Anonymization of longitudinal surveys in the presence of outliers

Hans-Peter Hafner
HTW Saar – Saarland University of Applied Sciences
Hans-Peter.Hafner@htwsaar.de

Rainer Lenz
HTW Saar – Saarland University of Applied Sciences
Technical University of Dortmund

Work Session on Statistical Data Confidentiality
Helsinki, 6 October 2015

Motivation

High demand for micro data of business surveys

NSI cannot respond to requests in adequate time (confidentiality requirements, staff shortage)

Synthetic data generated by CART models
►Not satisfactory for data containing outliers

Other anonymization methods are needed!

Overview

- **Linear Mixed Models and Extensions**
- **Data: German Cost Structure Survey**
- **Tested models**
- **Results analytical potential**
- **Results disclosure risk**
- **Conclusion and future work**

Linear Mixed Models and Extensions

Linear Mixed Model (LMM)

(1) $Y_i = X_i\beta + Z_i b_i + \varepsilon_i$
(2) $b_i \sim N(0, D)$
(3) $\varepsilon_i \sim N(0, \Sigma_i)$
(4) $b_1,\ldots,b_N, \varepsilon_1,\ldots, \varepsilon_N$ independent

β *fixed effects (constant across units)*
b_i *random effects (varying across units)*

Robust Linear Mixed Model (Koller 2014)
Lower weights for outliers

Generalized Linear Mixed Model (GLMM)
Exponential family instead of normal distribution

German Cost Structure Survey

Information on enterprises of the manufacturing sector

► Branch of economic activity, location of headquarter, number of employees, turnover, raw material consumption …

Sample at most 18.000 enterprises per year, 20+ employees

500+ employees: Complete survey

Panel 1999 to 2002: 13.227 enterprises

Tested models – 2 approaches

Approach 1

Same units for model building and estimation of anonymized values.

Approach 2

Two subsamples:
Estimate model with sample 1 and compute synthetic values using values of sample 2 and model coefficients from „nearest“ record of sample 1
(and the other way around)

„Nearest“ record:
Divide dataset in layers of 17 economic groups and old / new federal states.
Sort each layer by squaresum of the 4 values of the attribute to be anonymized
Unit with highest value -> Sample 1, unit with second highest -> sample 2 …
Assignment between samples: Units with same rank in layer

Tested models – 5 variations

For each approach 5 variations:

Variation 1
Common LMM for whole dataset respective whole sample.

Variation 2
Separate LMM for „normal“ data and outlier (Hampel rule).

Variation 3
Separate LMM for each of the 17 economic groups.

Variation 4
GLMM with Poisson distribution (for 17 economic groups).

Variation 5
Robust LMM (for 17 economic groups)

Analytical Potential: Criteria

Criteria

I) Deviations of single values

II) Deviations of means, standard deviations above 10% and deviations of correlations above 0,1 between waves for 17 economic groups

III) Deviations of means, standard deviations above 10% and correlations of change rates between waves above 0,1 for 17 economic groups

IV) Deviations of trends (increase / decrease) for 17 economic groups

Results: Analytical Potential 1

Approach 1

Variation 5 aborted because of memory problems

I) About 80% deviation less than 5 percent
II) 0 – 2%
III) 80 – 90%
IV) 12 – 18% single trends, 25 – 34% trends for combination of turnover and employees

I) – III) No significant differences between variations
IV) Variation 1 best results

Results: Analytical Potential 2

Approach 2

I) 12,5 – 22% deviation less than 5 percent (best results: variation 5)
II) About 13 – 50% (best results: variation 5)
III) About 50 – 85% (Variation 5 than all variations of approach 1)
IV) 18 – 26% single trends, 31 – 41% trends for combination of turnover and employees (best results: variation 5)

Variation 5 (robust LMM) has best analytical potential, but for criteria I) and II) higher deviations than for all variations of approach 1.

Disclosure Risk: Methodology

Scenario: Database Crossmatch

Anonymized dataset against original data
Minimization of distance measure between records
Blocks: 17 economic groups, old / new federal states, 6 employment size classes

Hit rate: Proportion of correct matched units in block

Useful value: Deviation from original value at most 10%

Disclosure risk in block
= (Hit rate * #useful values) / #values in block

Disclosure Risk: Results

Approach 1

For all blocks disclosure risk above 20 percent.

Number of correct matched enterprises in these blocks between 62 (variation 1) and 67 percent (variation 4) of all enterprises.

Approach 2

Disclosure risk above 20 percent for 3.5 to 4.6 percent of the blocks.

Number of correct matched enterprises in these blocks between 0.1 (variations 1 and 4) and 0.4 percent (variation 5) of all enterprises.

Conclusion

Approach 1: High analytical potential, high disclosure risk.

Approach 2: Low analytical potential, low disclosure risk

Most promising model: Approach 1, robust LMM

Possible modifications:

1) Assignment of units between the two samples.
2) Include correlation structure between waves into model.

THANK YOU FOR YOUR ATTENTION