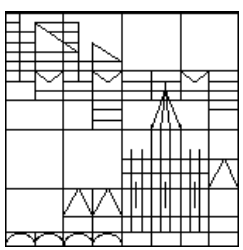
Second conference of the European Survey Research Association, 25-29 June 2007 in Prague, Czech Republic

Improving Record-Linkage-Software for Survey-Data

Rainer Schnell, Tobias Bachteler, and Jörg Reiher

Center for Quantitative Methods and Survey Research
University of Konstanz, Germany

June 25, 2007



CMS

Introduction

- Increasingly survey data is linked with individual administrative data.
- Such linkages may be used to improve data quality of surveys.
- An example are work histories, because respondents tend to underreport short spells of unemployment.

Outline

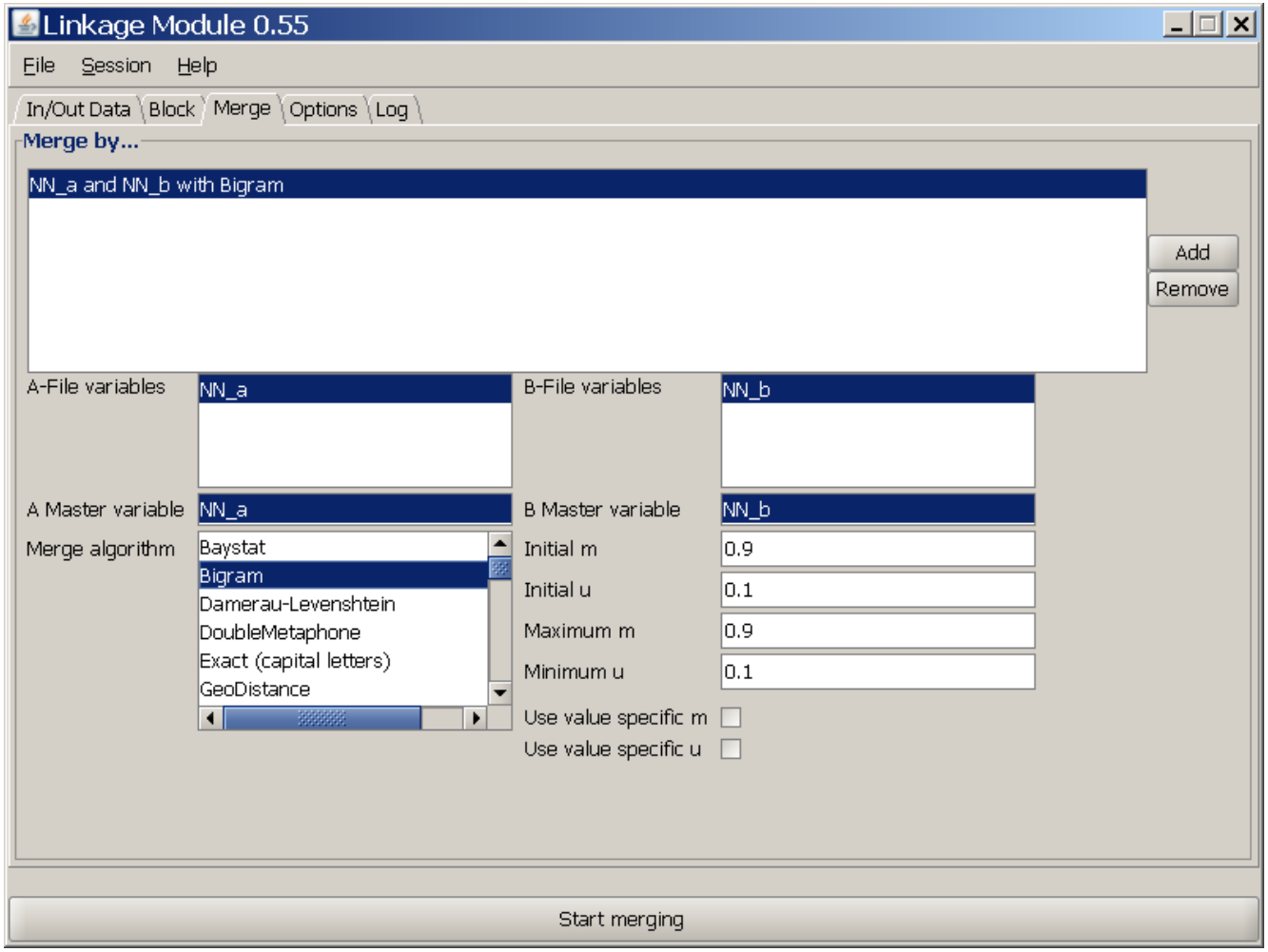
Main Problems of Applying Record Linkage on Survey Data

Two main problems have to be solved:

1. Data protection objections
2. Error prone common identifiers

- The first problem can be solved by explaining the purpose of the linkage to the respondents.
- The second problem can be solved by using special software.
  - We have implemented Merge Toolbox (MTB), a state-of-the-art record linkage software for the social sciences (free for academic use).
  - Further improvements supposedly data driven, by improving the data quality or by augmenting the existing data.
  - We intend to integrate the data augmentation process in MTB.

Merge Toolbox - MTB



Consenting Procedures in German Surveys

- Signed permission of each respondent in a survey.
- Data protection objections can be ameliorated by obtaining the informed consent of survey respondents, whether in written form or by telephone.
- Contrary to the popular belief, it is possible to get the consent to link the data by a large fraction of the respondents.
- The consent rate depends on details of implementation like the position in the questionnaire.
- To explain the purpose and importance of the linkage to the respondents is essential.

Example text of a consent form

In order to keep the interview as short as possible, we would like to use administrative data held by the Federal Employment Agency.

Such administrative data could be informations about previous periods of employment or unemployment, and participation in employment programs.

We would like to ask you to consent to the linkage of your administrative data with your interview data. Should these informations be analysed, it is absolutely guaranteed that all regulations of data privacy laws are strictly met.

Your consent is purely voluntary. You are free to revoke it at any time you like.

Adapted and translated from a questionnaire of the Institute for Applied Social Sciences (infas), Bonn.

Consenting Rates

Table 1: *Consenting rates to linkage requests in surveys conducted by infas*

Client	*Year*	*n*	*Rate*
Max Planck Institute for Human Development	1998-1999	3,000	80.6%
Federal Agency of Employment	2007	1,100	69.0%
Institute of Employment Research	1999-2004	9,000	78.0%
Institute of Employment Research	2000-2001	24,000	73.4%
Institute of Employment Research	2000-2004	4,083	79.0%
Institute of Employment Research	2005-2006	24,000	91.9%
Federal Ministry of Labour and Social Affairs	2002	6,183	69.0%
Federal Ministry of Labour and Social Affairs	2003-2006	1,500	87.7%
Federal Ministry of Labour and Social Affairs	2003-2006	10,000	97.5%
Federal Ministry of Labour and Social Affairs	2004-2006	24,000	84.9%

Source: Doris Hess (infas, Bonn), personal communication

Improving Record Linkage by Augmenting Data

- The second main problem when applying record linkage on survey data are error prone common identifiers.
- Existing algorithms can be improved by augmenting data.
- Main purpose is to obtain additional blocking variables.
- We will illustrate this by three examples.

A Bayes Classifier for Nationality

- Goal: to add information about the nationality of the respondents.
- We trained a naive Bayes classifier by calculating the probabilities of trigrams to be contained in a surname of persons with nationality n_i.
- A given surname is classified as of nationality n_i if the conditional probability of n_i given its trigram set is maximal.

Application

- In an experiment with 70,000 real names with known nationality of the persons, we tested the performance of the classificator.
- We tried to classify into 137 classes or nationalities.
- 81% of the names are correctly classified.
- Among the names of non Germans 44% are correctly classified.
- The PRE for the non German names amounts to 1/3.

Using Birth Place Information

- Sometimes the birth places of German respondents are available.
- If a birth place is not located in Germany, possibly the respondent is a naturalised person.
- Since most naturalised Germans were born in eastern european countries, we compiled a list of typical birth places there.
  - Region of birth as additional blocking variable
  - Additional usage: treating their names differently in similarity calculations
  - Further application: Sampling of special populations

Geocoding of Post Codes

- We have compiled geo-coordinates of nearly all of the about 30,000 German post codes.
- If post code is a common identifier, the geo-information can be linked.
- Based on the post codes, the distance for every record pair can be calculated.
- Usable as similarity measure or as a blocking variable.
- Underlying hypothesis: if respondents move they are more likely to move to a place nearby.
  - 3,737,000 German people moved across municipal borders in 2004.
  - 70% out of them moved to places within the same federal state, 30% moved across federal state borders.
    (Source: Federal Statistical Office, Data Report 2006)

Application

- In collaboration with the Bremen Cancer Registry we are going to match data from a mammography screening and a epidemiological cancer registry.
- Some of the women will have moved between the data collections.
- We will add geo-coordinates and use the distance as an additional matching variable.

Summary

There are two main problems of using record linkage with survey data

- Data protection objections
- Error prone identifier

For each, we presented a possible remedy:

- It is possible to get the consent to link individual data by a large fraction of survey respondents.
- Additional blocking and matching variables can be obtained by augmenting the available data.

Further Information

Literature

- Schnell, R., Bachteler, T., und Bender, S. (2003): Record Linkage Using Error Prone Strings. Proceedings of the joint statistical meeting, S. 3713-3717, American Statistical Association.
- Schnell, R., Bachteler, T. und Bender, S. (2004): A Toolbox for Record Linkage. Austrian Journal of Statistics, 33(1-2), 125-133.
- Schnell, R., Bachteler, T. und Reiher, J. (2005): MTB - Ein Record-Linkage-Programm für die empirische Sozialwissenschaft. ZA-Information, 56, 93-103.

Contact

- `recordlinkage@uni-konstanz.de`

Project home

- `http://www.uni-konstanz.de/Schnell/safelink.html`