Michaela Denk and Peter Hackl and Norbert Rainer

String Matching Techniques: An Empirical Assessment Based on Statistics Austria's Business Register

AUSTRIAN JOURNAL OF STATISTICS
Volume 34 (2005), Number 3, 235–249

String Matching Techniques: An Empirical Assessment Based on Statistics Austria's Business Register

Michaela Denk[1], Peter Hackl[2], and Norbert Rainer[3]

[1]ec3 – Electronic Commerce Competence Center, Vienna
[2]University of Economics and Business Administration, Vienna
[3]Statistics Austria, Vienna

Abstract: The maintenance and updating of Statistics Austria's business register requires a regularly matching of the register against other data sources; one of them is the register of tax units of the Austrian Federal Ministry of Finance. The matching process is based on string comparison via bigrams of enterprise names and addresses, and a quality class approach assigning pairs of register units into classes of different compliance (i.e., matching quality) based on bigram similarity values and the comparison of other matching variables, like the NACE code or the year of foundation.

Based on methodological research concerning matching techniques carried out in the DIECOFIS project, an empirical comparison of the bigram method and other string matching techniques was conducted: the edit distance, the Jaro algorithm and the Jaro-Winkler algorithm, the longest common subsequence and the maximal match were selected as appropriate alternatives and evaluated in the study.

This paper briefly introduces Statistics Austria's business register and the corresponding maintenance process and reports on the results of the empirical study.

Zusammenfassung: Die Pflege und Aktualisierung des Unternehmensregisters der Statistik Austria erfordert den Abgleich des Registers mit anderen Datenquellen, u.a. auch mit dem Register der Steuersubjekte des Bundesministeriums für Finanzen. Der Matching-Prozess beruht einerseits auf Stringvergleichen von Firmennamen und adressen mittels Bigrammen, und andererseits auf einem Qualitätsklassen-Ansatz, der Paare von Registereinheiten aufgrund ihrer Bigramm-Werte und der Vergleiche anderer Matching-Variablen, wie z.B. NACE-Code oder Gründungsjahr, in Klassen unterschiedlicher Übereinstimmung (d.h. Qualität) einteilt.

Basierend auf Ergebnissen des DIECOFIS-Projektes im Bereich Methoden für Datenintegration bzw. Datenabgleich wurde eine empirische Studie durchgeführt, um die bei Statistik Austria eingesetzte Bigramm-Methode anderen ausgewählten Stringvergleichsverfahren gegenüberzustellen: die Edit-Distanz, die Jaro- und Jaro-Winkler-Algorithmen, und die Verfahren Longest Common Subsequence und Maximal Match wurden als mögliche Alternativen in die Studie miteinbezogen.

Dieser Artikel stellt kurz das Unternehmensregister der Statistik Austria und den entsprechenden Abgleichsprozess vor und berichtet über die Ergebnisse der genannten empirischen Studie.

Keywords: Data matching, Record Matching, String Matching, Business Register, Register Maintenance, DIECOFIS.

1 Introduction

DIECOFIS (Development of a System of Indicators on Competitiveness and Fiscal Impact on Enterprises Performance, cf. DIECOFIS, 2003, and Roberti, 2004) was an EU-funded international research project, coordinated by the Italian national statistical agency ISTAT. The main goal of the project was to foster the development of "best" policy impact and evaluation techniques in the field of taxation. For that purpose, *micro-simulation models* for enterprise taxation have been developed. For the creation of a multi-source database of enterprise data as a basis of micro-simulations, *data integration*, mainly record matching, was a core issue of the project. Data integration has been the field in which the Austrian member of the consortium was engaged. Within 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 – Bundesanstalt Statistik Austria, and the Statistical Department of the Austrian Economic Chamber, were represented (Denk and Hackl, 2003).

A survey of available methods of data integration (Denk and Oropallo, 2002) has been provided as well as a discussion of relative merits of the various methods in the context of databases to be encountered in the national statistics context (Denk, Inglese, and Calza, 2003) and on the assessment of multi-source databases and related quality indicators (Denk, Inglese, and Oropallo, 2003). In addition, an empirical study has been performed (Weghofer, 2004) comparing the application of various string matching techniques in the context of ST.AT's business register and demonstrating the use of quality indicators for the assessment of the various techniques. An international workshop on data integration and record matching was organized in Vienna in November 2003 which was hosted by Statistics Austria (cf. Denk et al., 2004).

The paper gives a report on this empirical study. Section 2 briefly introduces Statistics Austria's business register and describes the updating process that is regularly performed. Section 3 gives some details of the empirical study and summarizes the main results. The final section 4 discusses some conclusions and sketches directions of further work in this field.

Following a referee's advice we would like to draw the reader's attention to the fact that the matching of datasets may cause problems with data protection and might even raise the question of legality of the whole exercise.

2 Statistics Austria's Business Register

The business register (BR) of ST.AT serves as a basic instrument for conducting surveys addressed to enterprises and similar economic units. It is therefore used for purposes of economic statistics, but also for certain surveys in social statistics. Because of the importance of the BR as a basis for high quality and internationally harmonised statistics, already in 1993 the European Council issued a Regulation on business registers for statistical purposes (No 2186/93) setting out the requirements that such registers have to fulfil.

In the past, the main use of the BR was supporting the survey processes by providing information on name, address and similar data of the enterprises that were selected for surveys, by building the frame for drawing samples and by helping in monitoring the survey process.

While these purposes are still important, further tasks have emerged. The BR is itself seen as a database for producing statistics on the size, structure and development of the register population. Secondly, the BR is used to replace or supplement economic and social surveys. If the requested information is already available in administrative registers and databases, there is no need to ask these data again from the enterprises which have delivered this information to government administrations, such as the tax and social security authorities. This aspect of the use of the BR has become quite important in the last years. The BR needs not only to provide high quality data as regards coverage, timeliness and accuracy but also to be linked to the various administrative data sources.

The BR covers all non-agricultural enterprises, their establishments and their local units, as well as the non-profit and the government institutions. For the regular maintenance and updating of the enterprise units four main administrative registers and databases are used:

- Register of the Austrian Federal Economic Chamber: This register provides information on the type of economic activity and on the existence of local units. However, not all economic activities are covered by this source as some activities are outside the competence of the Economic Chamber.
- Register of companies: This is a public register covering corporations. This database mainly provides the official name of the firm, the legal form and variables of enterprise demography. However, the register of companies covers only about 150.000 firms, which is about 40% of all enterprises, as most of the enterprises are sole proprietorship businesses which are not required to register in the company register.
- Register of Employers of the Social Security Institutions: This register covers all enterprises, non-profit and government institutions with employees. Enterprises with only self-employed persons are not included in this database.
- Tax Register: The tax register is the most comprehensive database, it even goes far beyond the scope of the BR; it includes all tax payers, not only enterprises but also individuals.

Each of the four main data sources provides specific information; none of these sources alone would be a sufficient database fulfilling the requirements of the BR. For example, the social security register provides data on the number of employees of each enterprise, the tax register the amount of turnover of each enterprise. Integrating these sources into the BR would not only help updating the BR information, but can also replace surveys on businesses by using the data given in the administrative sources via the link to the BR. Such an approach is already applied in the business cycle statistics in the service industries and has reduced costs and respondents' burden.

However, the main challenge in integrating the administrative sources is that there exists no common identifier over the various administrative registers that would therefore allow establishing a link very easily. Furthermore, the definitions of the units differ to some degree between the databases and the variables are not standardized so that even common variables in all these data sources may vary to some degree.

Therefore, in order to link the administrative data sources with the BR, string comparison algorithms are applied, using variables such as name and address of the enterprise. With the help of such algorithms, decisions are supported as to whether a specific unit that is found in an administrative database is already covered in the BR (match) or not (unmatch). The actions to be taken in the BR in cases of matches and unmatches depend on the kind of the administrative source. For example, a change in the legal form of a company results in certain cases into a new data record in the company register with a new identification code but has no consequence for the unit in the BR. A new data record in the social security register can indicate that either an existing enterprise has started to employ persons or the creation of a new enterprise with employees from the start on.

As there are about 400.000 enterprises in Austria, the basic matching procedures needs to be automated as much as possible. The matching algorithm should find matches and unmatches with as little error rates as possible. The matching procedure used in ST.AT consists of three main steps: analysis of the database that is to be matched with the BR, application of the bigram method, and categorising the results into classes of compliance based on the bigram similarity values.

The first step is a thorough analysis of the data file (administrative records) that should be matched against the BR. This step results into a standardised input file by applying various editing processes (such as deleting blanks, setting all characters in uppercase, convert special characters, convert typical abbreviations, split text strings into words, etc.) Theses pre-matching procedures also give hints for possibilities of blocking the mass of data records into distinctive groups as well as (first) experience on the general quality and peculiarities of the input data. It is the experience of ST.AT that this step of standardisation is of high relevance for efficient record matching.

The second step concerns the application of the matching method itself. ST.AT uses a matching method which is based on string comparisons via bigrams of name and address of the enterprise. Further variables are taken into account if available and of good quality. As usually a huge amount of data records are to be matched, blocking is applied if possible. Typically, the best blocking method is by regions (according to ZIP codes or the first two digits of the ZIP codes). However, remaining (unmatched) data records of one block will be matched with the remaining data records of the other blocks in a stepwise procedure. The main advantages of the bigram method are that it is simple and transparent, and it is robust against permutations of the words as well as against typing and phrasing errors.

The third step refers to the grouping of the matching results into classes of different compliance (i.e. matching quality). This is done by using the degree of similarity found by the bigram method, by considering compliance of further variables, if possible (for instance the NACE code), and by applying a weighting procedure. The categories of quality (in practice between three and five) are thus not only based on the degree of similarity of the string comparison procedure, but also on an assessment of other relevant variables and an evaluation of the quality of the input data by assigning weights to each of the components. The quality classes are thus defined for each matching procedure separately. Only matches of the highest quality class would automatically be considered in the BR, either by creating a new data record, by creating an additional link, or by taking other measures depending on the type of administrative data. Record pairs of the second quality level need to be checked by the BR staff. Matches of lower quality categories are not taken into account at all.

Matching the four above cited administrative data sources with the BR is a permanent task for the maintenance of the BR. Each of the four data sources has its strengths and weaknesses which needs to be considered in the matching process. Fortunately, in the past years the links between the administrative data sources have increased. For instance, the tax register now also provides the ID number of the company register, however, only for a subset of all enterprises. Such information is also utilised to check the links already created in the BR. As the links to the administrative data sources are of fundamental importance for the BR, every possibility needs to be considered to make sure that the links are correct and up to date.

Matching procedures are not only applied for linking administrative data sources to the BR. The same procedures are used to detect duplicates in the BR. Furthermore, in cases where certain variables of the BR such as the NACE code are to be linked to external data files, matching is a necessary step if no common identifier is available. For further details concerning the matching methods, see Haslinger (2004).

3 Empirical Analysis of Matching Algorithms

Major aims of the empirical study were the comparison of selected string matching techniques, in particular in contrast to the bigram method used by ST.AT, and the derivation of recommendations concerning Statistics Austria's matching process in this respect. Based on the methodological research carried out in the DIECOFIS project (cf. DIECOFIS, 2003; Denk and Oropallo, 2002; Denk, Inglese, and Calza, 2003; Denk, Inglese, and Oropallo, 2003; Denk and Hackl, 2003), appropriate alternatives were selected and implemented. Apart from the "pure" *bigram method* (i.e. without further pre-processing as in the ST.AT matching process; cf. Ukkonen, 1985; Frakes and Baeza-Yates, 1992), the classical *edit distance* (cf. Damerau, 1964; Levenstein, 1966; Hall and Dowling, 1980), two quite up-to-date methods developed and well proven by the U.S. Bureau of the Census (cf., for instance, Winkler, 1985, 1990, 1999; and Porter and Winkler, 1997), viz. the *Jaro algorithm* and an enhancement of the Jaro algorithm by Bill Winkler (henceforth termed *Jaro-Winkler algorithm*), and two methods widely used in information technology but not so much in statistical applications, the *longest common subsequence* (in the following abbreviated by LCS; cf. Hirschberg, 1977; Apostolico and Guerra, 1987; Gusfield, 1997) and the *maximal match* (Ehrenfeucht and Haussler, 1988), were evaluated. All string comparison measures were transformed to *similarity measures* and, if necessary, rescaled to the closed interval [0, 1]; perfect match corresponds to the value 1. For a description of all used algorithms see Weghofer (2004) and Denk (2002). Further string comparators are presented in Gill (2001) and Cohen et al. (2003).

ST.AT provided data from the tax register and from the business register for the province of Upper Austria, covering in total about 100.000 tax units and 70.000 business units. Creating the cross product of tax units and business units yields 7 billion record pairs. After blocking with ZIP code, i.e., restricting to record pairs with agreeing ZIP code, leaves about 400 million pairs. From these pairs, 51.906 are considered as matches, thereby following the final assignments of ST.AT's Business Register.

The comparison of matching techniques was separately conducted for matches and for unmatches; to reduce computing time a simple random sample of 58.020 unmatches was drawn from the 400 million unmatches, and all matches were considered. Finally, for the analysis of name matching only those pairs were included in the analysis that did not exactly agree on name (12.503 matches, 58.019 unmatches), since string matching is not required for identical names and, thus, the matching of identical names does not contribute to the quality of a matching technique. The same applies for the analysis of address matching, for which 13.137 matches and 57.984 unmatches that did not exactly agree on address were used. Of course, these record pairs have to be taken into consideration when evaluating the whole matching process, but they are of no relevance when assessing the performance of a string matching procedure. Figure 1 visualizes the data selection process.

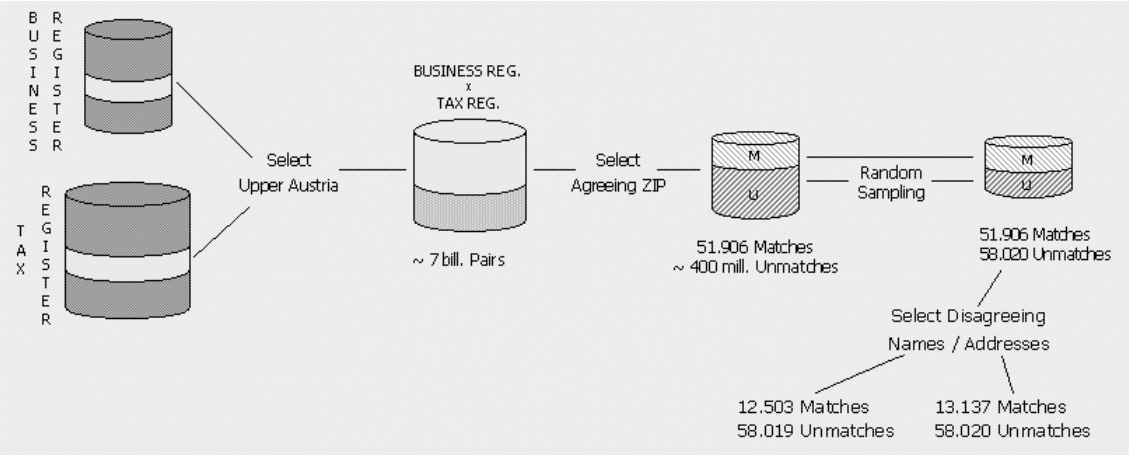

Figure 1: Data used in the analysis

To enable the comparison of implemented algorithms, ST.AT supplied not only the input data, but also intermediary datasets resulting from individual integration steps as well as the final assignments as used to update the business register; the latter dataset has been used as the above-mentioned set of matches and emanated from a comprehensive matching process with string comparison as one of several steps.

To give a first impression of the behaviour of the six string comparators, the distribution of the similarity measures is analysed separately for the classes of true matches and of true unmatches. The histograms in Figure 2 show the distributions of the similarity values for the matching variable name, Figure 3 contains the histograms for address. (Name and address were separately analysed.) Of special interest is of course the overlapping range of these distributions for unmatches and matches: The smaller the overlap the easier is the classification of a pair, i.e., the smaller are the false match and the false non-match rates.

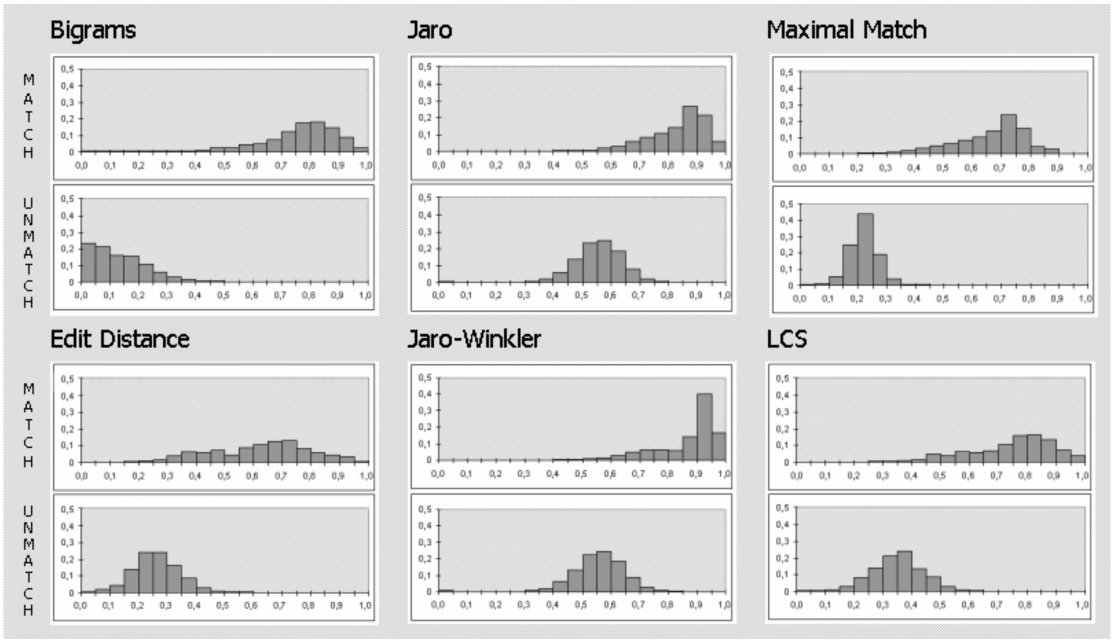

Figure 2: Distribution of string comparator values for matching variable *name*

For the matching variable *name*, the distributions of most comparators show less variability for unmatches than for matches. Apart from bigrams, the distributions of the similarity values obtained for unmatches are symmetric, whereas those for matches are left-skewed. In general, bigrams generate lower similarity values for unmatches (with a mean of 0.13) than other comparators. Maximal match also shows low values (mean 0.22) for unmatches, but with less variation. These two string comparators achieve the best separation of matches and unmatches which is tantamount to a high discriminatory power. In contrast, Jaro and Jaro-Winkler algorithms give rather high similarity values to unmatches, with the center of the distribution between 0.5 and 0.6, leading to a substantial overlap of the distribution in matches and unmatches. Though the edit distance also generates rather low values in unmatches (mean 0.27), its broad value range in matches starting from 0.2 up to 1 (with an average of 0.62) leads to a bad separation of matches and unmatches.

In addition to the separation of match and unmatch distributions and, thus, the discriminatory power of the comparators, the histograms also show that for a fixed error rate (false match or false non-match rate), the optimal similarity threshold (if only one threshold is used) will definitely be lower for bigrams and maximal match than for the edit distance and Jaro algorithms.

Figure 3 indicates that for the matching variable address the situation is much worse. Actually, none of the comparators yields a reasonable separation of matches and unmatches; all corresponding match and unmatch distributions show high variation and nearly completely overlap. Similar results are reported by, e.g., Winkler (1995) and Gill (2001), who found that addresses have less distinguishing power than names. However, this result also suggests that for variables with little discriminating power, differences between comparators diminish and, therefore, the choice of the algorithm is not that crucial.

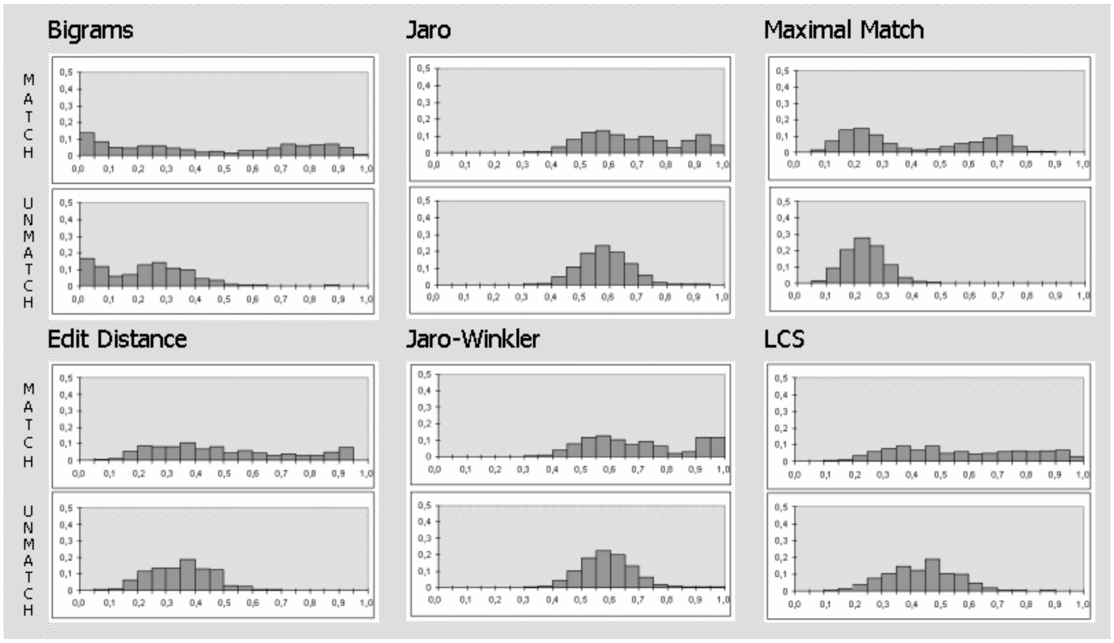

Figure 3: Distribution of string comparator values for matching variable *address*

Figure 4 displays the *correct match rate* (or *sensitivity*) versus the *false match rate* for the matching variable *name*, where the correct match rate is equal to *1 – false non-match rate* (Fellegi and Sunter, 1969; Jamieson et al., 1995). The correct match rate amounts to the proportion of detected matches with respect to all true matches (the false non-match rate is the proportion of matches that are not detected by the matching algorithm with respect to all matches). In contrast, the false match rate is defined as the proportion of unmatches that are erroneously denoted as matches by the matching method with respect to all unmatches. These measures were computed for all implemented matching algorithms for both matching variables for all feasible similarity thresholds. Note that Figure 4 shows the correct match rate only for the range from 0.7 to 1, and the false match rate from 0 to 0.3.

What could already be expected from the distributions of the string comparators is confirmed by the different error rates calculated: maximal match and bigrams are similar and dominate the other four comparators. Maximal match is slightly better for false match rates up to 8%. The Jaro and Jaro-Winkler algorithms are nearly the same, LCS is a little bit better (about 2.5 percentage points) than these two, in particular for a false match rate greater than 5%. The edit distance shows the worst performance: it detects less matches than all other comparators for a false match rate less than 20% and reaches approximately the same correct match rate as the Jaro and Jaro-Winkler algorithms for a false match rate greater than 20%. However, it must be kept in mind that the benchmark for all these comparisons is based on bigram matching, so the evaluation is biased in favour of the bigram method. Still, the assignments in the business register do not solely originate from bigram matching; they result from a comprehensive matching process, where bigram matching is only one step.

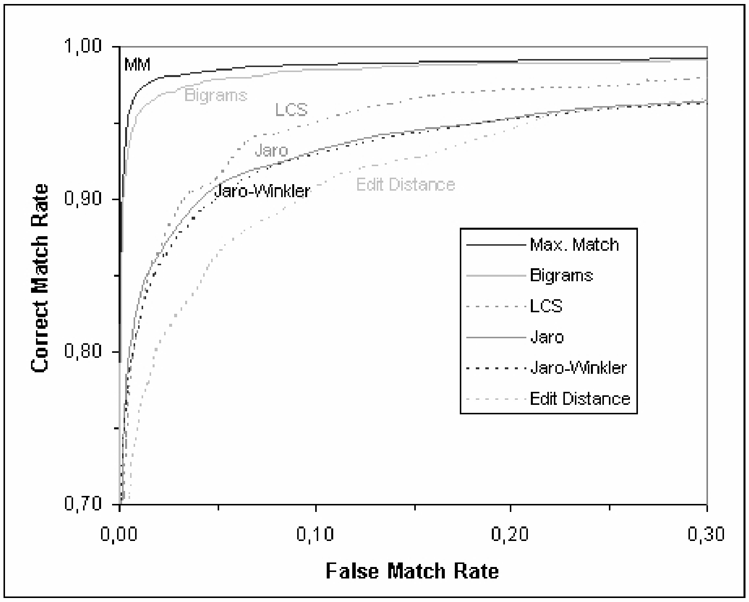

Figure 4: Correct match rate vs. false match rate of string comparators for *name*

A reason for the rather bad performance of the edit distance, the Jaro algorithms and the LCS technique which heavily rely on the sequence of different parts of strings (like first name and surname) might be that the data were standardised but not parsed. I.e., different string parts are not necessarily in the same order for all strings. This interpretation would stress the importance of these pre-processing steps for reasonable string matching results (cf. Winkler, 1995; Gill, 2001). However, according to this interpretation, one would rather expect lower comparator values for matches because of the – potentially – wrong order of string parts instead of higher values for unmatches as were observed for names (see Figure 2 above). For addresses, the interpretation seems to suit better (cf. Figure 3). Anyway, this interpretation militates in favour of the bigram and maximal match methods, since they yield better results even though the input data were not parsed.

The ordering of string comparators according to their error curves is the same for matching variable address; however, false match rates less than 20% are only achieved for false non-match rates greater than 50% which again shows the bad discriminatory power of address information.

A comparison of the sensitivity of bigrams and of the other five string comparators for matching variable name is given in Figure 5. The relative differences of the number of correct matches obtained by the five string comparators (edit distance, longest common subsequence, Jaro-Winkler and Jaro algorithm) and the bigram algorithm for selected false match rates (0.005, 0.01, 0.02) is displayed, that is the percentage of matches detected by bigrams but not by the ED, LCS, JW and J techniques, respectively, and the percentage of matches detected by maximal match but not by bigrams. In addition, the last group of bars shows the percentage of matches found by the entire ST.AT matching procedure (and thus fed into the BR) that is not detected by bigram name matching. I.e., 4-7% of the matches are gained by using additional information (bigram values of address, the NACE code, etc.) and clerical review.

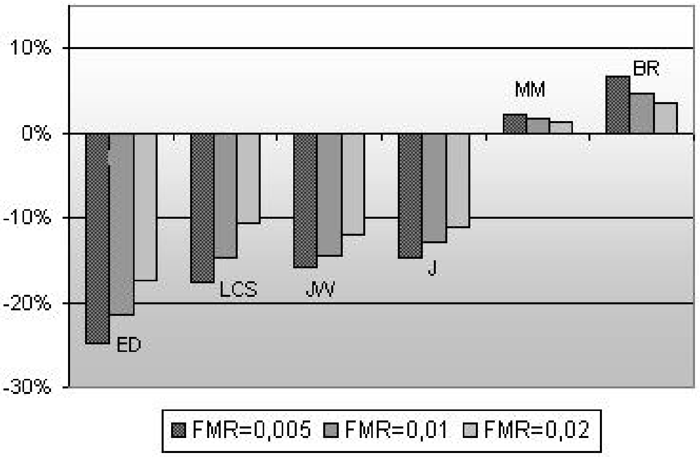

Figure 5: Advantage/disadvantage of matching methods in terms of correct matches relative to bigram name matching for selected false match rates (FMR)

Maximal match is the only algorithm that finds more matches than the bigrams; for example, the advantage is 140 matches for a false match rate of 2%. The larger the tolerated false match rate, the smaller is the gain or loss in terms of correct matches through the usage of other methods in comparison to bigrams. More absolute figures corresponding to the percentages in Figure 5 are provided in Table 1. For instance, for a tolerated false match rate of 0.5%, the Jaro and Jaro-Winkler algorithms detect approximately 1.800 matches less than the bigrams, while the edit distance even "loses" about 2.900 matches. The LCS method is slightly better than the Jaro and Jaro-Winkler algorithms for a false match rate of 0.5% and 1%. For FMR=2%, the order changes; i.e., LCS misses less matches than Jaro and Jaro-Winkler. Anyhow, in this area of very small false match rates, the performance of LCS, Jaro and Jaro-Winkler is quite similar, as shown in Figure 4 on the previous page. In general, Table 1 shows that the gain of using maximal match instead of bigrams is much smaller than the loss of using one of the other four methods instead of bigrams. Nevertheless, even a small reduction of the clerical effort may be valuable.

Table 1: Advantage/disadvantage of matching methods in terms of correct matches relative to bigram name matching for selected false match rates (FMR)

FMR	ED	LCS	JW	J	MM	BR
0.005	-2.906	-2.058	-1.853	-1.729	**+259**	**+793**
0.010	-2.509	-1.713	-1.695	-1.509	**+189**	**+545**
0.020	-2.032	-1.246	-1.403	-1.287	**+140**	**+404**

4 Conclusion and Further Work

The most obvious message of the results that were obtained in our empirical study is that maximal match and bigrams rather clearly dominate the other matching techniques: Given a certain false match rate, the dominating techniques show a similar and much higher correct match rate than the other investigated methods. However, this result must be seen with prudence.

- The final classifications that were used as benchmarks for the evaluation of the methods in our study are based on bigrams as applied by Statistics Austria; this fact might result in error rates that are biased in favour of the bigrams. The assessment of a pair in our study comes to the same decision as that taken by Statistics Austria, whether the classification is correct or erroneous, if bigrams are applied; this is not so if other techniques are investigated. Hence, a certain proportion of the correct match rate of bigrams is probably caused by this feature of our study. This effect could even be tremendously misleading if matching based on bigrams performed badly: Techniques that find the truly matching pairs with higher probability than bigrams could show in such a comparison a worse result than bigrams. However, one must bear in mind that the benchmarks do not directly stem from simple bigram matching, but rather from a comprehensive matching procedure where bigram matching is just one single matching step, as outlined in section 2.
- The reported results are conditional on the circumstances of the study. E.g., the different matching rates for *names* and *addresses* of the business units as reported in section 3 are an indicator for such a dependence upon circumstances. In general, it must be expected that the matching quality is highly dependent on certain features of the input data.

Jaro, Jaro-Winkler, and Edit Distance are affected by the mode of parsing, the preparation step applied to the data before matching is started. More generally, the experience of the individual who is responsible for the whole matching process (choice of matching variable, blocking, standardisation, parsing, choice of matching technique, etc.) determines to a great extent the quality of the results. Of course, the experience of individuals is not subject of a study like ours but has to be taken into account for understanding and interpreting the results.

- The comparison of the results that are obtained on the basis of the matching variable *name* with those based on the matching variable *address* suggests that the differences between the matching techniques also depend on the discriminating potential of the matching variable. The differences in error rates become smaller for a less discriminating matching variable.

As a consequence of the limited amount of prior experience and effort the results of our study lead us to further questions and problems that are waiting for closer investigations. Some of them are related to the matching techniques, some to their application in dataset integration.

- The most obvious question concerns the true relative performance of the various techniques, i.e., the matching rates unconditional of the Statistics Austria or any other benchmark.

- A systematic investigation of the effects of various features of the input data might deepen the experience and suggest new strategies for the matching process. These investigations might in particular take into consideration string characteristics, such as the character set the strings are based on or any characteristics of the composition and complexity of the strings, and aspects of data quality, like the amount of missing data in the matching variables, and experiment with various ways of treating missing values in the matching process. In this context we also see the comment of a referee who mentions that according to his or her experience, it might be worth to include character transposition in the bigram pairs to cope with some of the keyboarding errors that occur during data preparation.
- Other matching approaches might be considered; e.g., probabilistic record linkage that makes use of the result of string comparisons as input for the estimation of conditional probabilities of a record pair being a match or unmatch might improve the chances of finding matching pairs. An obvious alternative to bigrams are trigrams or other "n-grams"; the referee, however, suggested this option, also reports poor performance of trigrams in a related study.
- The additional use of a second matching technique besides bigrams typically results in pairs of units that are considered as matching by the second method but not by the bigrams (or vice versa). The analysis of such differently classified pairs might be a fruitful source for learning more about the matching potential of the bigrams and the second technique.

For the application of matching techniques a closer look at the following issues might prove helpful:

- A quality class approach like that used at Statistics Austria in order to separate sets of pairs for customized further processing could be refined, e.g., by using additional similarity measures.
- Alternative blocking strategies could be analysed with regard to "lost" matches and the potential of reducing the number of pairs that go into the actual matching procedure. For instance, Baxter et al. (2003) and Gu and Baxter (2004) propose alternatives to "standard blocking", like the sorted neighbourhood method, fuzzy blocking or canopy blocking, and performance measures for blocking methods, such as the reduction ratio, the pairs completeness and the trade-off between these two measures.
- The task of updating the business register by matching its records with those of the tax register results (i) in a set of matching units present in both registers for which a link is established in the BR, (ii) in a set of inactive business units for which no matching record can be found in the tax register and (iii) in a set of new units which are only present in the tax register. These latter business units are candidates for being appended to the business register as a result of the updating process. Up to now, the empirical analysis focussed primarily on finding matching units; however, the quality of different matching procedures should also be analysed concerning the ability of detecting such candidate units, i.e., finding true unmatches.

These points should make clear that, although Statistics Austria is very successful in routinely updating the business register by means of bigrams, clarifying some of these questions might lead to a sharper targeted and even more efficient updating process. Apart from potential methodological improvements, especially, a higher degree of automisation of the matching process and advances in the quality of administrative data sources also including links established between different data sources would contribute to an efficiency increase.

Acknowledgements

The authors are grateful to the editor for the easy and constructive cooperation and the referee for his or her highly competent advice.

References

Apostolico, A., and Guerra, C. (1987). The longest common subsequence problem revisited. *Algorithmica*, *2*, 315-336.

Baxter, R., Christen, P., and Churches, T. (2003). A comparison of fast blocking methods for record linkage. In *Proceedings of the First Workshop on Data Cleaning, Record Linkage, and Object Consolidation, 9th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Washington, DC.*

Cohen, W. W., Ravikumar, P., and Fienberg, S. E. (2003). A comparison of string distance metrics for name-matching tasks. In *Proceedings of the 18th International Joint Conference Workshop on Information Integration on the Web.*

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, *7*, 171-176.

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

Denk, M., Fröschl, K., Hackl, P., and Rainer, N. (Eds.). (2004). Special issue on data integration and record matching. *Austrian Journal of Statistics*, *33*, 1-264.

Denk, M., and Hackl, P. (2003). Data integration and record matching: An Austrian contribution to research in official statistics. *Austrian Journal of Statistics*, *32*, 305-321.

Denk, M., Inglese, F., and Calza, M. G. (2003). *Assessment of different approaches for the integration of sample surveys.* DIECOFIS Deliverable 1.2, ISTAT, Rome.

Denk, M., Inglese, F., and Oropallo, F. (2003). *Report on statistical indicators for the assessment of multi-source databases.* DIECOFIS Deliverable 1.3, ISTAT, Rome.

Denk, M., and Oropallo, F. (2002). *Overview of the issues in multi-source databases.* DIECOFIS Deliverable 1.1, ISTAT, Rome.

DIECOFIS. (2003). *DIECOFIS web site.* http://petra1.istat.it/diecofis/index.html.

Ehrenfeucht, A., and Haussler, D. (1988). A new distance metric on strings computable in linear time. *Discrete Applied Mathematics*, *20*, 191-203.

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

Frakes, W. B., and Baeza-Yates, R. (Eds.). (1992). *Information retrieval: Data structures and algorithms.* Upper Saddle River, NJ: Prentice-Hall.

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

Gu, L., and Baxter, R. (2004). Adaptive filtering for efficient record linkage. In *Proc. siam 2004 international conference on data mining, orlando, florida.*

Gusfield, D. (1997). *Algorithms on Strings, Trees, and Sequences.* Cambridge University Press.

Hall, P. A. V., and Dowling, G. R. (1980). Approximate string matching. *ACM Computing Surveys*, *12*, 381-402.

Haslinger, A. (2004). Data matching for the maintenance of the austrian business register. *Austrian Journal of Statistics*, *33*, 55-67.

Hirschberg, D. S. (1977). Algorithms for the longest common subsequence problem. *Journal of the ACM*, *24*, 664-675.

Jamieson, E., Roberts, J., and Browne, G. (1995). The feasibility and accuracy of anonymized record linkage to estimate shared clientele among three health and social service agencies. *Methods of Information in Medicine*, *34*, 371-377.

Levenstein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Sov. Phys. Dokl.*, *10*, 707-710.

Porter, E., and Winkler, W. (1997). *Approximate string comparison and its effect on an advanced record linkage system.* RR97-02, U.S. Bureau of the Census. (Available at http://www.census.gov/srd/www/byyear.html)

Roberti, P. (2004). International research into developing integrated and systematized information systems (eisis) for eu business policy impact analysis. *Austrian Journal of Statistics*, *33*, 3-33.

Ukkonen, E. (1985). Algorithms for approximate string matching. *Information and Control*, *64*, 100-118.

Weghofer, E. (2004). *Beurteilung ausgewählter Stringvergleichsalgorithmen zur Eignung für Record Linkage an Hand einer empirischen Datenbank.* Unpublished master's thesis, Vienna University of Economics and Business Administration, Department of Statistics.

Winkler, W. (1985). Preprocessing of lists and string comparison. In B. Kilss and W. Alvey (Eds.), *Record Linkage Techniques* (p. 181-187). FCSM, Washington, DC.

Winkler, W. (1990). String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In *Proc. Section on Survey Research Methods* (p. 354-359). American Statistical Association.

Winkler, W. (1995). Matching and record linkage. In B. Cox and et al. (Eds.), *Business Survey Methods* (p. 355-384). New York: J. Wiley.

Winkler, W. (1999). *The state of record linkage and current research problems.* RR99-04, U.S. Bureau of the Census. (http://www.census.gov/srd/www/byyear.html)

Author's addresses:

Dr. Michaela Denk
ec3 – Electronic 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
E-mail: michaela.denk@ec3.at
http://www.ec3.at/

Univ.-Prof. Dr. Peter Hackl
Department of Statistics
Vienna University of Economics and Business Administration
Augasse 2-6
A-1090 Vienna
Austria

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

Dr. Norbert Rainer
Registers, classifications and methods division
Statistics Austria
Guglgasse 13
A-1110 Vienna
Austria

Tel. +43 1 71128 / 7703
Fax +43 1 71128 / 7053
E-mail: norbert.rainer@statistik.gv.at
http://www.statistik.at/