Peter Hackl and Michaela Denk

Data Integration: Techniques and Evaluation

Austrian Journal of Statistics
Volume 33 (2004), Number 1&2, 135-152

Data Integration: Techniques and Evaluation

Michaela Denk* and Peter Hackl**
* ec3 – Electronic Commerce Competence Center, Vienna
** University of Economics and Business Administration, Vienna

Abstract: Within the DIECOFIS framework, ec3, the Division of Business Statistics from the Vienna University of Economics and Business Administration and ISTAT worked together to find methods to create a comprehensive database of enterprise data required for taxation micro-simulations via integration of existing disparate enterprise data sources. This paper provides an overview of the broad spectrum of investigated methodology (including exact and statistical matching as well as imputation) and related statistical quality indicators, and emphasises the relevance of data integration, especially for official statistics, as a means of using available information more efficiently and improving the quality of a statistical agency's products. Finally, an outlook on an empirical study comparing different exact matching procedures in the maintenance of Statistics Austria's Business Register is presented.

Zusammenfassung: Im Rahmen von DIECOFIS gab es eine Zusammenarbeit von ec3, der Abteilung für Wirtschaftsstatistik der WU Wien und ISTAT zur Analyse von Datenintegrationsmethoden zur Erstellung einer umfassenden Unternehmensdatenbasis für Steuer-Mikro-Simulationsstudien durch Verknüpfung bestehender disparater Datenquellen. Dieser Artikel gibt einen Überblick über das breite Spektrum der untersuchten Methoden (v.a. Verfahren des exakten und statistischen Matching, aber auch Imputationstechniken) und relevante statistische Qualitätsindikatoren. Er betont außerdem die Bedeutung der Datenintegration, insbesondere für die amtliche Statistik, als Möglichkeit, vorhandene Daten effizienter zu nützen und die Qualität der Produkte eines öffentlichen Statistikanbieters zu verbessern. Abschließend wird ein Ausblick auf eine empirische Studie zum Vergleich unterschiedlicher exakter Matchingverfahren beim Abgleich des Unternehmensregisters der Bundesanstalt Statistik Österreichs gegeben.

Keywords: DIECOFIS, Official Statistics, Data Integration, Record Matching, Exact Matching, Statistical Matching, Quality Indicators.

1 Introduction

DIECOFIS (Development of a System of Indicators on Competitiveness and Fiscal Impact on Enterprises Performance, cf. DIECOFIS, 2003, and Roberti, 2004) is an EU-

funded international research project, coordinated by the Italian national statistical agency ISTAT. The main goal of the project is to foster the development of "best" policy impact and evaluation techniques in the field of taxation, to further the Lisbon objectives and EU governance. One problem in this policy area is that "facts" on the impact of taxation are charted with a high degree of approximation, in spite of extensive discussions, experts' and working groups' meetings and a crowd of reports; tax indicators have well-known pitfalls.

The basic idea of DIECOFIS is to develop a system of micro-founded indicators. It aims at (i) assembling a wide ranging system of statistical information including data from economic, tax and social insurance sources into an *integrated multi-source enterprise database*, and (ii) creating *micro-simulation models* for enterprise taxation in two European countries, Italy and the UK, with a view to eventually producing an "EU demonstrator" as a foundation for the development of similar models in the whole EU. For the creation of such a multi-source database of enterprise data as a basis of micro-simulations, *data integration*, mainly record matching, is a core issue of the project. The project shows the importance of data integration as a means of generating comprehensive statistical databases as a sound foundation for deliberate decision making.

The Austrian member of the consortium (the Division of Business Statistics from the Vienna University of Economics and Business Administration, ec3 – Electronic Commerce Competence Center, a non-profit research corporation, ST.AT – Statistics Austria, and the Statistical Department of the Austrian Economic Chamber) was mainly engaged in the first step of the project concerning data integration (a summary of results is given by Denk and Hackl, 2003). The data integration project surveyed available methods of data integration, to provide a critical assessment of different data integration methods with a focus primarily on statistical issues and to provide an overview of statistical indicators for quality measures of multi-source databases. All these activities have been seen in view of the concrete application within DIECOFIS. Denk and Oropallo (2002) surveyed the available methods of data integration. Denk, Inglese, and Calza (2003) discussed the relative merits of the methods in the context of their application to national statistics databases. Quality indicators for assessing multi-source databases were provided by Denk, Inglese and Oropallo (2003). An empirical study has been designed comparing the applicability of various integration procedures in the context of the ST.AT's business register and demonstrating the use of quality indicators for multi-source databases.

This paper discusses the relevance of data integration, especially for official statistics (Section 2), and provides an overview of the broad spectrum of investigated methodology (Section 3) and related statistical quality indicators (Section 4). Section 5 gives an outlook on the empirical study that is carried out at ST.AT. Finally, concluding remarks are summarized in Section 6.

2 Relevance of Statistical Data Integration

There is a trend towards sensible re-use of available databases in official statistics and other fields. For instance, Froeschl and Grossmann (2000) stress the increased re-use of administrative data sources. An example involving the *Dutch Virtual Census* is reported

in this volume (Linder, 2004). From an official statistics' point of view, data integration is of major interest as a means of using available information more *efficiently* and improving the quality of a statistical agency's products. By using integration methods, the *value added* that can be extracted from the existing stock of information can be greatly augmented. *Responder burden* may be reduced considerably, and time and money may be saved as additional collecting of data can be avoided, for instance, by substituting new surveys with (integrated) available data; in many situations, it is not possible to obtain the required data by new surveys. The creation and maintenance of *registers* and census and survey frames are another important and typical application of data integration methods.

As stated above, apart from the better exploitation of data resources, data integration techniques can also contribute to the quality of available data. *Different definitions* of a variable can be compared. *Data reliability* can be evaluated by combining datasets originating from different surveys containing the same variable. *Missing* or *invalid* data can be replaced. By matching a dataset with itself, *duplicates* can be detected and removed.

A situation where it is not possible to obtain the required data by new surveys occurred in the DIECOFIS project. The data necessary for micro-simulation modelling concerning enterprise taxation and performance are all available in distributed and heterogeneous sources. Enterprises would refuse to provide all the information once again. Hence, the information pieces have to be put together in the right way to enable micro-simulation. In that case, data integration is a means of generating a comprehensive statistical database as a sound foundation for deliberate decision making.

Statistics Austria, being an early adopter of data integration methods, contributes actively in the project by collaborating in the empirical study on measuring multi-source data quality obtained by different exact matching methods as outlined in Section 5.

3 Statistical Data Integration Methodology

Data integration is a broad field of research and can be viewed from various perspectives. In DIECOFIS, the main emphasis was on *statistical* data integration *methodology* and *quality indicators* for the assessment of different approaches and applications. However, also some *technical considerations* need to be addressed for multi-source database integration. In addition to technical considerations, semantic discrepancies and similarities of data sources need to be analyzed before the application of statistical methods generally and integration methods in particular: *Data source integration* as a prerequisite of *dataset integration*. A *metadata* oriented approach for the detection and formalised representation of semantic heterogeneities following the ideas and concepts of IDARESA (e.g., IDARESA, 1997, 1998, Denk and Froeschl, 2000, and Denk, Froeschl and Grossmann, 2002) was proposed. In this volume, Froeschl (2004) discusses the possible contributions of meta-computing, i.e., the processing of metadata alongside the accompanied statistical data as well as procedures for controlling the integration process based on metadata, to the integration of statistical

data and metadata. For a discussion of technical and metadata related integration aspects in DIECOFIS see Denk and Hackl (2003).

According to D'Orazio, Di Zio and Scanu (2001), two broad classes of statistical integration procedures can be distinguished, viz. (i) *micro procedures* integrating datasets at record level by combining records representing the same (or a similar) real-world entity in different datasets, and (ii) *macro procedures* where the main interest is on aggregates of the integrated data. Several different terms are used for micro data integration: the most common seem to be *object* or *instance identification* (e.g., Neiling, 1998, Neiling and Lenz, 1999, Wang and Madnick, 1989), *record matching* (e.g., Fellegi and Sunter, 1969, Winkler, 1995, Fair and Whitridge, 1997, FCSM, 1980, or Alvey and Jamerson, 1997), and *data fusion* (e.g., Raessler, 2002). *Exact* and *statistical matching* procedures as well as *imputation* methods fall into this category, while the macro category encompasses all kinds of *weighting* procedures (e.g. adapting estimates resulting from surveys in order to comply to population structures or parameters) and procedures for combining summary level data into one single table, as for instance Malvestuto's *Universal Table Model* (e.g., Malvestuto, 1989, 1991, 1993).

Exact matching is used when datasets with substantial overlap (with regard to observed entities as well as variables) are integrated, and matching of records belonging to identical entities (in our case: enterprises) is the goal. If this is not possible (or not essential for the intended usage of the combined dataset), e.g., because of different survey samples that rarely overlap, statistical matching (as an approximation of exact matching) can be used. (For a discussion of exact and statistical matching see, for instance, FCSM, 1980). Imputation is applied to replace missing or invalid values (e.g., item non-response, failed edits) by valid values. It is also closely related to statistical matching: imputation replaces missing or obviously erroneous values in a dataset, while statistical matching inserts values for variables not originally included in the survey.

In DIECOFIS micro integration strategies were investigated. For the creation of the integrated and systematised enterprise statistical information system needed for micro-simulation purposes, exact matching was used to combine administrative data (from the business register, commercial accounts, tax returns and foreign trade) and survey data. Statistical matching was relevant to integrate different ISTAT surveys (like structural business statistics and industrial production) that do not contain the same enterprises in order to reduce responder burden. Imputation was applied to complete still missing data. In the Austrian empirical study where ST.AT's business register is integrated with tax authority data, only exact matching was used.

3.1 Exact Matching

In case of availability of identifiers valid in all datasets to be combined, integration simply amounts to a natural database join on the basis of these identifiers. Yet, this ideal situation is rather unlikely. Even if datasets contain identifiers, their equivalence across datasets of different data sources is not necessarily provided. Usually other identifying characteristics (such as names or addresses of persons or enterprises) have to be taken into account which, in general, does not allow unique identification of identical units. Basically, exact matching methods classify all record pairs that can be built from source datasets into *non-links*, *possible* (i.e. *indeterminate*) *links*, and *links*. Possible links are

then clerically reviewed, and in most cases, linked pairs are checked to obtain a 1:1-assignment of records. In practice, in order to reduce the number of pairs that have to be investigated by the matching procedure, the set of all record pairs is decomposed into (i) *blocks* containing candidate pairs that agree on selected blocking variables which are then further analysed, and (ii) a residual set of determinate non-linked pairs that do not satisfy blocking criteria. A description of many of the practical problems that have to be dealt with in applying matching methods give Nikles and Müllauer (2003).

The following subsections briefly introduce exact matching methods. However, in most real-world applications, a combination of available methods seems to work best. A quite common pragmatic approach is to use deterministic linkage, followed by probabilistic linkage (including string comparators, if necessary), and followed by clerical review (Gill, 2001).

3.1.1 Quality Classes

In the quality class approach record pairs are assigned to different *compliance* or *quality classes* of record pairs based on their extent of agreement or disagreement on specified matching variables. By this means, a hierarchy of compliance classes is established. Record pairs in classes with high compliance ("*high quality match*") are linked, those in classes with low compliance are designated as non-links. Pairs in between are sent to clerical review. For a description of the quality class approach applied at Statistics Austria see Nikles and Müllauer (2003).

Usually, selection of variables as well as definition of classes is based on experience. Otherwise, the method is ad-hoc which makes the results hard to interpret. It is quite easy to implement and easy to use; yet, a disadvantage is that the clerical review region might be large. There is no underlying statistical model. Anyhow, there is danger of overfitting, since there are many parameters to be set. These parameters include the selection of variables, the combination of variables used, the setting of thresholds for each class and designation of classes as being link/non-link. Matching systems working with compliance classes might have to be adapted very often. If training samples with true matching status are available, a justification of used class definitions might be achieved by statistical classification algorithms, such as classification trees (cf. Breiman et al., 1984).

3.1.2 String Comparator Metrics

When comparing values of string variables like names or addresses, it usually does not make sense to just discern total agreement and disagreement. Typographical error may lead to many incorrect disagreements. Several methods for dealing with this problem have been developed: string comparators are mappings from a pair of strings to the interval [0, 1] measuring the degree of compliance of the compared strings (Winkler, 1990). String comparators may be used in combination with other exact matching methods, for instance, as input to probabilistic linkage, discriminant analysis or logistic regression. The simplest way of using string comparators for exact matching is to define compliance classes based on the values of the string comparator.

In order to make reasonable comparisons of string variables, adequate pre-processing by *standardizing* (i.e., replacing words of little distinguishing power with consistent abbreviations) and *parsing* (decomposing a string variable into a set of string components which are then individually compared) the strings is essential (cf. Winkler, 1995). This holds, in particular, when matching business data, since inconsistencies of name and address information are typically even greater for this kind of data (Winkler, 1999). Problems with addresses are due to the different types that might be used by an enterprise in different situations, such as the mailing address, the physical address, or the address of the lawyer. Apart from hybrid similarity methods (discussed below), the only "basic" string comparator that will work even if the order of different components of a string variable is not fixed for all records is the bigram method.

This method consists in comparing the *bigrams* that two strings have in common. A *bigram* is two consecutive letters of a string. The return value of the bigram function is the total number of common bigrams in the two strings divided by the average number of bigrams in the two strings (Porter and Winkler, 1997). Other bigram variants use a different denominator: Instead of the average number of bigrams the number of bigrams in the first (or in the second) string is used. Bigrams are known to be a very effective, simply programmed means of dealing with minor typographical errors. They are widely used by computer scientists working in information retrieval (Frakes and Baeza-Yates, 1992). Porter and Winkler (1997) have shown empirically that bigrams work well, and ST.AT has successfully used the bigram method in the update process of the business register.

An early string comparator is the *Damerau-Levenstein* (D-L) *Metric* (Damerau, 1964, Levenstein, 1966), which is in fact only one instance of a metric from the class of *edit distance metrics*. Its basic idea is the fact that any string can be transformed into another string through a sequence of changes via substitutions, deletions, insertions, and possibly reversals. The smallest number of such operations required to change one string into another divided by the maximum length of the two compared strings is a measure of the difference between them which is easily converted to a string comparator rating the degree of agreement of the two strings. For a discussion of several enhancements of the D-L metric see Hall and Dowling (1980).

Jaro (see for instance Winkler, 1985, 1990) introduced a string comparator more straightforward to implement and maybe more closely related to the type of human decisions in comparing strings than the D-L metric. Basically, it accounts for the proportion of common characters in both strings and the number of transpositions that have to be made to create the sequence of common characters of one string from the sequence of common characters of the other string. Several enhancements to the Jaro comparator have been developed, in particular by Winkler (e.g. Porter and Winkler, 1997).

Standard computer science string similarity measures, as, for instance, the longest common substring or the *longest common subsequence* (e.g., Hirschberg, 1975) can also be used as basis for string comparator metrics for exact matching.

Hybrid similarity measures can be regarded as second level similarity measures, as they take recourse to some other "basic" string comparator. The similarity of each component (separated by blanks) of the first string to each component of the second string is computed using the selected basic string comparator. Then, for each component of the first string, the maximum similarity to one of the components of the second string

is determined, and the similarity between the two strings is computed as average of these maximum similarities. This way, hybrid similarity measures help overcome problems occurring when strings are not properly parsed.

Several further string comparators are introduced in Gill (2001) and Cohen, Ravikumar and Fienberg (2003). In this volume, Schnell, Bachteler and Bender (2004) present a record-linkage toolbox for the comparison of the performance of various string similarity measures for German surnames. Another toolbox of string comparator methods is the SecondString package (SecondString, 2004, Cohen et al., 2003).

3.1.3 Probabilistic Record Linkage

In probabilistic record linkage (cf. Fellegi and Sunter, 1969, Kilss and Alvey, 1985, Alvey and Jamerson, 1997), the conditional probabilities of observing agreement (disagreement) on a matching variable given a pair is actually a match (or a non-match, respectively) are used to define *matching weights* measuring the evidence that a pair is a match or not. Usually, the dual logarithm of the likelihood ratio of these conditional probabilities is used as weight, with the probability given a true match in the enumerator. Each matching variable is associated an *agreement* and a *disagreement weight*. The individual variable weights are assembled to a composite matching weight for each record pair. *Weight thresholds* are then determined for the classification of record pairs into links, possible links and non-links based on fixed error levels.

To simplify the estimation of conditional probabilities a *conditional independence assumption* is made. More specifically, the comparison outcomes for different matching variables are assumed to be mutually statistically independent with respect to each of the conditional distributions. For instance, matching variables might include all variables that relate to names (name, middle name, surname, initials, etc.), and those relating to addresses (city name, street name, house number, etc.). Concerning this example, the conditional independence assumption says that in matches, errors in names are independent of errors in addresses, and that in non-matches, accidental agreement on name is independent on accidental agreement on address. In practice, this assumption is often violated.

The kind of linkage rule defined by Fellegi and Sunter (1969) is optimal in the sense that the number of possible links is minimised for fixed error levels. It is also intuitively appealing. If a particular comparison outcome consists primarily of agreements, then it is more likely to occur among matches than non-matches and the corresponding weight will be large. On the other hand, if the comparison outcome consists mainly of disagreements, the matching weight will be small.

In practice, matching weights are computed using some variant of the EM algorithm (Dempster, Laird and Rubin, 1977, Wu, 1983, Meng and Rubin, 1993).

3.1.4 Classification Methods

Micro data integration can also be viewed as a well-known statistical problem, viz. a *classification problem*. Record pairs have to be assigned to the class of matches or the class of non-matches, respectively. However, there is one problem: A training sample must be available to enable estimation of classification rules.

A classical choice of classification model is discriminant analysis. One approach based on discriminant analysis is the Belin-Rubin method (Belin and Rubin, 1995) which tries to predict class membership conditional on the matching weight assigned to record pairs. Usage of discriminant analysis based on original values of identifying characteristics or comparison outcomes instead of matching weights is also conceivable. Non-parametric methods whose applicability is independent of distribution assumptions, such as nearest neighbour approaches or classification trees, are often used (cf. Neiling 1998).

Another classification method that might be used is logistic regression. Again, comparison outcomes or matching weights may serve as input variables. Chatterjee and Segev (1992, 1994) suggest fitting a logistic regression model to estimate matching weights.

3.2 Statistical Matching

In statistical matching the linkage of data for the same real-world entity either is not sought or is not essential to the procedure (FCSM, 1980). Usually, datasets have very few (or no) entities in common. Thus, the linkage of data for similar entities rather than for the same entity is acceptable as a goal. Actually, except in rare cases, linked records do not represent real-world entities, but rather what is referred to as a synthetic entity (Rodgers and DeVol, 1981), as opposed to exact matches, where, apart from erroneous assignments, linked records refer to identical entities.

Statistical matching originated in the field of economics, initially primarily targeting the combination of income data and data on tax returns (e.g., Okner, 1972, 1974, Radner, 1978, Radner and Muller, 1977). Statistically matched datasets have been used extensively in micro-simulation modelling (e.g., Cohen, 1991) to examine the impact of policy changes on population subgroups, and, hence, this suggests the suitability of statistical matched datasets for DIECOFIS tax simulation studies.

Among statistical matching methods, there are (i) techniques separating datasets into equivalence classes and then selecting records to be linked randomly, (ii) distance measures for the selection of most similar records, and (iii) regression-based techniques (see Kadane, 1978, Moriarity and Scheuren, 2001, Rodgers, 1984, or Raessler, 2002). Imputation techniques are very closely related to statistical matching (e.g., Kovar, Whitridge 1995). Essentially, statistical matching differs from imputation only with regard to its purpose: In a statistical match two different datasets are matched and (in almost all cases) the purpose is the addition of variables not present for any entity in the base dataset, whereas in imputation often only one dataset is used and values missing for several entities are completed.

3.3 Imputation

Imputation is used to reconstruct values missing for a record (item non-response, partial missing answers). If a full unit non-response (total missing answers) occurred (i.e., there is no record in the dataset for a sampled unit) usually macro integration procedures

(such as weighting) are utilized. A broad introduction to imputation and other types of missing data analysis is given in Little and Rubin (1987).

The simplest imputation approach is *deterministic* imputation, where all missing values of a variable are replaced with the same value, such as the mean, median or mode of the variable. If a large portion of a dataset has to be imputed, this method yields extremely unrealistic distributions with high peaks at the imputed values.

Model-based methods hypothesize a probabilistic relation between the variable with missing values and the matching variables. An auto-regression model is often used, so that the variable itself (taken from previous surveys) supplies the information. The probabilities for the occurrence of observed values of a variable are estimated. The imputation value is then randomly drawn from this probability distribution.

In donor-based approaches like *hot-deck* or *nearest-neighbour* imputation, the imputation value is taken from a so-called *donor*, which is a complete and correct record that is similar to the incomplete record. The similitude between donor and receiving record is determined via matching variables. Several donors might be available for the same record – then, one of them is chosen randomly.

Multiple imputation (Rubin 1987) is a simulation-based approach to the statistical analysis of incomplete data. Each missing value is replaced by $m>1$ simulated values. The resulting m versions of the complete data are then analysed by standard complete data methods, and the results combined to produce inferential statements (e.g. interval estimates or p-values) that incorporate missing data uncertainty. So, actually, multiple imputation is not one particular imputation algorithm, but rather a means of evaluation of imputation results.

4 Quality Assessment

No matter what the objective of data integration actually is, an evaluation of the applied procedures and the resulting multi-source database is indispensable. The quality of the integrated database will depend on various factors such as the quality of source data and the methods and variables used for integration. The evaluation may contain measures on the variability and reliability of parameters of the resulting database like standard error and bias, as well as method-specific or application-specific measures.

4.1 General Quality Criteria

We now discuss general data quality measures for input data. These quality measures include coverage of the underlying populations, correspondence of statistical units and congruence of variable and value definitions. Three further measures are completeness, which takes into account the proportion of missing data, consistency, which considers the proportion of observations failing edits, and the proportion of duplicates.

The quality of matching variables is crucial to any of the integration procedures presented. For its assessment, a precise definition of the concept captured, the amount of missing data, the discriminating power and the reliability should be reported (for measures of discriminating power and reliability see, for example, Hassard, 1986). Also

the correlation with target variables is important. For a reasonable overview of the usability of particular personal characteristics as matching variables see Jabine and Scheuren (1986) or Gill (2001). For enterprises, Winkler (2001) provides some empirical evidence.

Concerning the dataset resulting from integration, the distributions of matching and other variables can be tested with regard to deviations from their source distributions. Of course, for informative missing values, the preservation of the original distributions is not aimed at. To evaluate the completeness of the integrated database, the proportion of missing values and the number of records can be calculated as quality indicators and compared to the respective measures in the input datasets, and, if available, the population size.

Mean square errors of estimates (e.g., of correlation or regression coefficients) based on the integrated dataset are indicators for the quality of the estimates, and thus, also of the underlying dataset. Little work has been done to date on the evaluation of the accuracy of estimates of model parameters based on integrated data. Heuristic procedures for the estimation of the variability of the estimates are available. For instance, D'Orazio, Di Zio and Scanu (2001) propose the folded database procedure, a heuristic approach to get an idea of the bias of the association among the integrated variables introduced by statistical matching techniques. Sensitivity analysis can be used to estimate the variance of parameter estimates in an integrated database, as, for instance, suggested by Rubin (1987) with his idea of multiple imputation or the steepest ascent approach proposed by Winkler (1989) to be used in record linkage.

One further aspect to be taken into account when comparing alternative integration procedures is the complexity of the method, the corresponding implementation effort and the required computing time. Of course, the complexity (or simplicity) of an algorithm is not itself a measure of the quality of the algorithm. However, when comparing alternative procedures, a lower quality dataset, for instance in terms of a larger bias in estimates or a less accurate preservation of distributions of variables, may be accepted if computational requirements can be kept at a lower level.

4.2 Method-Specific Evaluation

In exact matching, misclassification rates and the size of the grey zone of possible matches are of particular interest. Depending on the matching aim, gross or net error rates may be considered. The accuracy of the estimation of error rates mainly depends on the availability of training data with known matching status, for instance, from similar previous applications, or a sample from the current data for which the true matching status is determined via clerical review. Moreover, quality indicators for individual processing stages (like blocking or 1:1-assignment) are available (e.g., Baxter, Christen and Churches, 2003). For the blocking stage, the reduction of matching error rates, the reduction of potential matches (i.e. record pairs that have to be considered in further processing), but also match completeness (that is, the proportion of matches "surviving" the blocking stage) can be estimated.

In statistical matching, where the linkage of records belonging to similar entities is sought for, error rates are not defined, since there is no "true matching status". Rather,

distributions of distances of linked records or the number of times individual records are used in linkage (in case that multiple linkage is enabled) are used as quality indicators.

Due to the similarity of imputation methods (especially donor-based imputation) and statistical matching procedures, imputation results may be judged by the same or at least similar criteria as statistical matching results.

5 Empirical Study

In order to fulfil the requirements of the European Union's register regulation, Statistics Austria has made efforts to create an Austrian business register (UR) from existing files of enterprises, taking into account several economic surveys and member data of the Austrian Chamber of Commerce, and further data sources of enterprise data, such as the commercial register, the register of agriculture and forestry, and the tax register.

The business register includes business units of three different types, viz. enterprises (organisational unit, in most cases also conforming to one legal unit), establishments, and local units. It encompasses identifying variables (such as ID, type of unit, name, legal form, status (active/inactive), date of formation); address variables (such as ZIP code, NUTS 3 code, street address); classification variables (e.g. NACE code); reference variables, which are basically keys to external business data sources (such as tax ID, ID in the commercial register); shipping variables (such as additional addresses, telephone or fax numbers, names of contact persons); and demographic variables (such as dates of closing or new formation of enterprises).

In the creation and maintenance of the register, one of the major tasks is the integration of the various data sources that are available to Statistics Austria. For business units that have already been entered into the register and for which linkages to different external data sources have already been established, further linkages may be simply achieved by using the respective foreign key contained in the UR which is tantamount to deterministic linkage. Other exact matching techniques must be applied to detect new business units and to find linkages to other data sources (i.e., to identify the foreign key of a unit in another database) for business units for which these particular linkages have not yet been set up. Currently, data integration is based on a compliance class approach where similarity of records is determined using the bigram method. A description of the implementation of the EU's register regulation at ST.AT is given in Schaumann (1999). A detailed presentation of the creation and maintenance process of the business register is given by Haslinger (2004) in this volume, or Nikles and Müllauer (2003) and Müllauer (2003).

The ST.AT data integration study within the framework of the DIECOFIS project is designed (a) to link business units in the UR to business units in the tax register in order to update the UR data with tax data, such as VAT- or income tax data and (b) to identify new business units in the tax register that are not yet contained in the UR so that the UR can be updated with respect to the set of tax-paying business units. This exercise allows the assessment of the applicability of several relevant database integration techniques, the comparison of already used methods with new ones, and gaining experience by assessing the integrated databases with various measures of database quality.

Based on methodological research carried out within the first author's dissertation project (cf. Denk, 2002) and within DIECOFIS (cf. Denk, Inglese and Calza, 2003), appropriate methods have been selected to be implemented and evaluated, to see if there is room for improvement or to derive recommendations for future applications, respectively.

In contrast to the integration task at ISTAT, where finding and linking identical units (enterprises) from different data sources is not necessarily sought for, or not even possible, and thus, statistical matching procedures are applied, the Austrian integration study requires the linkage of identical enterprises. For this reason, only exact matching procedures are envisaged, which also puts some constraints on the criteria that might be used to assess the quality of the matching result. Since, essentially, only string variables "name" and "address" are supplied for matching, the spectrum of applicable methods is rather limited. The applicability of string comparison methods is quite obvious.

In the blocking stage, the decision was made to stick to ST.AT's procedure using the ZIP code as primary blocking variable. Very large blocks occurring particularly in urban areas are then further subdivided using the initial letter of the street address. It should be mentioned that the implementation of efficient blocking procedures requires a high degree familiarity with the data.

In the string comparison stage, the Jaro algorithm as well as variants thereof, such as the one defined by Winkler, the edit distance and the longest common subsequence are used instead of ST.AT's bigram algorithms to compare name and address variables. Hybrid string comparators could also be applied.

In the matching stage, ST.AT uses a system of compliance classes defined by thresholds on bigram outcomes and agreement or disagreement on other variables like NACE code. Pairs in different classes obtain different follow-up processing. Those with the highest compliance are matched, for others, further variables may be taken into consideration or manual checks may be applied, and those with lowest compliance are dropped (for details cf. Nikles and Müllauer 2003). In the integration study, the simplest approach is the usage of the same compliance class definitions in order to solely evaluate the different string comparison algorithms. Another feasible approach is the usage of alternative compliance class definitions, but again, familiarity with the data is a crucial prerequisite for a good decision.

The application of matching procedures working with training samples drawn from the "benchmark dataset" provided by Statistics Austria is also taken into consideration. In particular, the estimation of a logistic regression function on the comparison outcomes and the classification of pairs via discriminant analysis, seem promising. For instance, CART or nearest neighbour approaches could be used to determine decision rules classifying the record pairs as links or non-links.

To enable the comparison of methods applied in the study to results of Statistics Austria, not only input data are supplied, but also intermediary datasets resulting from individual integration steps as well as the final assignment currently contained in the business register (which may have been attained by clerical review). This final assignment is accepted as being correct and used as the benchmark in the assessment of the validity of the assignments made by the tested integration methods.

The quality assessment of tested methods will include the computation of error rates (gross & net errors, false match rate, false nonmatch rates) and the number of links, non-links, and possible links (or the size of different quality classes, respectively), and

numbers of equal and deviating classifications, as well as the proportions of correct and erroneous classifications among these deviating classifications, and computing effort / complexity of algorithms. Different diagrams may be used to illustrate various error rates for different methods under different assumptions.

6 Concluding Remarks

DIECOFIS is an EU-funded international research project, coordinated by ISTAT. The objectives are the development of an appropriate methodology for the construction of a system of indicators on competitiveness and fiscal impact on enterprises performance and the illustrating application of the developed methods. Data integration, mainly record matching, and the generation of multi-source databases that are to be used as a basis of micro simulations are a core issue of the project. The project shows the importance of data integration as a means of generating comprehensive statistical databases as a sound foundation for deliberate decision making.

The Austrian member of the consortium is mainly engaged in the issues of database integration. Contributions include the surveying of available methods, a critical assessment of different data integration methods with a focus primarily on statistical issues, and an overview of assessment criteria for multi-source databases from a theoretical perspective, in particular statistical indicators of multi-source database quality. These contributions are made with a focus on DIECOFIS goals and requirements. An empirical study has been designed that compares the applicability of various integration procedures in the context of the Austrian business register and that demonstrates the use of quality indicators for multi-source databases. Preliminary results on string comparisons indicate that the choice of methods is not crucial in determining the overall quality of resulting dataset. We will wait for the final results before making firm, reliable conclusions. What can be stated is the importance of (i) input data quality, (ii) adequate data pre-processing, and (iii) knowledge of the data.

Particularly for official statistical agencies, the integration of datasets is of major interest as a means of using available information more efficiently and of improving the data quality. The quality of the results of using integration methods is naturally limited, as matching errors are, of course, inevitable. Hence, assessment of integration results with respect to appropriate quality criteria is strongly recommended and should also be reported to users of the multi-source dataset.

References

W. Alvey and B. Jamerson, editors. *Record Linkage Techniques*. Federal Committee on Statistical Methodology (FCSM), Washington, DC, 1997.

R. Baxter, P. Christen, and T. Churches. A Comparison of Fast Blocking Methods for Record Linkage. To appear in *Proc*. *First Workshop on Data Cleaning, Record Linkage, and Object Consolidation, 9th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining*, Washington, DC, 2003.

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

L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Wadsworth, Monterey, 1984.

A. Chatterjee and A. Segev. Resolving Data Heterogeneity in Scientific Statistical Databases. In H. Hinterberger and J.C. French, editors, *Proc. 6th Int. Conf. on Scientific and Statistical Database Management*, pages 145-159. ETH Zürich, 1992.

A. Chatterjee and A. Segev. Supporting Statistics in Extensible Databases: A Case Study. In H. Hinterberger and J.C. French, editors, *Proc. 7th Int. Conf. on Scientific and Statistical Database Management*, pages 54-63. IEEE Computer Society, 1994.

S. Cohen. Micro-simulation of Firm Investment. Presented at the *Symposium on Economic Modelling*, London University, 1991.

W.W. Cohen, P. Ravikumar, and S.E. Fienberg. A Comparison of String Distance Metrics for Name-Matching Tasks. Submitted to *18th International Joint Conference Workshop on Information Integration on the Web*, 2003. Also available at *http://www.niss.org/dg/technicalreports.html*.

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

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

M. Denk *Statistical Data Combination: A Metadata Framework for Record Linkage Procedures*. Dissertation, Department of Statistics and Decision Support Systems, University of Vienna, 2002.

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

M. Denk, K.A. Froeschl and W. Grossmann. Statistical Composites: A Transformation-bound Representation of Statistical Datasets. In J. Kennedy, editor, *Proc. 14th Int. Conf. Scientific and Statistical Database Management* (Edinburgh, UK), pages 217-226. IEEE Computer Society Press, Los Alamitos, Ca., 2002.

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

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

M. Denk, F. Inglese, and M.G. Calza. *Assessment of Different Approaches for the Integration of Sample Surveys*. DIECOFIS Deliverable 1.2, ISTAT, Rome, 2003.

M. Denk, F. Inglese, and F. Oropallo. *Report on Statistical Indicators for the Assessment of Multi-source Databases*. DIECOFIS Deliverable 1.3, ISTAT, Rome, 2003.

DIECOFIS. DIECOFIS Web Site, *http://petra1.istat.it/diecofis/index.html*, 2003.

M. D'Orazio, M. Di Zio, and M. Scanu. Statistical Matching: a tool for integrating data in National Statistical Institutes. In *Proc. of the Joint ETK and NTTS Conference for Official Statistics*, Crete, 2001.

M.E. Fair and P. Whitridge. Tutorial on Record Linkage. In W. Alvey and B. Jamerson, editors, *Record Linkage Techniques*, pages 457-479. FCSM, Washington, DC, 1997.

FCSM – Federal Committee on Statistical Methodology. *Report on Exact and Statistical Matching Techniques*. Statistical Policy Working Paper 5, U.S. Department of Commerce ,Washington, DC, 1980.

I.P. Fellegi and A.B. Sunter. A Theory for Record Linkage. *JASA*. 64:1183-1210, 1969.

W.B. Frakes and R. Baeza-Yates, editors. *Information Retrieval: Data Structures and Algorithms*. Prentice-Hall, Upper Saddle River, NJ, 1992.

K.A. Froeschl. A Sketch of Statistical Meta-Computing as a Data Integration Framework. *Austrian Journal of Statistics, Special Issue on Data Integration and Record Matching*, 33: ???-???, 2004.

K.A. Froeschl and W. Grossmann. The Role of Metadata in Using Administrative Sources. *Research in Official Statistics*. 3(1):65-82, 2000.

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

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

A. Haslinger. Data Matching for the Maintenance of the Austrian Business Register. *Austrian Journal of Statistics, Special Issue on Data Integration and Record Matching*, 33: ???-???, 2004.

T.H. Hassard. Writing the Book of Life: Medical Record Linkage. In Brook, et al., editors, *The Fascination of Statistics*, pages 25-46. Dekker, New York, 1986.

D.S. Hirschberg. A Linear Space Algorithm for Computing Maximal Common Subsequences. *Communication of the ACM*. 18:341-343, 1975

IDARESA. *The Data Model – Final Version*, Deliverable 3.4.2, Dept. of Statistics, University of Vienna, 1997.

IDARESA. *IDARESA Tandem Structures*, TPR–*viu*–3.4.2/3, Dept. of Statistics, University of Vienna, 1998.

T.B. Jabine and F.J. Scheuren. Record Linkages for Statistical Purposes: Methodological Issues. *Journal of Official Statistics*. 2(3):255-277, 1986.

J.B. Kadane. Some Statistical Problems in Merging Data Files. In *1978 Compendium of Tax Research*, pages 159–171. US Dept. of the Treasury, 1978. (Reprinted in *Journal of Official Statistics*. 17(3):423-433, 2001.)

B. Kilss and W. Alvey, editors. *Record Linkage Techniques*. FCSM, Washington, DC, 1985.

J.G. Kovar and P.J. Whitridge. Imputation of Business Survey Data. In B. Cox et al., editors, *Business Survey Methods*. John Wiley, New York, 1995.

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

F. Linder. The Dutch Virtual Census 2001: A New Approach by Combining Administrative Registers and Household Sample Surveys. *Austrian Journal of Statistics, Special Issue on Data Integration and Record Matching*, 33: ???-???, 2004.

R.J.A. Little and D.B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, New York, 1987.

F.M. Malvestuto. A Universal Table Model for Categorical Databases. *Information Sciences*. 49:203-223, 1989.

F.M. Malvestuto. Data Integration in Statistical Databases. In Z. Michalewicz, editor, *Statistical and Scientific Databases*, pages 201-232. Ellis Horwood, Chichester, 1991.

F.M. Malvestuto. A Universal-Scheme Approach to Statistical Databases Containing Homogeneous Summary Tables. *ACM Transactions on Database Systems*. 18:678-708, 1993.

X.L. Meng and D.B. Rubin. Maximum Likelihood Estimation via the ECM Algorithm: A General Framework. *Biometrika*. 80(2):267-278, 1993.

C. Moriarity and F. Scheuren. Statistical Matching: A Paradigm for Assessing the Uncertainty in the Procedure. *Journal of Official Statistics*. 17(3):407-422, 2001.

R. Müllauer. *TST-2 – Benutzerhandbuch (in German)*. Internal Report, Statistics Austria, Vienna, 2003.

M. Neiling. Data Fusion with Record Linkage. Presented at the *3rd Workshop* "*Föderierte Datenbanken*", Magdeburg, 1998. Also available at *http://wwwiti.cs.uni-magdeburg.de/fdb98/online-proc/.*

M. Neiling and H.J. Lenz. The Creation of the Register Based Census for Germany in 2001: An Application of Data Integration. In *Betriebswirtschaftliche Reihe: Diskussionsbeiträge des Fachbereichs Wirtschaftswissenschaft der FU Berlin* 34. Freie Universität Berlin, 1999.

S. Nikles and R. Müllauer. *TST-2 – Textsuchtetool Version 2: Beschreibung der Methode zum Abgleich von Dateien (in German)*. Internal Report, Statistics Austria, Vienna, 2003.

B.A. Okner. Constructing a New Data Base from Existing Microdata Sets: the 1966 Merge File. *Annals of Economic and Social Measurement*. 1:325-342, 1972.

B.A. Okner. Data Matching and Merging: An Overview. *Annals of Economic and Social Measurement*. 3(2):347-352, 1974.

E. Porter and W. Winkler. *Approximate String Comparison and its Effect on an Advanced Record Linkage System*, RR97-02, U.S. Bureau of the Census, 1997. Available at *http://www.census.gov/srd/www/byyear.html*.

D.B. Radner. The Development of Statistical Matching in Economics. In *Proc. Social Statistics Section*, pages 503-508. American Statistical Association, 1978.

D.B. Radner and H.J. Muller. Alternative Types of Record Matching: Costs and Benefits. In *Proc. Social Statistics Section*, pages 756-761. American Statistical Association, 1977.

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

P. Roberti. The DIECOFIS Project: Progress and Lessons. *Austrian Journal of Statistics, Special Issue on Data Integration and Record Matching*, 33, ???-???, 2004.

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

W.L. Rodgers and E.B. DeVol. An Evaluation of Statistical Matching. In *Proc. of the Survey Research Methods Section*, pages 128-132. American Statistical Association, 1981.

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

R. Schaumann. Die Registerverordnung der EU und deren Umsetzung in Österreich (in German). In: N. Rainer, editor, *Österreichs Statistik in der Europäischen Integration*, ÖstASt Nr. 2, ÖSTAT, Wien, pages 91-99, 1999.

R. Schnell, T. Bachteler, and S. Bender. A toolbox for record linkage. *Austrian Journal of Statistics, Special Issue on Data Integration and Record Matching*, 33: ???-???, 2004.

SecondString. SecondString Project Page http://secondstring.sourceforge.net/, (2004).

Y.R. Wang and S.E. Madnick. The Inter-Database Instance Identification Problem in Integrating Autonomous Systems. In *Proc. of the 6th International Conference on Data Engineering*, Los Angeles, pages 46-55. IEEE, 1989.

W. Winkler. Preprocessing of Lists and String Comparison. In B. Kilss, W. Alvey, editors, *Record Linkage Techniques*, pages 181-187. FCSM, Washington, DC, 1985.

W. Winkler. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. In *Proc. Section on Survey Research Methods*, pages 354-359. American Statistical Association, 1990.

W. Winkler. Matching and Record Linkage. In B. Cox et al., editors, *Business Survey Methods*, pages 355-384. J. Wiley, New York, 1995.

W. Winkler. *The State of Record Linkage and Current Research Problems*, RR99-04, U.S. Bureau of the Census, 1999. See *http://www.census.gov/srd/www/byyear.html*.

W. Winkler. *Quality of Very Large Databases*, RR2001/04, U.S. Bureau of the Census, 2001.

C.F.J. Wu. On the Convergence Properties of the EM-Algorithm. *Annals of Statistics*. 11(1):95-103, 1983.

Authors' addresses:

Dr. Michaela Denk
ec3 – E-Commerce Competence Center
Donau-City-Straße 1
A-1220 Vienna
Austria

Tel. +43 1 522 71 71 / 19
Fax +43 1 522 71 71 / 71
Elec. Mail: michaela.denk@ec3.at
http://www.ec3.at/

Univ.-Prof. Dr. Peter Hackl
Department of Statistics, WU Wien
Augasse 2-6
A-1090 Vienna
Austria

Tel. +43 1 31336 / 4751
Fax +43 1 31336 / 711
Elec. Mail: peter.hackl@wu-wien.ac.at
http://eeyore.wu-wien.ac.at/stat4/hackl/home.html