GENERALIZED SOLUTIONS FOR DATA VALIDATION AND DATA EDITING AT SURS

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

Workshop
ValiDat Foundation - Towards a Common Validation Policy in the ESS
Wiesbaden, Germany, 10-11 November 2015
Contents

• Introduction
• Development of statistical data processing at SURS
• Main characteristics of data validation and data editing in MetaSOP
• MetaSOP IT environments
• Basic architecture of the editing module
• Ad hoc program for logical checks
• Graphical interfaces and examples of interfaces for logical checks
• Impacts of the new generalized approach
• Future challenges
Introduction

- Data validation at SURS is incorporated into data editing process
- 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, aggregation, tabulation (in production), being tested or developed modules: standard error estimation, tabular protection, quality indicators
Main characteristics of data validation and data editing in MetaSOP

- The module consists of three parts:
  - logical checks
  - deterministic systematic and individual error corrections
  - imputations

- 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 on how the change has been performed)
  - validation checks can be run on any record status (phase of record change)

- Meta-data driven approach
- Microdata databases have to be designed according to standard rules
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
  - logical checks and edit rules in SAS syntax which are entered in the .Net application
Basic architecture of the editing module

MICRODATA DATABASES

... → TABLES STRUCTURES

AD HOC PROGRAM
input SAS work tables

DATABASES OF PROCESS METADATA

GENERAL SAS CODE

AD HOC PROGRAM
output SAS tables - edited data and “shadow“ table with variable statuses
Ad hoc program for logical checks

- Input SAS work table: basic table + joined variables from other tables that are needed for logical checks
- General SAS macro for logical checks
- Output microdata table with indicator variables for each logical check which is exported to a network drive
Graphical interfaces in the editing module

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 (derived variables and aggregated variables)
  - for definition of processes (phases) – logical checks can be run on any phase

- 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
Interface for logical checks

- Interfaces for selection and preparation of the survey and survey instance:
  - for importing metadata from previous instances
  - for variables metadata and definition of new "process" variables
  - for definition of process stages

- 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

LIST OF LOGICAL CHECKS

LOGICAL CHECK

TABLES
Interface for process running

- TABLES
- PROCESSES
- BASIC INFORMATION ABOUT PROCESSING
- PROCESS RUNNING
- LINK TO LOG AND DIFFERENT OUTPUTS
- BASIC INFORMATION ABOUT PROCESS RESULTS
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 logical checks and edit 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 (standard error estimation, tabular protection, quality indicators)

• Implementation of data editing for all surveys at SURS with MetaSOP Application

• Procedures for logical checks on macrodata