional data sources covering the target population where the different data sets contain different target variables and common variables able jointly to identify the units. This privacy preserving linkage step is also needed in the basic data configurations 4 and 5; the former is characterised by multiple cross-sectional data sources covering the target population with overlapping units between the different data sources, the latter is characterised by under coverage of the target population in the different integrated sources.

2.2. Statistical tasks (deliverable 2): The privacy preserving probabilistic record linkage can be adopted for the *Creation of joint statistical data, combining data at micro level belonging to the same unit*.

2.3. Possible competing methods or related methods could be quoted in this part: The standard techniques for probabilistic record linkage (see Deliverable 5, XX) can be an alternative to the hashing and anonymization techniques used in privacy preserving probabilistic record linkage, if there are not constrains in terms of privacy.

3. Description of the method

Several types of anonymisation techniques can be listed:

3.1. Secure hash encoding. One-way hash encoding functions (Schneier-1996) converts a string value into hash-code such that having access to only a hash-code will make it nearly impossible with current computing technology to learn its original string value. A major limit of this technique is that only exact matches can be found (Dusseree-1995).

3.2. Embedded space, based on the idea of mapping based blocking. Attribute values are embedded (mapped) into a metric space, while the instances between values are preserved (Scannapieco-etal-2007).

3.3. Pseudo random functions. A pseudo-random function is a deterministic that when given an n-bit seed k, and an n-bit argument x, it returns an n-bit string such that it is infeasible to distinguish it for random k from a truly random function (Luby-1986).

3.4. Phonetic encoding. A phonetic encoding algorithm groups values together that have a similar pronunciation, and has the main advantage of inherently providing privacy by the encoding itself.

3.5. Bloom filters. A bit-string data structure of length l bits where all bits are initially set to 0. k independent hash functions, h1,h2, . . . ,hk, each with h range1, . . .l, are used to map each of the elements in a set s into the Bloom filter by setting k corresponding bit positions to 1 (Bloom-1970).

Other privacy techniques doesn't require anonymization of variables/strings, for instance: Secure multi-party computation (SMC), where a computation is secure if at its end no party knows anything except its own input and the final results. SMC techniques employ some form of encryption schemes to allow secure computation; Generalization techniques, where data are generalized in such a way that re-identification from the perturbed data is not possible. A type of generalization is k-Anonimity, which is defined for tables in relational databases. Let us consider as a quasi-identifier an attribute that can be used to identify individual entities, a table satisfies the k-Anonimity criteria if every combination of quasi-identifier attributes is shared

by at least k tuples; Random values, that consist of adding random noise in the form of extra records to the data sets (this is called a data perturbation technique) (Kargupta et al 2003); Differential privacy, a recent technique (Inan et al 2010) that allows parties to interact with each other's databases using statistical queries, and only the perturbed results of a set of statistical queries are then discarded to other parties; Reference Values. Reference values result from reference lists that can be constructed either with random faked values, or values that for example are taken from a public telephone directory, such as all unique surnames and town names. This list of reference values can be used to calculate the distances between their attribute values and the reference values.

Finally, a part of the privacy techniques, other dimentions of the privacy requirements of PPRL among organizations mainly refer to the number of parties and the adversary model. As far as the number of parties, solutions may be classified into those that require a third party and those that do not. Two party protocols are more secure than three parties protocols, because there is no possibility of collusion between one of the data set owners and the third party. At the same time, they typically are more complex. Regarding the adversary model, two scenarios are possible, the former is the honest-but-curious behaviour, where parties are curious, i.e. they follow the protocol, but try to find out information about the other party's data; the latter the malicious behaviour, where parties can behave arbitrarily. Several surveys on PPRL have recently been published, e.g. Trepetin (2008) and Verykios (2009). The most comprehensive survey is Vatsalan and Verykios (2013) and Schnell (2015).

4. Examples

A relevant example of Hashing and anonymization techniques in official Statistics is the Program Beyond 2011 (ONS, 2013), where the UK NSI tested the reliability of matching procedure applied on pseudo-anonymised large dateset including 2011 Population Census and some administrative registers. Other interesting applications come from healthy registers (cancer registers) in Germany (Hundepool et al, 2012) and UK (Smith and Shlomo, 2015).

5. Input data (characteristics, requirements for applicability)

Hashing and anonymization techniques require as input data micro-level values for the variables one wants to make anonymous.

6. Output data (characteristics, requirements)

As output data the Hashing and anonymization techniques provide values at micro-level that avoid to reveal their initial meaning but can be compared with standard record linkage techniques performing approximate matching.

7. Tools that implement the method

Some codes already available in R, Python, Java, to adapt to specific usage.

8. Appraisal

8.1. From a practical perspective, there are some relevant aspects to be considered when applying hashing functions or anonymization techniques: first of all, the importance of pre-processing (standardising, cleaning etc.) of data prior to it being hashed (or equivalently anonimized); secondly the need of comparing hashed (or anonymised) data in a way that still allows us to detect agreement between similar but different values (e.g. John and Jon), i.e. allowing approximate matching. Many of the

illustrated techniques have indeed the purpose of "preserving" distances in order to perform approximate matching.

8.2. Theoretical analysis and evaluation of anonymitazion techniques make reference to Linkage quality is analysed in terms of fault tolerance of the matching technique to data errors. Fault-tolerance to errors can be addressed by using approximate matching or pre-processing. Assuming that truth data are available (which is frequently not the case in PPRL applications), the linkage quality can be assesses using any of the common RL measures. Main privacy vulnerabilities include frequency attack, dictionary attack and collusion between parties. Privacy evaluation can be performed with various measures, we mention two of them: A. Entropy, Information gain (IG) and relative information gain (RIG). Entropy H(Y) measures the amount of information contained in a message X. The information gain IG assesses the possibility of inferring the original message Y, given its enciphered version X. The RIG measure normalizes the scale of IG to the interval (0..1). Since RIG values are normalized between 0 and 1, they provide a scale for comparison and evaluation. B. Security/simulation proof. The proof of privacy can be evaluated by simulating the solutions under different adversary models. If under a certain adversary model a party learns no information except its input and output, the technique can be proven to be secure and private. Finally, another dimension to consider in evaluating an anonymization technique is the scalability, because we have also to take into account communication costs between parties.

9. References

Fienberg S., Hall R.,(2010) Privacy-preserving record linkage, in: Privacy in Statistical Databases, Springer Lecture Notes in Computer Science, vol. 6344, Corfu, Greece, 2010, pp. 269–283.

Christen, P. (2012). Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Heidelberg: Springer, Chapter 8: Privacy Aspects of Data Matching.

Batini C., Scannapieco M. (2006) Data Quality: Concepts, Methodologies and Techniques. Data-Centric Systems and Applications, Springer, ISBN 978-3-540-33172-8

Batini C., Scannapieco M. (2014) Information Quality: from Data to Big Data. Springer

Trepetin S. (2008) Privacy-preserving stringcomparisons inrecordlinkage systems: a review, Information Security Journal: A Global Perspective 17, 253–266.

Verykios V.S., Karakasidis A., Mitrogiannis V. (2009) Privacy preserving record linkage approaches, International Journal of Data Mining, Modelling and Management 1 (2) 206–221.

Vatsalan, Christen, Verykios (2013): A taxonomy of privacy-preserving record inkage techniques Information Systems 38 946–969

Schneier B. (1996), Applied Cryptography: Protocols, Algorithms, and Source Code in C, 2nd edition, John Wiley & Sons Inc., New York

Dusserre L., Quantin C., Bouzelat H. (1995), A one way public key cryptosystem for the linkage of nominal files in epidemiological studies, Medinfo 8 644–647.

Scannapieco M., Figotin I., Bertino E., Elmagarmid A. (2007), Privacy preserving schema and data matching, in: ACM SIGMOD, Beijing, China, pp. 653–664.

Kargupta H., Datta S., Wang Q., Sivakumar K., (2003) On the privacy preserving properties of random data perturbation techniques, in: IEEE ICDM, Florida, USA, pp. 99–106.

Inan A., Kantarcioglu M., Ghinita G., Bertino E. (2010) Private record matching using differential privacy, in: EDBT, Lausanne, Switzerland, pp. 123–134.

Luby M., Rackoff C., (1986) How to construct pseudo-random permutations from pseudo-random functions, in: CRYPTO, vol. 85, , p. 447.

Bloom B., (1970) Space/time trade-offs in hash coding with allowable errors, Communications of the ACM 13 (7) 422–426

Al-Lawati A., Lee D., McDaniel P., (2005) Blocking-aware private record linkage, in: International Workshop on Information Quality in Information Systems, Baltimore, MD, USA, pp. 59–68.

Christen P., Pudjijono A., (2009) Accurate synthetic generation of realistic personal information, in: PAKDD, Lecture Notes in Artificial Intelligence, vol. 5476, Springer, Bangkok, Thailand, 2009, pp. 507–514.

Christen, P. (2012). Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Heidelberg: Springer, Chapter 8: Privacy Aspects of Data Matching.

ONS (2013) Beyond 2011: Matching Anonymous Data http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/index.html

Schnell, R. (2014) An efficient privacy-preserving record linkage technique for administrative data and censuses , Statistical Journal of the IAOS , 30(3), 263–270. DOI 10.3233/SJI-140833. Accessed October 1, 2014 from: http://iospress.metapress.com/content/16ux7285j3811466/fulltext.pdf

Schnell R. (2015). Privacy-preserving record linkage, in Methodological Developments in Data Linkage, by Harron K, Goldstein H, Dibben C, Wiley

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K. and de Wolf, P.-P., (2012). Statistical Disclosure Control. Chichester: John Wiley & Sons, Ltd.

Smith D. Shlomo N., 2015 Privacy Preserving Probabilistic Record Linkage NTTS https://ec.europa.eu/eurostat/cros/system/files/Smith-etal_NTTS2015_944_abstract_PPPRL_unblinded.pdf.