Administrative data - editing

Danilo Dolenc

GCC-STAT Regional Workshop

Muscat, Oman, 22-24 September 2019

Introduction

- General about strategy on editing administrative sources
  - Does differ from editing of surveys
- Identifiers – to be or not to be of the register-based statistics
- Individual administrative sources – how far we can go
- Integrated administrative data editing – the crown
  - Brilliant from the crown – register-based census

Four strategy scenarios

- At least two administrative sources available

Edit separately each single adm. source	No editing at all	“Light“ editing of each single adm. source	Edit selected adm. sources
			No editing of selected adm. sources
Integration of edited sources	Integration of non-edited adm. sources	Integration of edited sources	Integration of edited and non-edited adm. sources
Edit the integrated sources	Edit the integrated sources	Edit the integrated sources	Edit the integrated sources

Drivers for choosing strategy (1)

- Good knowledge on data sources performances and quality of input data
  - Set in advance some quality indicators for each variable
    - Share of missing data (including unknown category)
- Trade-off between expected quality, time needed and resources available
  - **No statistical editing is perfect**
    - Too much data and inter-dependency between variables

Drivers for choosing strategy (2)

- Experience of subject-matter experts
  - The most responsible for data editing
- Experience of general methodologists
  - To propose the methods for data editing
- Statistical processing organization inside NSI
  - Are there some generic solutions
  - IT experts must provide support and develop adequate tools (programmes) for execution of methodological guidelines

Drivers for choosing strategy (3)

- Strategic paper about role of subject-matter specialists, general methodologists and IT expert
  - Modernisation of statistical processing at SURS

Editing administrative vs survey data – single source (1)

- Editing of single source not differs a lot
  - All errors are interpreted as content errors only
  - Not possible to contact responding unit in case of business data
    - Persons in sample surveys are normally not re-contacted
      - Except follow-up of interviewer work
  - Data from admin. sources are pre-edited
    - We expect / assume better quality
  - Data quality from surveys depend also on mode of data collection and build-in checks
    - Paper questionnaire, CAPI, CATI, web services…

Editing administrative vs survey data – single source (2)

- Traditional methods of (mostly) automatic data editing are applied for administrative data
  - Range of values and outliers detection
  - Matching with classification used for each variable
  - Distributional check
  - Duplicate records detection
  - Comparison with distribution known from other sources
  - Consistency check
    - Relation between several variables of the same unit

Editing administrative vs survey data – single source (3)

- Imputation = replace missing values
  - Just single variable (not available in more than one source) imputed
    - Additional information for editing available in other sources
      - Case: formal marital status
  - Methodological decision needed to choose variables to be imputed or not
    - Depend on diversity of categories
      - Case: sex, age vs occupation
        - Working abroad

Editing administrative vs survey data – single source (4)

- Macro-editing or output editing
  - Based on historical data (previous outputs)
  - External (aggregated) sources could be used
  - Main aims
    - To analyse the outliers
    - To discover influential errors
      - The simplest way – pivoting corresponding variables

Editing integrated sources

- Creating new derived variables
    - Complex in case of several sources
- Consistency editing of variable errors
    - Confronting same variable from different sources
        - Most common – no variable error in a single source
- Consistency editing of object errors
    - Matching different units with same identifier
        - Important if identifiers are not standardized

Editing rules – general (1)

- Rule
  - Condition that should be satisfied that some statement is TRUE
    - IF AGE = 10 THEN Activity_Status = Child
    - IF AGE = 10 THEN Edu_Participation = Primary_school

Editing rules – general (2)

- Rule
  - Condition that should be satisfied that some statement is FALSE
    - IF AGE = 10 AND Activity_Status = Employed
    - IF AGE = 10 AND Edu_Participation = missing
      - Hard (fatal) and soft rules from processing point of view
      - Influential and non-influential errors from dissemination point of view

Editing rules – automated editing (3)

- Firstly - inventory of rules for checking consistency
  - The most important relations between variables – to cover influential errors
    - Age vs labour force status / educational attainment
    - Year of birth vs year of immigration
- Secondly – the corrected (TRUE) value is determined

Editing rules – automated editing (4)

- The order of automated corrections is very important
    - Determine which variable is “dominant“ to be corrected first
        - Educational attainment vs participation in formal education
- In some cases, more than one step is needed
    - Following logical connections between variables
        - Citizenship vs country of birth vs country of previous residence

Editing rules – imputation (5)

- Several methods of imputation exist
  - Most of them are used in business statistics
    - Logical, mean value, historical, structural, regression, distributional, donor
  - In social statistics hot deck (internal donor) method is dominant
    - Value is taken from another record in database
      - Donor could be determined randomly within a large group of units (e.g. students)
      - More often we search for similarity of recipient and donor with respect to more matching variables

Editing rules – imputation (6)

- Hot deck method - imputing labour force status
  - Define stratum – the large group from which donor will be selected
    - Non-nationals of chosen citizenship
  - Define matching variables
    - Age (could be single age or broader age group)
    - Sex
  - Define minimum number / share of donors
    - If number below threshold imputation is not executed

Identifiers – some basics

- Register-based statistics depends on exact matching
  - Primary and secondary keys
    - Primary key in basic source must be unique
      - Case: PIN of person vs PIN of parents
  - Missing identifier in individual source
    - Missing record (under-coverage)
- To collect PIN's in field survey or not
- Identifiers in register-based census
  - Combining primary and secondary keys
- How to construct identifier

Missing personal identifiers

- Persons without identifiers (=not being registered) could be find in surveys only
  - Two possible options to solve
    - To generate new “artificial“ identifier
    - To impute identifier
      - Intentional object error
  - Collecting PIN‘s in the field is not recommended
    - Application for determination PIN set up inside NSI
      - Based on address, name, surname, date of birth, sex
      - Probabilistic approach for non-exact math

Application for determination PIN

- The whole history of CPR = donor database
  - 3.6 mio unique PIN's
  - Updated monthly
  - The matching results depend on quality of field work
    - SILC 2019 results – 9,000 new entries
      - 97.4% - full match
      - 1.2% - random match with high probability
      - 1.4% - no match or below probability threshold (125 cases)
        - 118 found manually by adding other variables to search
          - Place of birth, relations between children and parents
        - 7 records not possible to match

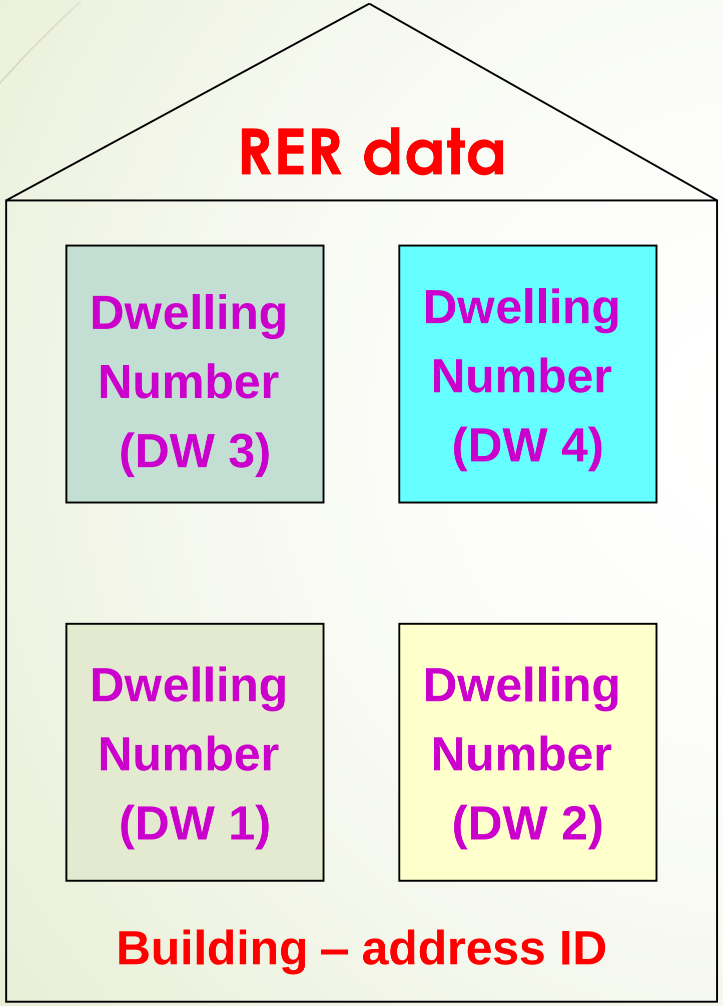
Distinguishing Power Concept (1)

- Creating identifiers if they do not exist
  - Distinguishing power relates to uniqueness of the values of variables intended for matching key
    - High distinguishing power variables
      - Full name, address, date of birth
    - Low distinguishing power variables
      - Sex, age, citizenship
    - Variables with less changeability more appropriate
  - The same topic must be available in all sources foreseen for matching

Distinguishing Power Concept (2)

- Practical example from our donor database
  - Variables joined together using function CAT in SAS
    - First name + first surname
      - 50% unique, 14% duplicates, 36% triplicates or more
    - First name + first surname + date of births
      - 99.93% uniqueness - 2,538 duplicates
    - First name + first surname + date of births + sex
      - 99.94% uniqueness - 2,009 duplicates
    - First name + first surname + date of births + sex + address
      - 99.98% uniqueness - 686 duplicates

Census data integration



RER data

Dwelling Number (DW 3)

Dwelling Number (DW 4)

Dwelling Number (DW 1)

Dwelling Number (DW 2)

Building – address ID

PIN	Address ID	DW		PIN	Address ID	HH
108979529	23470898	3		108979529	23470898	1
123457805	23470898	3		123457805	23470898	1
250789532	23470898	3		250789532	23470898	1
498230857	23470898	3		498230857	23470898	1
897600036	23470898	2		897600036	23470898	2
345678149	23470898	2		345678149	23470898	2
340090023	23470898	2		340090023	23470898	2
987650128	23470898	2		987650128	23470898	2
145092232	23470898	4		145092232	23470898	3
567725951	23470898	4		567725951	23470898	3
658735773	23470898	4		658735773	23470898	4
100089700	23470898	4		100089700	23470898	4
789568391	23470898	4		789568391	23470898	4
135790740	23470898	4		135790740	23470898	4

CRP data

HR data

Census data integration – process and identifiers – step by step (1)

- 1. Usual residence population derived from CPR (T+3) = basic census table PERSONS
  - PIN (no missings, primary key)
  - PIN_S (spouse, secondary key)
  - PIN_M (mother, secondary key)
  - PIN_F (father, secondary key)
  - Address ID - A_ID (no missings, secondary key)
  - Dwelling ID - D_ID (missings, secondary key to A_ID)
- 2. Administrative data (CPR - T+0) used for update of missing D_ID
  - PIN(P) = PIN(2) AND A_ID(P) = A_ID(2) THEN D_ID(P) = D_ID(2)

Census data integration – process and identifiers – step by step (2)

- 3. Integration of household data (T+0)
  - PIN (no missings, primary key)
  - Address ID - A_ID (no missings, secondary key)
    - Dwelling ID due missings not used as identifier
  - Household ID (H_ID) (no missings, secondary key)
  - Relation to the reference person (HH) – missings
    - Special key used for automated derivation of family data
    - Matrixes of unique relations in the household prepared in advance
  - PIN(P) = PIN(3) AND A_ID(P) = A_ID(3) THEN H_ID(P) = H_ID(2) AND HH(P) = HH(2)

Census data integration – process and identifiers – step by step (3)

- 4. Determination of D_ID and H_ID for collective living quarters
  - Address based list distinguishing six large groups
    - Student residences, old people's homes, social welfare institutions (for adults, for younger population), penal and correctional institutions, religious institutions
  - Address ID - (no missings, primary key)
  - Dwelling ID – statistically determined special code
  - Household ID - statistically determined special code
  - HH - statistically determined special code
  - A_ID(P) = A_ID(4) THEN D_ID(P) = D_ID(4) AND H_ID(P) = H_ID(4) AND HH(P) = HH(4)

Census data integration – process and identifiers – step by step (4)

- 5. Extracting building and dwelling data from Real Estate Register (T+0) = set up basic census table DWELLINGS
  - Building ID – B_ID (no missings, primary key)
    - For simplicity reason we equalize B_ID and A_ID here
  - D_ID – (no missings, secondary key)
  - Derived variable - Type of use of building part – assigned to differ dwellings and other non-dwelling parts
    - TYPE = 1 – dwelling
    - TYPE = 2 – non dwelling

Census data integration – process and identifiers – step by step (5)

- 6. Update DWELLINGS table with D_ID from address-based list (step 4)
  - New record imputed
  - A_ID(D) = A_ID(4) THEN D_ID(D) = D_ID(4) AND TYPE(D) = 2
- 7. Linkage PERSONS and DWELLINGS table by using A_ID and D_ID as composed key to detect
  - Persons (PIN‘s) without D_ID
  - Not matched D_ID(P) and D_ID(D)
    - In most cases error in population database
  - Empty dwellings

    A_ID(P) = A_ID(D) and D_ID(P) <> D_ID(D) AND TYPE = 1

Census data integration – process and identifiers – step by step (6)

- 8. Automated editing of missing identifiers in table PERSONS (D_ID, H_ID, HH)
  - Key – A_ID
    - Several rules starting from simple (deductive) to very complex solutions

BEFORE		AFTER	
D_ID	H_ID	D_ID	H_ID
1	5	1	5
1	5	1	5
1		1	**5**

BEFORE		AFTER	
D_ID	H_ID	D_ID	H_ID
7	3	7	3
7	3	7	3
	3	**7**	3

Census data integration – process and identifiers – step by step (7)

- 8. Automated editing of missing identifiers in table PERSONS (D_ID, H_ID, HH)
  - Key – A_ID
    - Very important is the order of execution of the rules
    - Table EMPTY_DWELLINGS created for imputation of D_ID
      - No imputations if there is no empty dwelling at the address
    - For missing H_ID and HH identifier the rules based on relations were used
    - No imputations for children 0-17 years without PIN‘s of parents
      - Non-nationals mostly

Census data integration – process and identifiers – step by step (8)

- 8. Automated procedures for H_ID and HH based on PIN‘s
  - Used for replace missing values
  - Used also for checking correctness
    - At least one link to at least one other household member must exist

BEFORE					
H_ID	HH	PIN	PIN_S	PIN_M	PIN_F
3	00	A			
		C	D	A	B
		D	C	E	

AFTER					
H_ID	HH	PIN	PIN_S	PIN_M	PIN_F
3	00	A			
3	03	C	D	A	B
3	08	D	C	E	

Census data integration – process and identifiers – step by step (9)

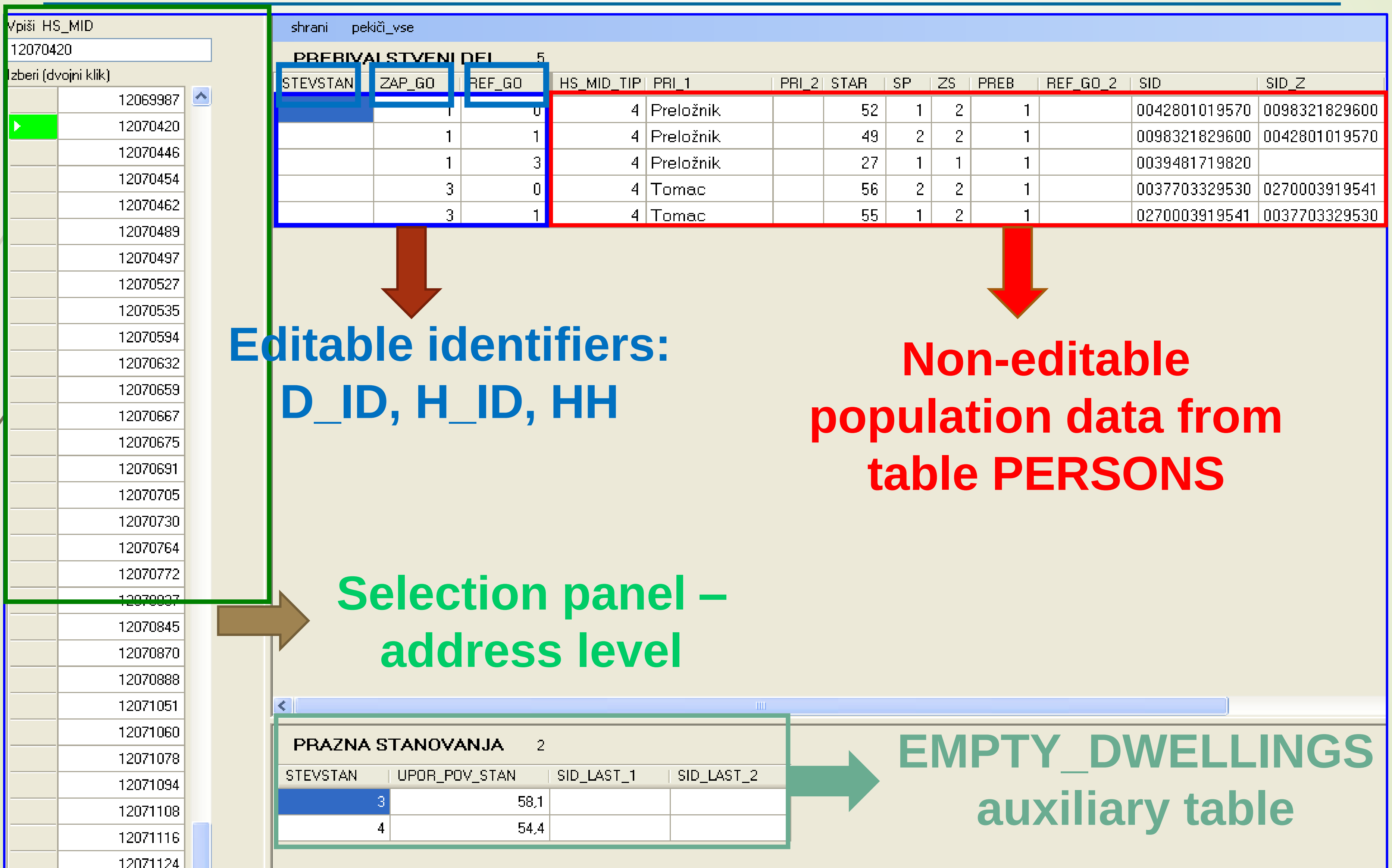
- 8. Automated procedures for H_ID and HH based on PIN's
  - Relations depend on selection of HH

HH	Sub-matrix 1	HH	Sub-matrix 2	HH	Sub-matrix 3
00	Reference person	00	Reference person	00	Reference person
03	Daughter	01	Spouse (husband)	01	Spouse (wife)
08	Son-in-law	05	Mother	06	Mother-in-law

Census data integration – process and identifiers – step by step (10)

- 9. Manual editing of missing identifiers in table PERSONS (D_ID, H_ID, HH)
  - Key – A_ID where at least one identifier is missing
    - Very important – surnames were used for connecting children with parents
    - Interface prepared for manual data entry
    - Possible to correct already edited data
      - But only identifiers could be corrected

Interface – manual editing



Vpiši HS_MID

12070420

Izberi (dvojni klik)

12069987
12070420
12070446
12070454
12070462
12070489
12070497
12070527
12070535
12070594
12070632
12070659
12070667
12070675
12070691
12070705
12070730
12070764
12070772
12070837
12070845
12070870
12070888
12071051
12071060
12071078
12071094
12071108
12071116
12071124

PREBIVALSTVENI DEL 5

STEVSTAN	ZAP_GO	REF_GO	HS_MID_TIP	PRI_1	PRI_2	STAR	SP	ZS	PREB	REF_GO_2	SID	SID_Z
	1	0	4	Preložnik		52	1	2	1		0042801019570	0098321829600
	1	1	4	Preložnik		49	2	2	1		0098321829600	0042801019570
	1	3	4	Preložnik		27	1	1	1		0039481719820	
	3	0	4	Tomac		56	2	2	1		0037703329530	0270003919541
	3	1	4	Tomac		55	1	2	1		0270003919541	0037703329530

Editable identifiers: D_ID, H_ID, HH

Non-editable population data from table PERSONS

Selection panel – address level

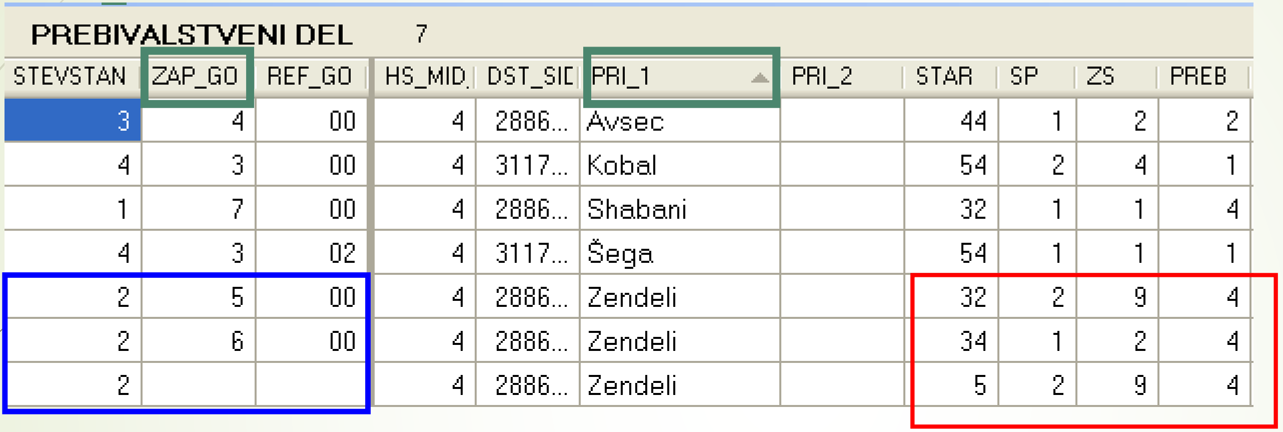
PRAZNA STANOVANJA 2

STEVSTAN	UPOR_POV_STAN	SID_LAST_1	SID_LAST_2
3	58,1		
4	54,4		

EMPTY_DWELLINGS auxiliary table

Interface – example

H_ID Surname

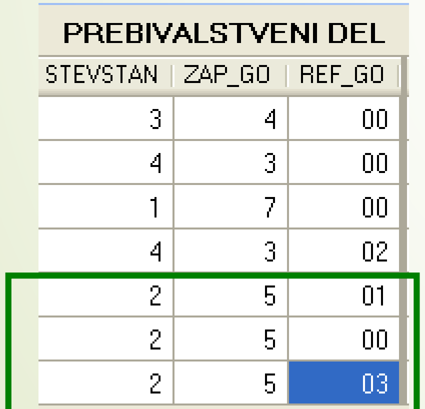


PREBIVALSTVENI DEL 7

STEVSTAN	ZAP_GO	REF_GO	HS_MID	DST_SIC	PRI_1	PRI_2	STAR	SP	ZS	PREB
3	4	00	4	2886...	Avsec		44	1	2	2
4	3	00	4	3117...	Kobal		54	2	4	1
1	7	00	4	2886...	Shabani		32	1	1	4
4	3	02	4	3117...	Šega		54	1	1	1
2	5	00	4	2886...	Zendeli		32	2	9	4
2	6	00	4	2886...	Zendeli		34	1	2	4
2			4	2886...	Zendeli		5	2	9	4

Demographic data

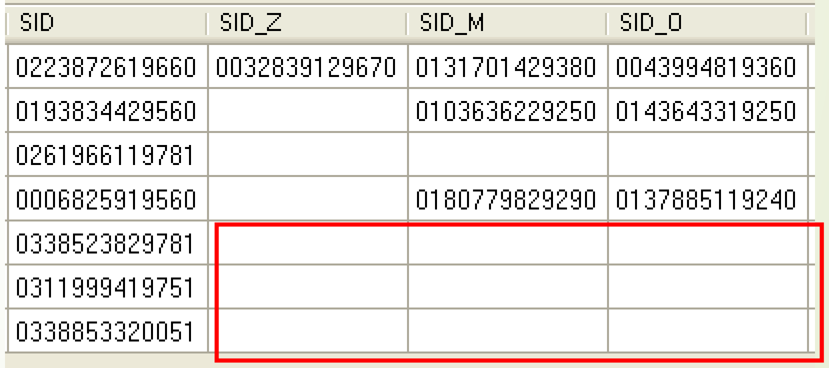
Input data after automated procedures



PREBIVALSTVENI DEL

STEVSTAN	ZAP_GO	REF_GO
3	4	00
4	3	00
1	7	00
4	3	02
2	5	01
2	5	00
2	5	03

Output data – manual correction



SID	SID_Z	SID_M	SID_O
0223872619660	0032839129670	0131701429380	0043994819360
0193834429560		0103636229250	0143643319250
0261966119781			
0006825919560		0180779829290	0137885119240
0338523829781			
0311999419751			
0338853320051			

No data on PIN‘s of father/mother/spouse

Census data integration – process and identifiers – step by step (11)

- 10. Final manual editing of inconsistencies between identifiers in table PERSONS (D_ID, H_ID) using interface
    - Key – A_ID + D_ID
        - Household ID's with two or more different dwelling ID's
            - (COUNT(DISTINCT(D_ID)) GROUP BY H_ID) >1

Creating derived variables from multisources (1)

- Methodological problem first
  - Depend on content of the variable and data sources available
- Possible approaches
  - The highest value is chosen (if numeric)
    - Case: Number of live-born children
  - The most quality value is chosen (if character)
    - Case: Educational attainment
  - Priority is given to the most trustable source
    - Case: Annual tertiary graduates

Creating derived variables from multisources (2)

- Possible approaches
    - The most timely updated source is used
        - Case: Marital status from CPR
    - The sub-population source fitted the most to the statistical concepts is used first
        - Case: Statistical Employment Register
    - Qualitative and quantitative analyses of each source taking into account objective criteria – a dream goal
        - But at the end also subjective decision is often needed to prioritize data sources

Case: Educational attainment (1)

- The main methodological problems
  - Population over 14 years observed
    - Different periods of education
    - Not comparable school systems
  - No sources available
    - For pupils finished obligatory elementary school
    - For pupils graduated from short-term vocational programmes
      - Information deduced from enrolment data

Case: Educational attainment (2)

- Basic editing principles
  - The hierarchy of the sources as a general rule
    - Modified in some particular combinations of levels of educational attainment available from two or more sources
  - Pre-editing - the highest education in case of several records for same person in the same source
    - Tertiary education graduates from 1989-2010
      - Object errors possible (but not identifiable)
  - Not harmonized classifications in sources
    - First step – re-coding to the national classification standard KLASIUS

Case: Educational attainment (3)

- Basic editing principles
  - The hierarchy of the sources as a general rule
    - Modified in some particular combinations of levels of educational attainment available from two or more sources
  - Pre-editing - the highest education in case of several records for same person in the same source
    - Tertiary education graduates from 1989-2010

SID	INPUT DATABASES AND PRIORITY								
	DIPL_TERC	MATURA	UN_EMPL	CHAMBER	STUD_TERC	PRIM	SOL_STIP	SRE	CENSUS
	1	2	3	4	5	6	7	8	9
A	17002							17002	17001
B		15001							12001
C	18202							17002	17003
D					15001		15002		15001
E			14002	14002					17002
F						12001	15002		
G									11002

DATA INTEGRATION						
SID	Nr. of values	MIN	MAX	DERIVED	SOURCE	COMMENT
A	3	17001	17002	17002	1	Trustable source
B	2	12001	15001	15001	2	Trustable source
C	3	17002	18202	18202	1	Trustable source
D	3	15001	15002	15001	5	Same level of education, value with higher count selected, source with higher priority indicated
E	3	14002	17002	14002	3	Census data are the less trustable, source with higher priority indicated in case of same value
F	2	12001	15002	15002	7	Higher value even lower priority in case of combination of sources 6 and 7
G	1	11002	11002	11002	9	Only one source

Case: Educational attainment (3)

- Sources by hierarchy (population 15+)

Priority	Owner	Source content	Period	Share from source		
				2011	2015	2018
1	SURS	Tertiary education graduates	1989 - 2010	11.1	12.5	14.2
2	NEC	Graduates of matura	2002 - 2010	9.1	9.3	9.4
3	Chambers	Vocational/masters exam	2002 - 2010	0.2	0.2	0.2
4	SURS	Students education at enrolment	2002/03-10/11	2.6	2.1	1.9
5	NEC	Primary school exam	2006 - 2010	4.6	4.6	4.6
6	SURS	Scholarship recipients	2006 - 2010	0.5	0.5	0.5
7	SURS	Educational attainment SRE	1986 - 2010	55.9	57.0	57.4
8	ESS	Registered unemployed persons	1.1.2011	0.8	1.5	1.5
9	SURS	2002 Census education	31.3.2002	13.6	10.7	8.7
10		Imputation		1.6	1.6	1.6

Case: Educational attainment (4)

- Annual update
  - The same sources and same methodology used on yearly basis (except 2002 Census)
    - Short and even period between two consecutive stocks is desirable
  - The educational attainment can‘t be decreased
    - Exception – the imputation in the previous year
  - The source indicator is changed in case of the same level of educational attainment but the priority of source is higher

Case: Educational attainment (5)

			CHANGES / IMPROVEMENT						
AGE	CENSUS 2011		2012-2016		2017		CENSUS 2018		COMMENT
	EDU	SOURCE	EDU	SOURCE	EDU	SOURCE	EDU	SOURCE	
42	17002	1					17002	1	No change
30	15001	2			16002	1	16002	1	Improvement
55	18202	1					18202	1	No change
32	15001	9			15001	4	15001	4	Change of source
60	14002	10					14002	10	No change
23	15002	8	17002	1			17002	1	Improvement
85	11002	IMP					11002	IMP	No change
40	15001	IMP			14002	9	14002	9	Change of value - higher value imputed previously
21			14001	3	15002	2	15002	2	Improvement

Conclusion

- Data processing in a register-based system (census) is a complex system including
  - Methodological issues
    - Usual residence population is a base
  - Defining the processing stages
    - Step by step
  - Data integration
  - Editing (data cleansing)
  - Outcomes evaluation

Register-based census – the future

- Register-based census method using several administrative and statistical sources is the answer to key objectives for future of the censuses
  - Negligible costs
  - Adequate quality of outputs
  - No respondent burden
  - Privacy
  - Frequency

Future of the traditional census

- Is a traditional census conducted every 10 years still feasible?
- Is there still a future for the traditional censuses beyond 2021?
- Every country must find its own way
  - The road is open for all