



GENERALIZED SOLUTIONS FOR DATA EDITING AT SURS

Rudi Seljak, Kaja Malešič
Statistical Office of the Republic of Slovenia

4th Baltic-Nordic Conference on Survey Statistics
Helsinki, August 24 - 28, 2015



Contents

- Introduction
- Development of statistical data processing at SURS
- Main characteristics of data editing in MetaSOP
- MetaSOP IT environments
- Basic architecture of the editing module
- Graphical interfaces and examples
- Impacts of the new generalized approach
- Future challenges



Introduction

- Data editing is a process aimed at detecting and correcting errors
- Rationalization of statistical processes
- The need for transition:
 - from custom made solutions for surveys (stovepipe approach) to generalized process solutions
 - from domain oriented to process oriented production



Development of statistical data processing at SURS

- 2007 generic SAS programs
 - smaller generalized solutions for different parts of process (“building blocks“)
 - meta-driven approach
- 2014 - **MetaSOP** application (SOP =Slovenian acronym for Statistical Data Processing)
 - a general tool for data processing: data editing (in production), being tested or developed modules: aggregation and standard error estimation, tabulation and tabular protection, quality indicators



Main characteristics of data editing in MetaSOP

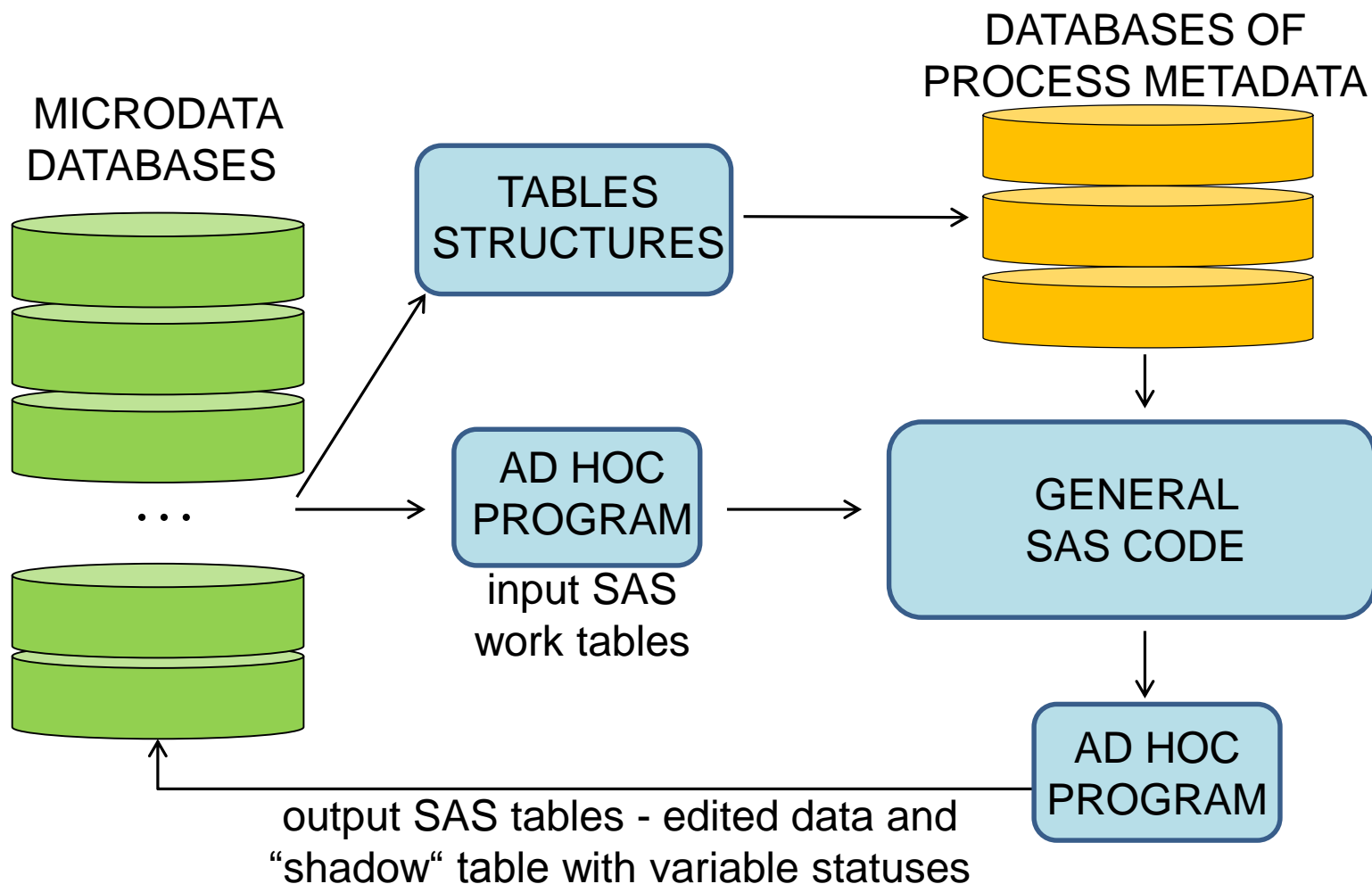
- Traceability, repeatability and reproducibility
 - all changes in data are recorded with statuses of the record (information in which process (phase) the record was changed) and statuses for each variable (information how the change has been performed)
- Meta-data driven approach
- Microdata databases have to be designed according to standard rules
- Standard quality indicators are used to monitor the process



MetaSOP IT environments

- .Net WPF application - the visible part of the system used by subject-matter statisticians
- SAS macros as general programs for data processing
- ORACLE database of process metadata
 - information about tables and variables
 - edit checks and rules in SAS syntax which are entered in the .Net application

Basic architecture of the editing module



Graphical interfaces

Groups of graphical interfaces:

- Interfaces for selection and preparation of the survey and survey instance:
 - for importing metadata from previous instances
 - for information on variables and definition of new “process” variables
 - for definition of processes (phases)
- Interfaces for management of the process metadata
 - logical checks
 - deterministic systematic and individual corrections
 - imputations
- Interface for running the processing (running the SAS macros)
- User administration interface

Example - Interface for logical checks

The screenshot shows the 'Kontrole' (Checks) section of the S256 - EU-SILC / 2014 (letna) - MetaSOP application. The interface is divided into several parts:

- Navigation Menu (Left):** A list of categories including DOHODNINA, GOSP, INT_GOSP_V, INT_OSEB_V, OSEB, OSTALL_VIRI, and SRDAP. An arrow points from this menu to the text 'TABLES'.
- Table of Checks (Center):** A table with columns for Oznaka (Label), Pogoj (Condition), Opis (Description), and Opomba (Note). An arrow points from this table to the text 'LIST OF LOGICAL CHECKS'. The table contains several rows of checks, including LK069, LK070, LK070A, LK071, LK072, LK073, LK074, and LK074A.
- Detailed View (Bottom):** A form showing the details for check LK069. An arrow points from this form to the text 'LOGICAL CHECK'. The form includes fields for Veljavnost (checked), Oznaka (LK069), Vrsta (DRUGO), Pogoj (Condition), Opis (Description), and Opomba (Note).

Oznaka	Pogoj	Opis	Opomba
LK069	$((AK8 \leq 5) \text{ or } (AK8 > 40))$	Otroci, (stari od 6 do 12 let), ki so bili v šoli manj kot 5 ur ali več kot 40 ur na teden	popravimo sistemsko
LK070	$AK10 > 40 \text{ and not } (AK8 > 40)$	Otroci, (stari od 6 do 12 let), ki so bili v šoli manj kot 5 ur ali več kot 40 ur na teden	
LK070A	$(AK8 + AK10 > 55) \text{ ar } (AK8 > 40)$	Otroci, (stari od 6 do 12 let), ki so bili v šoli manj kot 5 ur ali več kot 40 ur na teden	
LK071	$AK12 > 45 \text{ and not } (AK8 > 40)$	Otroci, ki so bili v varstvenem centru, ki so bili v varstvenem centru manj kot 5 ur ali več kot 40 ur na teden	
LK072	$AK14 > 50 \text{ and not } (AK8 > 40)$	Otroci, ki so bili v varstvenem centru, ki so bili v varstvenem centru manj kot 5 ur ali več kot 40 ur na teden	
LK073	$AK16 > 60 \text{ and not } (AK8 > 40)$	Otroci, ki so bili v varstvenem centru, ki so bili v varstvenem centru manj kot 5 ur ali več kot 40 ur na teden	popravimo na zgornji
LK074	$(AK6 > 0) \text{ and } (AK8 > 1)$	Če je otrok v vrstcu ne glede na pretekla leta	
LK074A	$(AK6 > 60) \text{ and not } (AK8 > 40)$	Trinisti otroci, ki en teden	

Veljavnost:

Oznaka: LK069

Vrsta: DRUGO

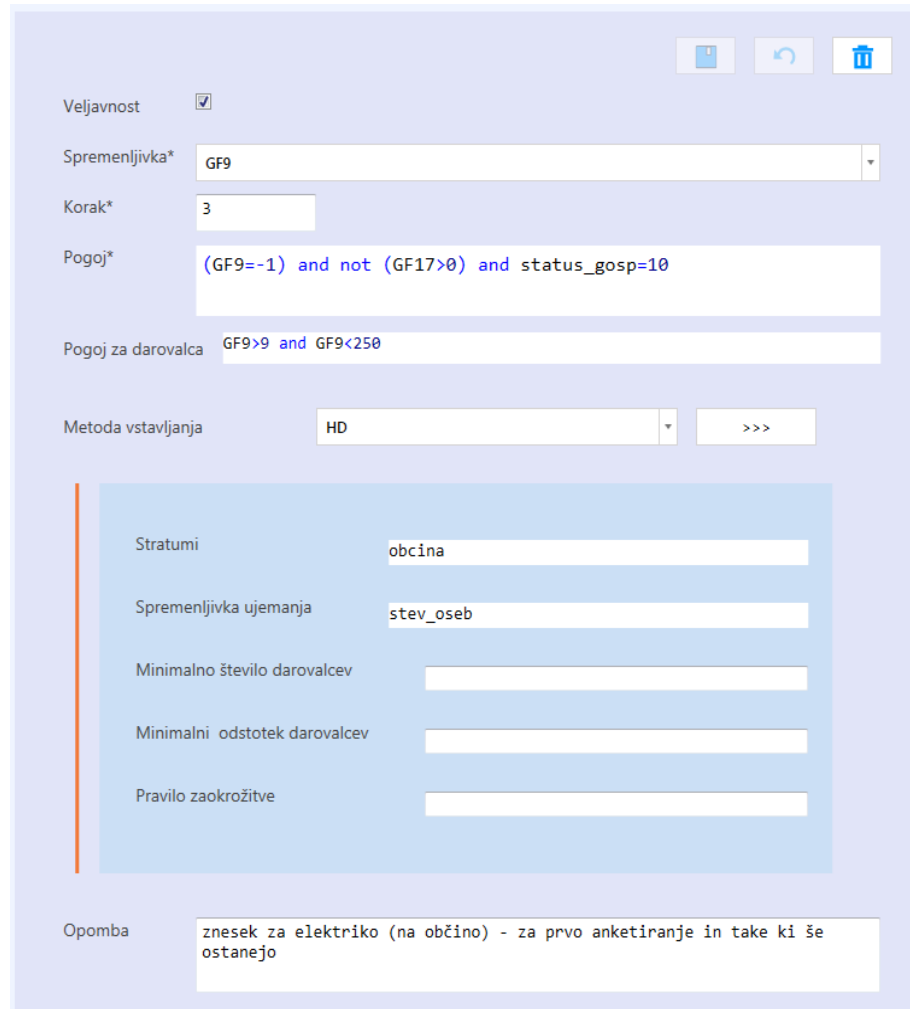
Pogoj: $((AK8 \leq 5) \text{ or } (AK8 > 40)) \text{ and not } (AK8 \text{ in } (-2 -1)) \text{ and } (STAR \text{ in } (6 7 8 9 10 11 12))$

Opis: Otroci, (stari od 6 do 12 let), ki so bili v šoli manj kot 5 ur ali več kot 40 ur na teden

Opomba: popravimo sistemsko; ker ne more biti otrok na teden v šolo manj kot 5 ur v povprečju (glej popravki_sist)

Example - Parameters for a „hot-deck“ imputation

- Variable
- Step (of the imputation process)
- Condition (determines for which units the imputation procedure will be performed)
- Condition for donor (determines which units can serve as the suitable donors)
- Method of imputation (HD in this case)
- Stratum (stratification variables for donors)
- Matching variable
- Minimum number of donors
- Minimum percentage of donors
- Rounding rule
- Note



The screenshot shows a configuration interface for hot-deck imputation. It includes the following fields and options:

- Veljavnost**:
- Spremenljivka***: GF9
- Korak***: 3
- Pogoj***: `(GF9=-1) and not (GF17>0) and status_gosp=10`
- Pogoj za darovalca**: `GF9>9 and GF9<250`
- Metoda vstavljanja**: HD
- Stratumi**: občina
- Spremenljivka ujemanja**: stev_oseb
- Minimalno število darovalcev**: [empty field]
- Minimalni odstotek darovalcev**: [empty field]
- Pravilo zaokrožitve**: [empty field]
- Opomba**: znesek za elektriko (na občino) - za prvo anketiranje in take ki še ostanejo



Example - Interface for process running

TABLES

PROCESSES

PROCESS RUNNING

BASIC INFORMATION ABOUT PROCESSING

BASIC INFORMATION ABOUT PROCESS RESULTS

LINK TO LOG AND DIFFERENT OUTPUTS

Status	Zacete	Konec	Datum
Končano	10:52:41	10:52:47	16.10.2014
Končano	10:51:54	10:52:01	16.10.2014
Napaka	10:49:57	10:50:00	16.10.2014
Napaka	10:47:56	10:47:58	16.10.2014
Napaka	10:47:46	10:47:48	16.10.2014
Napaka	10:46:50	10:46:54	16.10.2014
Končano	14:33:00	14:33:06	15.10.2014

kontrola	POGOJ	Stevilo
LK2	IF EB = 2	56
LK_5	IF CESNJE>10	91
LK_55	IF x<-5 and cebula<-1 and TEST_DUMMY=1	0
LK_6kk	IF ZELJE>3700 and KROMPIR>5000	0



Impacts of the new generalized approach

- Main goal is rationalization of statistical processes and overall improvement in data quality.
- Changes in work organization:
 - Different distribution of work among subject-matter statisticians, general methodologists and IT experts.
 - Subject-matter statisticians are trained to write the edit checks and rules themselves in the form of SAS syntax.
 - Surveys are less dependent on an individual subject matter statistician, methodologist or IT expert.
- Elimination of errors in consistency between entered rules and variables.
- Microdata databases must be designed according to standard rules.



Future challenges

- All modules of MetaSOP introduced into production (aggregation and standard error estimation, tabulation and tabular protection, quality indicators)
- Procedures for selective editing
- Implementation of data editing for all surveys at SURS with MetaSOP Application