member of
swissuniversities

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Home Affairs FDHA
Federal Statistical Office FSO

Monitoring statistical data preparation

Beat Hulliger and Daniel Kilchmann

FHNW School of Business and Swiss Federal Statistical Office

Schweizer Tage der öffentlichen Statistik, 15.9.2016, Neuchâtel

Content

1. Introduction and Statistical Data Preparation Process (SDP Process)
2. Swiss Structural Survey 2013 SDP Process
3. Flags, Structural Missingness, Indicators
4. Results for Structural Survey 2013
5. Conclusions

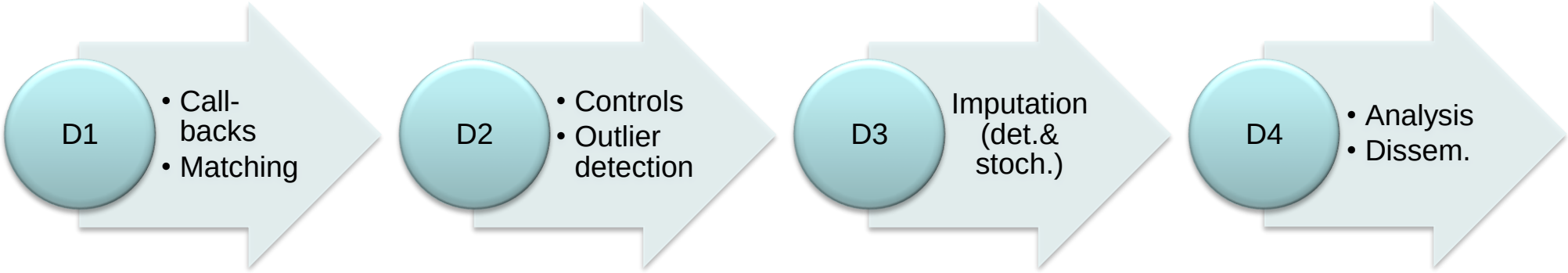
Project

- Study commissioned to Fachhochschule Nordwestschweiz (FHNW) by Swiss Federal Statistical Office (FSO)
- Collaboration with Monika Ferster, Jean-Paul Kauthen, Daniela Lussmann, Olivier Wirz (all from FSO) and Juan-David Berdugo, Marc Bill, Ruedi Niederer (all from FHNW)
- Data: Swiss Structural Survey 2013
- Objectives:
  1. analysis of statistical data preparation process (SDPP)
  2. investigating potential for improvement
  3. develop indicators for the users of the data

Swiss Structural Survey 2013

- Yearly survey to complement the register based census in Switzerland
- 280’000 persons
- mail and online
- Person questionnaire: language, religion, migration, education, activity and occupation, commuting
- household questionnaire: household composition and dwelling including rent
- SDP process and methods developed by FSO

Data sets



Dx	Description	Observations	Variables
D1	Raw	283’926	449
D2	Matched	283’926	442
D3	Controlled	281’991	406
D4	Final	281’990	461

User is interested in change from D1 to D4, i.e. from raw to final data

Producer is interested in all changes, D1 to D2 to D3 to D4, i.e. in the process

Variables

- Questionnaire variables (person and household, each tick one variable)
- Imputation flags (established by FSO) indicating a change compared to the preceding stage.
  - Binary flags =1 if change, =0 if no change
  - Complex flags with three categories indicate deterministic, stochastic (nearest neighbour) or mixed imputation
- Weights:
  - Initial weight for raw data
  - Person and household weight for final data

Questions, Variables and Variable Groups

1. Main language (Q1): multiple response question with 10 items
2. Completed education (Q8): multiple response question with 13 items
3. Current activity status (Q11): multiple response question with 9 items
4. Status in employment (Q13): single response question with 10 items
5. Net rent (rentnet) (Q33): quantitative variable

- First four questions (from person questionnaire) are treated as response groups (e.g. all 10 items of mainlanguage form a response group)
- **rentnet** is a household variable

11. Welches ist Ihre gegenwärtige Situation auf dem Arbeitsmarkt? (mehrere Angaben möglich)

Kreuzen Sie alle zutreffenden Antworten an; zählen Sie auch kleine Gelegenheitsjobs dazu.

Sie sind erwerbstätig, wenn Sie:

- mindestens eine Stunde pro Woche einer bezahlten Arbeit nachgehen,
- oder im Betrieb eines/einer Familienangehörigen unbezahlt arbeiten,
- oder Ihrer Arbeit vorübergehend fernbleiben (Ferien, Krankheit oder bezahlter Mutterschaftsurlaub, Militär-/Zivildienst), ansonsten jedoch unselbstständig oder selbstständig erwerbstätig sind.

Lehrlinge kreuzen sowohl «Erwerbstätigkeit» als auch «in Ausbildung» an.

1 ☐ eine Erwerbstätigkeit (Vollzeit) → durchschnittliche Anzahl Stunden pro Woche: ☐☐

2 ☐ eine Erwerbstätigkeit (Teilzeit) → durchschnittliche Anzahl Stunden pro Woche: ☐☐

3 ☐ mehrere Erwerbstätigkeiten (Teilzeit) → durchschnittliche Anzahl Stunden pro Woche: ☐☐

4 ☐ auf Stellensuche (bei der Arbeitslosenversicherung gemeldet oder nicht)

5 ☐ in Ausbildung (Schule, Studium, Lehre)

6 ☐ Hausfrau/-mann

7 ☐ invalide oder teilinvalide Person (z.B. IV-Rentner/in)

8 ☐ pensioniert (AHV, andere Pension) oder Rentner/in ausser Invalidenrenten

9 ☐ andere Situation ohne Erwerbstätigkeit

Die Fragen 12, 13, 14 richten sich nur an Erwerbstätige, inklusive Lehrlinge.

Structural missingness flags

Structural missingness occurs when a question is filtered out. E.g. unemployed persons do not have to give their status in employment.

Input:

- Filtering variables (e.g. current activity status)
- Filtering condition (e.g. no tick in first three items)
- Filtered variables (e.g. status in employment)

Output for status in employment as the filtered variable:

- $b_{sie} = 0$ if **not** structurally missing (default)
- $b_{sie} = 1$ if structurally missing (equivalent to a response)

Notation

- $\hat{y}_{ij}$ is the value of variable j of observation i.
- flags: r for response, b for structural missingness, g for imputation (change).
- weights: w
- Set of observations: S , group of variables: A
- E.g. imputation ratio on final data set D4 with global imputation flag g_{14} and raw response flag r_{14}:

$$IMROR = \frac{\sum_{i \in S} w_{4i} \sum_{j \in A} r_{14ij} \left(1 - b_{4ij}\right) \; g_{14ij} \hat{y}_{ij}}{\sum_{i \in S} w_{4i} \sum_{j \in A} \left(1 - b_{4ij}\right) \hat{y}_{ij}}$$

Implementation in 

Software and environment for statistical calculations (Version 3.2.2)

- Indicators and utilities implemented as an R-package `sdap` with documentation
- Processes in R scripts

Indicators

- Unit response rate (URR)
- Item response rate (IRR)
- Imputation rate (IMR) and imputation rate for responded items (IMRR)
- Item response ratio (IRO) and item response ratio for resp. items (IROR)
- Imputation ratio (IMRO) and imupation ratio for resp. items (IMROR)
- Imputation impact (IMI) and imputation impact for resp. items (IMIR)
- Structural missingness rate (SMR)

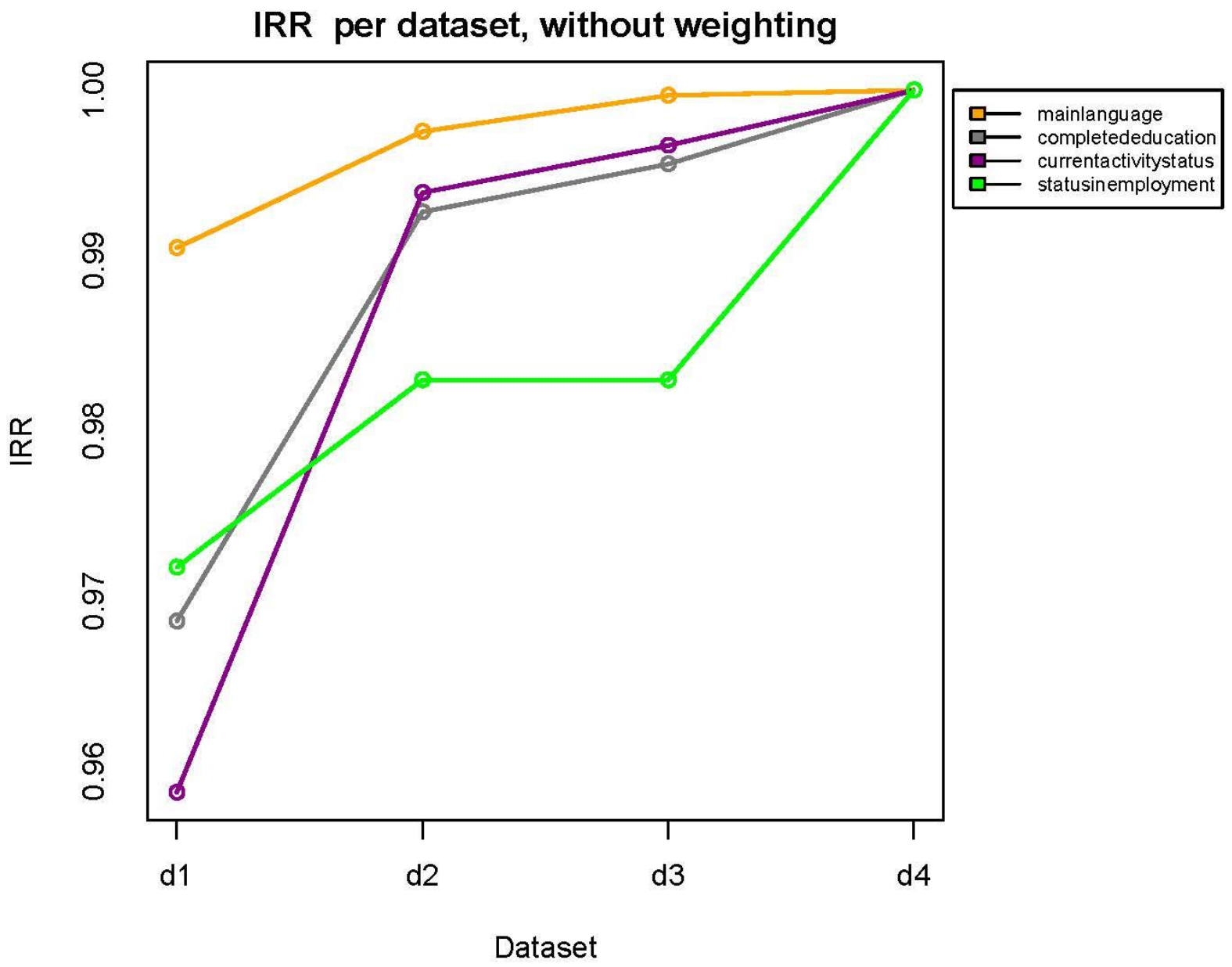
Results on final data set D4

raw response flag r_{14}, global imputation flag g_{14}, weighted

D4-r14-g14-w	urr=irr	iro	imr	imrr	imro	imror
rentnet	0.8852	0.8081	0.2633	0.0674	0.2539	0.0621
statusinemployment	0.9698	0.9522	0.0095	0.0040	0.0521	0.0043
completededucation	0.9723	0.9809	0.0115	0.0068	0.0301	0.0110
currentactivitystatus	0.9616	0.9658	0.0170	0.0113	0.0796	0.0454
mainlanguage	0.9928	0.9934	0.0051	0.0033	0.0190	0.0124

IMI(statusinemployment)	0.0138
IMIR(statusinemployment)	0.0015
SMR(statusinemployment)	0.3691
SMR(rentnet)	0.4138

IRR


IRR per dataset, without weighting
IRR
1.00
0.99
0.98
0.97
0.96
d1
d2
d3
d4
Dataset
mainlanguage
completededucation
currentactivitystatus
statusinemployment

Number of imputations

# imputations*	0	1	2	3	4
main language	274393	6234	1363	0	0
completed education	260904	20778	308	0	0
current activity status	250341	30571	1078	0	0
status in employment	266976	14988	25	1	0
rentnet	115719	142469	3555	19951	296

* Including coding for structurally missings.

Conclusions for Structural Survey

- Person variables
  - Reasonable number of missing values
  - Low imputation ratios for individual variables. Highest with current activity status (8%)
- Quantitative variable rentnet
  - difficult to respond (information retrieval and exact definition)
  - difficult to treat (outliers, only soft control rules)
  - Important imputation ratio (25%) does not show the change due to imputation (much smaller!)
- Efficiency of SDPP is high, no obvious potential for improvement!

Conclusions for SDPP indicators

- Core set of indicators (URR, IRR, IRO, IMR, IMRO, IMI, SMR) is useful
- Application to other variables is possible
- Application to other surveys is desirable
- Full value of the indicators for
  - comparison between editions of the same survey
  - effect of changes in SDPP (methods, parameters)
- Documentation and archiving of indicators for perodic surveys to be developed!

Some References

[Kilchmann2014]: KILCHMANN, D.: Statistischer Datenaufbereitungsprozess im BFS (Draft), Bericht, Swiss Federal Statistical Office, 2014

[Luzi2007]: LUZI, O.; WAAL, T. D.; HULLIGER, B.; ZIO, M. D.; PANNEKOEK, J.; KILCHMANN, D.; GUARNERA, U.; HOOGLAND, J.; MANZARI, A. & TEMPELMAN, C.: Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys. In: ISTAT, CBS, S. E. (Hrsg.): *Italian Statistical Institute ISTAT,*., 2007

[essqual14]: QUALITY TEAM OF EUROSTAT: ESS Guidelines for the Implementation of the ESS Quality and Performance Indicators (QPI): *European Commission, Eurostat*., 2014

[R2015]: R CORE TEAM: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2015