



Data Preparation for Business Surveys

Beat Hulliger

University of Applied Sciences Northwestern Switzerland

Daniel Kilchmann

Statistical Methods Unit/Federal Statistical Office

Swiss Statistics Meeting 2007

16.11.2007





Contents

Introduction: SDP and EDIMBUS-Project

EDIMBUS-Recommended Practices Manual

Implementation at FSO

Conclusions



Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Home Affairs FDHA
Federal Statistical Office FSO



University of Applied Sciences Northwestern Switzerland
School of Business

Introduction: SDP and EDIMBUS-Project

Swiss Statistics





Statistical Data Preparation (SDP)

- ▶ SDP covers all steps from raw data in coded and electronic form to final data which is ready to be analysed.
- ▶ Statistical Data Preparation is traditionally called "Editing and Imputation" (D: Plausibilisierung, Validierung F: Plausibilisation, Validation).
- ▶ SDP is crucial for the quality of results from surveys.
- ▶ SDP is often expensive and slow (up to 40% of resources of survey, up to 1 year until publication).





EDIMBUS Project

- ▶ "Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys in the ESS".
- ▶ ISTAT (Italy), CBS (Netherlands), SFSO (Switzerland)
- ▶ Grant of 40'000€ from Eurostat for RPM.
- ▶ Estimated Total Project Cost 165'000€.
- ▶ 1.1.2006-30.6.2007.
- ▶ Target readers: Survey managers, methodologists, computer scientists (largely applicable to social surveys, too).
- ▶ State of the art survey, Draft, Referee-process, Final.



Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Home Affairs FDHA
Federal Statistical Office FSO



University of Applied Sciences Northwestern Switzerland
School of Business

EDIMBUS-Recommended Practices Manual

Swiss Statistics





Objectives of Statistical Data Preparation

1. Quality control of data.
2. Improvement of **survey** quality.
3. Data fit for use.



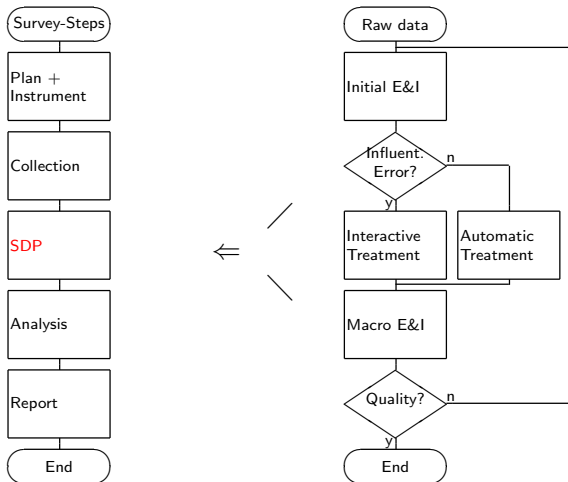
Structure of the EDIMBUS-RPM

1. General Framework of Editing and Imputation in Cross Sectional Business Surveys
2. Designing and tuning Editing and Imputation
3. Detection of Errors
4. Treatment of Errors
5. Subsequent Analysis and Estimation
6. Documenting Editing and Imputation processes
7. Synthesis and General Recommendations
8. Appendix (Notation, Methods, Indicators, Glossary)





Surveys and Statistical Data Preparation





Influential errors: selective editing

- ▶ Selective editing: Treat influential errors intensively and non-influential errors automatically (Micro-editing!).
- ▶ Selection criterium: Score to predict impact on results.
- ▶ Basic scores: $s_i = w_i |y_i - \tilde{y}_i|$ (\tilde{y}_i an anticipated value).
- ▶ Score functions mainly consider totals and ratios, but combinations and more complex versions exist (Hidioglou-Berthelot for changes).
- ▶ Choice of scales and cut-offs for scores must be tested.



Impact on Horvitz-Thompson estimator

- ▶ Impact: Sensitivity-Curve for a particular statistic at a particular observation. (Macro-editing concept!)
- ▶ Horvitz-Thompson estimator $T_H = \sum_{i \in S} w_i y_i$, with $w_i = 1/\pi_i$. Impact is

$$SC(y_i; T_{HT}, y_S, i) = n w_i (y_i - \hat{y}_i),$$

where

$$\hat{y}_i = \frac{\sum_{k \in S \setminus i} w_k y_k}{\sum_{k \in S \setminus i} w_k}$$

is the Hájek-estimator based on the rest of the observations.





Impacts, scores and outliers

- ▶ Basic score of selective editing approximates the HT-impact (\tilde{y}_i replaces \hat{y}_i).
- ▶ Only a few statistics can be reflected in a score function.
- ▶ Outliers may be simpler to detect than influential observations.
- ▶ But outliers are tied to a model, not to a statistic.
- ▶ An efficient approach will use **selective micro-editing** plus **outlier detection** and **impact measures** in macro-editing.



Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Home Affairs FDHA
Federal Statistical Office FSO



University of Applied Sciences Northwestern Switzerland
School of Business

Implementation at FSO

Swiss Statistics





Implementation at FSO

Preparation of the implementation within GUS
("Gesamtprogramm Unternehmensstatistik")

1. Dissemination
2. FSO-recommendation (1st version: 31.12.2008)
3. Standard indicators
4. Pilot

Implementation at FSO: 2009?



FSO-recommendation

Some **draft** examples (not necessarily the most important):

- ▶ Design: SDP must be a part of each project design report (part of the approval of the project design report).
- ▶ Testing: e.g. SDP has to be tested before production.
- ▶ Monitoring: e.g. use detection and treatment flags.



FSO-recommendation (continued)

Separate detection and treatment!

1. **Detection:** Flag errors, **no** change to data
2. **Decision:** Choose treatment
3. **Treatment:** Change data, flag changes.
4. **Control:** Treatment satisfactory? Else **loop back**.



FSO-recommendation (continued)

Some **draft** examples (not necessarily the most important):

- ▶ Archiving: e.g. restoration of phase archives (Initial, Micro-, Macro-E&I) should be efficient at any time of SDP.
- ▶ Documentation: e.g. steering indicators of the process have to be documented together with its respective decision.



Conclusions

- ▶ Design of the strategy from a process view.
- ▶ Testing and monitoring (better control of the process and resources) → enhancement of the survey quality.
- ▶ Documentation and archiving are crucial.
- ▶ Acquisition and exchange of further experiences.
- ▶ First step towards standardization and harmonization.

EDIMBUS web-site: <http://edibus.istat.it/EDIMBUS1>