WP.4
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ljubljana, Slovenia, 9-11 May 2011)

Topic (i): Editing of administrative and Census data

Editing Process in the Case of the Slovenian Register-Based Census

Invited Paper

Prepared by Danilo Dolenc[1], Milojka Krek[2], Rudi Seljak[3], Statistical Office of the Republic of Slovenia

I. Introduction

1. The 2011 Slovenian Census of Population, Households and Housing will be for the first time completely register-based, meaning that since all the micro-data will be obtained by linking several administrative and statistical sources, no fieldwork will be carried out. Compared with the previous census, the main advantage is obviously the essential costs reduction (the field enumeration together with data processing represented more than 70% of costs).

2. The 2002 census budget was approximately EUR 10 million. Besides the saving of such a large amount of money, response burden reduction and shorter time needed for processing are another two aspects of advantage. The transition from the field census to the register-based census represents a major development achievement of the Slovenian statistics. With this, Slovenia joins few European countries that have already conducted such censuses (Denmark, Finland, the Netherlands, and Iceland) or will implement them for the first time in 2011 (Austria, Sweden, and Norway).

3. The editing part of the statistical process will consist of two main parts. In the first part the custom-made software programs will be used to integrate different sources, do the initial data cleaning and derive the composite variables. In the second part the general application, together with the interface for manual editing, will be used to complete the editing work.

4. In the first part of the paper we describe the main features of the census methodology, while in the second part we focus on the description of the planned editing process. The main steps of the process will be described from the methodological as well as from the more technical point of view. Also the main expected “critical points” and the measures taken to overcome them will be presented.

II. Statistical process in the previous censuses

[1] Danilo Dolenc, Statistical Office of the Republic of Slovenia, Parmova 33, 1000 Ljubljana, Slovenia, danilo.dolenc@gov.si, phone +386 1 23 40 876
[2] Milojka Krek, Statistical Office of the Republic of Slovenia, Parmova 33, 1000 Ljubljana, Slovenia, milojka.krek@gov.si, phone: +386 1 2340 672
[3] Rudi Seljak, Statistical Office of the Republic of Slovenia, Vožarski pot 12, 1000 Ljubljana, Slovenia, rudi.seljak@gov.si, phone: +386 1 24 15 294

5. The classical statistical process in the previous censuses (with the exception of the 2002 Census) was very similar to the processes of the "classical" statistical sampling surveys, which usually consist of the following main phases:

(a) Data collection (collection performed with around 10,000 trained enumerators collecting data by face-to-face interviews)
(b) Manual editing (manual coding, supplementing missing data, mostly identifications)
(c) Transfer of data from paper questionnaires to the electronic form (manual key entry or later also optical reading)
(d) Data processing in the narrower sense of the word (automated corrections, consistency checks mostly performed manually, imputations)
(e) Preparation of the final micro-data files for dissemination and later use

6. The statistical process in the 2002 Census introduced several novelties as a consequence of the technological development and introduction of a new approach to enumeration (combined method using administrative data and field enumeration). The main changes in the data processing stage were:

(a) No more manual editing (exclusively automated coding, optical reading of questionnaires and verification of data done simultaneously)
(b) Most consistency checks automated, for the first time complete imputation for some variables
(c) Use of digitalized questionnaires (paperless statistical processing)

II. Characteristics of the statistical process in the 2011 register-based census

7. The main methodological features that had to be taken into consideration before setting up the whole system will be shortly presented.

A. Gradual data processing

8. All foreseen data sources are not available at the same time, so the integration of the input data had to be adapted to timeliness of the sources. To ensure the adequate quality of input data, some keepers of the sources need more time depending on the updating of data in the registers or the method of data collection and editing. A data source could be available at the reference date of the 2011 Census (1 January) as data from the Household Register are (data in this register are in fact just a cross-section in one point of time). Two most important census sources (Central Population Register and Real Estate Register) are classic administrative registers and their data are available 3 months after the reference date (according to the law 3 months is the legal period for updating the events). Many mostly statistical sources and other external databases (students, graduates, taxation data) are available in the second half of the year (6 to 9 months after the reference date).

9. Availability of data is also the basis for the dissemination plan for the 2011 Census data, which is foreseen in three phases in 2011 and also the statistical process is adjusted to this plan:

(a) Integration of input data for population, households and housing (first release of some final census data at the end of April 2011)
(b) Processing of household and family data (first release not later than 30 September 2011)
(c) All other census topics (economic and educational characteristics, migration, occupied dwellings) are processed lastly (by calendar the first release on 30 December 2011)

10. Data that were already disseminated are no longer subject to change in the later stages of the process, so special metadata tables are prepared to ensure tracking of changes in the process. The aim is that the last status of individual record is copied into the final census database.

B. Administrative and statistical concepts

11. At the beginning of the statistical process the distinction between administrative and statistical concepts had to be recognized and the data integration methodologically follows the differences. The administrative registers are maintained on the legal basis but the census methodology is in many cases more advanced.

12. A typical example of conceptual "non-quality" is the definition of a household, which is by the Household Register possible only at the permanent address. The consequence is that for the foreigners with temporary residence in Slovenia these data (identification of household, household members and relations between them) do not exist at all. Because of that, the complete statistical process of deriving household and family data (by the way this is also the most demanding, complex and time consuming process in the 2002 Census) had to be performed in a special sub-process.

C. The same topics in several sources

13. Data for some census topics are available in more than one source, so the primary goal was first of all to detect the quality and relevance of data in sources which could be very heterogeneous. For example, the data for the topic education were foreseen in 11 different sources, very similar also for the activity status data. In such cases direct data integration was not possible, so auxiliary tables were formatted to process data and the final data only were transferred to the main process tables.

III. Statistical process – main principles

14. When designing the statistical process, we followed two main demands: traceability and repeatability. In other words, all the changes in the data performed during the statistical process must be transparently and clearly recorded. Such a transparent record of all the changes is also needed for the purposes of quality report preparation.

15. The structure and the content of the quality report is determined by the European implementing regulation, which should come into force very soon. The quality report will contain the qualitative and quantitative information, describing all the well known quality dimensions of the quality assessment model, accepted in the European Statistical System. The clear record of the data changes will be of special importance when the quantitative part (also called the quality indicators) of the quality report is prepared. Correct calculation of the indicators such as the editing rate or the imputation rate is feasible only with such an evident system.

16. To enable the above mentioned demands, the following rules are always followed through the process:

(a) For each record, for which any data is changed in the particular part of the process, a new version of the record is created and inserted into the database. The different versions of the record are designated by the values of the special variable, also called the status of the record. The initial record version hence gets status 1, the record for which the data have been changed through the manual editing process status 2, etc.

(b) Each table with the data in the database has a copy-table which contains the so called status of the variable. This status contains information about the data collection method and the information whether the data were corrected through the statistical process or not. The values of the status are assigned according to the standard 4-digit classification used at SORS. The simple rule is: if the data are changed, the associated status codes are also changed in accordance with the code list. The status is in fact a kind of metadata which should enable a better overview over the statistical process and also enable easy calculation of the process quality indicators.

IV. Statistical process – main steps

A. Input databases loading

17. Approximately 20 different sources stored in almost 30 databases are being prepared. The data stage area and the data model are based on the Oracle 10g database, the steps of filling are developed with Oracle tool SQL Loader and programming language PL / SQL.

18. Most of the data are already used for statistical analysis and are already present in the Oracle database. In order to protect personal data, it is necessary for all sources to be converted from the code of the Personal Identity Number (EMSO) to the Statistical Identifier (SID). The conversion process requires that consultants prepare the data as a txt file. As a consequence, almost all the resources which are copied to the database using Oracle SQL Loader appear in this format. For data of the Real Estate Register, which also contains a list of Owners, the data from the 2002 Census and the data on income of the population, which are already SID based in the spreadsheets in the Oracle database, the database PL/SQL procedures are used.

B. Data integration

19. Three main Oracle tables (Buildings, Dwellings, Population) and auxiliary Population Household table are set up. Besides that, metadata tables for every basic table are formed. The most comprehensive process here is the intersection of population, household and dwelling data and transfer of data between tables.

20. Integration of resources takes place in several consecutive steps, which are developed in Oracle PL/SQL Procedures as advised by methodological guidelines. In the first phase we prepare two main sets of spreadsheets:

(a) For each record for which any data are changed in the particular part of the process, a new version of Population is prepared where we combine residents and other persons who are not residents, but residing in Slovenia, the latter being used only during the preparation of data for the purpose of determining the density of housing. These records are the basis on which we can link all other sources. The first auxiliary source that we connect to the spreadsheet of Population is the spreadsheet of Households.

(b) Data of the Real Estate Register:

- Buildings,
- Parts of buildings (apartments, offices, auxiliary spaces) and
- Owners

21. The Real Estate Register, administrated by the Surveying and Mapping Authority of the Republic of Slovenia, and the auxiliary spreadsheet of Households, administrated by the Ministry of the Interior, are new sources in statistical surveys and are not yet in use and are still at the stage of development.

22. Due to the above said, both sources are still missing a lot of information; some of it is even incorrect. In the absence of a unique identifier to connect people with apartments and given the fact that often the existing data on the address and apartment number are not compatible, we needed to develop a series of algorithms, which have been included in the integration process in order to come to a solution for density of population or households and apartments.

23. With these procedures, based on existing data from other records, by virtue of ownership of apartments and additional resources we supplement the missing information on addresses, apartment numbers and households, which enables us to connect residents with apartments or parts of buildings. Additionally, in the process of integration of resources, we carry out a series of variables either to help us in data linking or to control the sums or to help us in the further process.

24. The second phase of the integration of resources is:

(a) Activity of residents

(b) Education

(c) Migrations

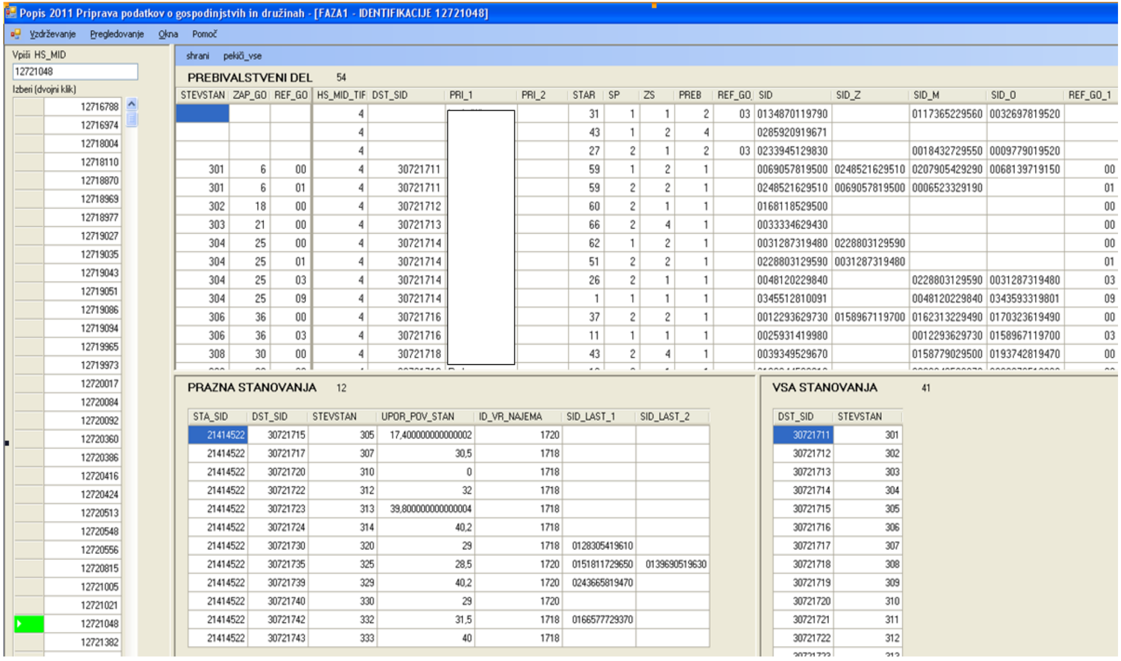
(d) Number of live births, where, for each of these groups, data from multiple sources (up to 11) are combined in an auxiliary spreadsheet. According to the prescribed methodology, we take data and the data source from which the information was taken, and integrate both data into the spreadsheet of Population

C. Data corrections

25. Corrections of data in the census database will be performed in two different ways. For the main Oracle tables (Buildings, Dwellings, Population) only automated corrections are foreseen, but for the auxiliary Population Household table also manual editing will be used.

26. At the first stage methodological rules for automated corrections of identifications (sequel number of dwelling and household in the building) and relation to reference person of the household will be applied in auxiliary Population Household table. At the second stage the inconsistent data will be corrected manually by using the custom made graphical interface, which will enable easier management with large amount of the data. Manual editing will be used only for the households and family data with the main purpose to improve the quality of output data. Two main problems had to be solved by manual editing: connecting children to their parents in case of foreigner's households (mostly missing identifications of households) and family formation in household with several members where automation is not rational (manual coding of family status). The interface which also includes the surnames allows correction of very few strictly approved variables.

Figure 1: Part of the interface for manual corrections


Popis 2011 Priprava podatkov o gospodinjstvih in družinah - [FAZA1 - IDENTIFIKACIJE 12721048]
Vzdrževanje
Pregledovanje
Okna
Pomoč
shrani
pekiči_vse

Vpiši HS_MID

12721048

Izberi (dvojni klik)

12716788
12716974
12718004
12718110
12718870
12718969
12718977
12719027
12719035
12719043
12719051
12719086
12719094
12719965
12719973
12720017
12720084
12720092
12720360
12720386
12720416
12720424
12720513
12720548
12720556
12720815
12721005
12721021
12721048
12721382

PREBIVALSTVENI DEL 54

STEVSTAN	ZAP_GO	REF_GO	HS_MID_TIF	DST_SID	PRI_1	PRI_2	STAR	SP	ZS	PREB	REF_GO	SID	SID_Z	SID_M	SID_O	REF_GO_1
			4		[redacted]		31	1	1	2	03	0134870119790		0117365229560	0032697819520	
			4		[redacted]		43	1	2	4		0285920919671				
			4		[redacted]		27	2	1	2	03	0233945129830		0018432729550	0009779019520	
301	6	00	4	30721711	[redacted]		59	1	2	1		0069057819500	0248521629510	0207905429290	0068139719150	00
301	6	01	4	30721711	[redacted]		59	2	2	1		0248521629510	0069057819500	0006523329190		01
302	18	00	4	30721712	[redacted]		60	2	1	1		0168118529500				00
303	21	00	4	30721713	[redacted]		66	2	4	1		0033334629430				00
304	25	00	4	30721714	[redacted]		62	1	2	1		0031287319480	0228803129590			00
304	25	01	4	30721714	[redacted]		51	2	2	1		0228803129590	0031287319480			01
304	25	03	4	30721714	[redacted]		26	2	1	1		0048120229840		0228803129590	0031287319480	03
304	25	09	4	30721714	[redacted]		1	1	1	1		0345512810091		0048120229840	0343593319801	09
306	36	00	4	30721716	[redacted]		37	2	2	1		0012293629730	0158967119700	0162313229490	0170323619490	00
306	36	03	4	30721716	[redacted]		11	1	1	1		0025931419980		0012293629730	0158967119700	03
308	30	00	4	30721718	[redacted]		43	2	4	1		0039349529670		0158779029500	0193742819470	00

PRAZNA STANOVANJA 12

STA_SID	DST_SID	STEVSTAN	UPOR_POV_STAN	ID_VR_NAJEMA	SID_LAST_1	SID_LAST_2
21414522	30721715	305	17,400000000000002	1720		
21414522	30721717	307	30,5	1718		
21414522	30721720	310	0	1718		
21414522	30721722	312	32	1718		
21414522	30721723	313	39,800000000000004	1718		
21414522	30721724	314	40,2	1718		
21414522	30721730	320	29	1718	0128305419610	
21414522	30721735	325	28,5	1720	0151811729650	0139690519630
21414522	30721739	329	40,2	1720	0243665819470	
21414522	30721740	330	29	1720		
21414522	30721742	332	31,5	1718	0166577729370	
21414522	30721743	333	40	1718		

VSA STANOVANJA 41

DST_SID	STEVSTAN
30721711	301
30721712	302
30721713	303
30721714	304
30721715	305
30721716	306
30721717	307
30721718	308
30721719	309
30721720	310
30721721	311
30721722	312

27. The automated data corrections in main tables will also be performed in two consecutive steps. For the first step the custom-made procedure, directly using SQL language in the ORACLE database, will be used. Here predominantly those corrections where several records are processed for one correction are performed. In the second step the simple micro-data corrections where rules refer only to one individual record will be done. For this purpose the generic metadata driven application will be used. The application was already described in the paper for the "Neuchatel working session" (Seljak, 2009). Here we only present an example of the record in the metadata table which is used to provide the application of the information about what should be corrected and how it should be corrected.

Figure 2: Metadata table for systematic data correction

Variable	Table	Condition	Condition_Table	New_Value	Step	Comment

AKT	PREB	IF AKT =" AND PODL IN (5, 19)	PREB	"02"	1	xxxx

D. Missing data imputation

28. For small parts of the target population data in administrative and statistical sources are not available at all. For these missing parts data imputation methods will be used in order to complete the micro-data files. For some cases the logical imputations, based on the rules derived from other already known values of connected topics, will be used (e. g. imputation of data on first residence in case of missing data on last migration, activity status could be directly derived on the basis of age and health insurance validity, household identification determined on the basis of existing dwelling number or family relation as parents–children or spouses). For other missing values mostly donor-based imputation methods will be used.

29. Also in the case of imputation procedures, the general metadata driven application will be used. Compared with the application described in (Seljak, 2009), we now added several additional imputation methods and also the system for checking of completeness and consistency of metadata provided. In fact, due to the performance problems, the general application has to be slightly adapted for the census purposes. Because especially donor- based procedures required too much processing time, we had to perform additional optimization. So, we in fact prepared a special “census-version” of the application.

IV. Statistical process – schematic presentation

30. Here we provide a complete overview of the statistical process presented in the below graphical scheme. For better understanding of the scheme, we provide some additional explanations.

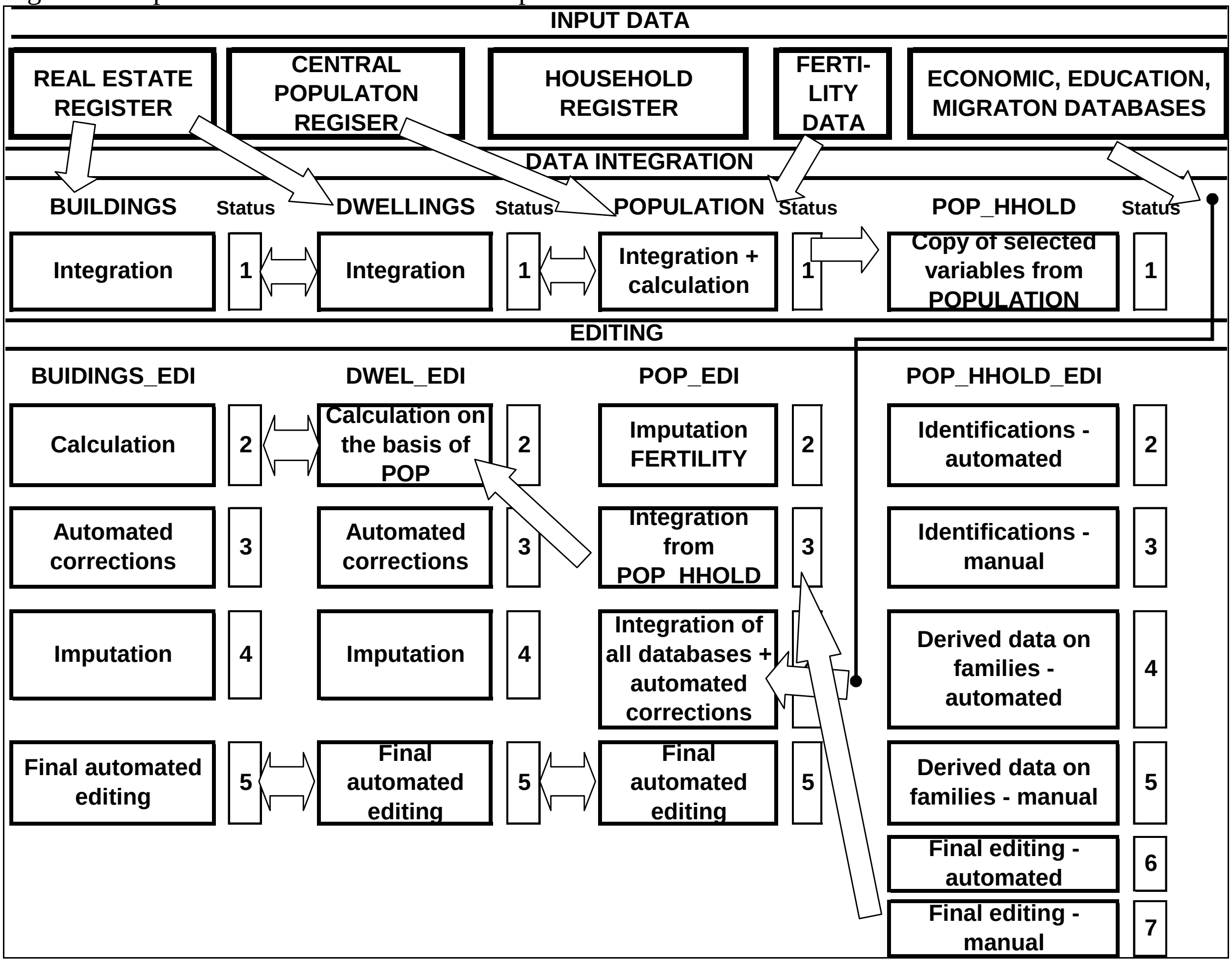
(a) Tables in the database can be roughly divided into two parts:

- Tables with the input data and tables with the data after the integration process. All these tables will be created at the beginning of the process and the data from these tables will not be changed through the statistical process.
- Tables created for the purposes of the data editing and imputation process. All these tables have the same structure as the tables from the first part, but the data in these tables will be changed constantly through the process. Each record for which at least one data item will be changed in the process will be written in these tables. Tables are named by taking the table name of the input table and adding the suffix _EDI.

(b) For each particular process, a new version of the changed record is created. These versions are in the database denoted with the variable *status of the record*. The records in the input table have status 1, changed after the first process status 2, etc.

(c) Each table with the data has a “shadow-table” with the statuses of the variables. These statuses are changed every time the data have been changed. In addition to the information that the data have been changed, these statuses also contain information how the change has been performed. These tables are not presented in the scheme.

Figure 3 Simplified scheme of the statistical process



References

1. Banff Support Team: Functional Description of the Banff System for Edit and Imputation System, Statistics Canada, Quality Assurance and Generalized Systems Section Technical Report
2. Dolenc, D. (2003), “Register based 2002 Census of Population, Households and Housing in Slovenia and New Solutions in Data Processing”, paper presented at the Joint ECE-EUROSTAT Work Session on Population and Housing Censuses, Ohrid, The Former Yugoslav Republic of Macedonia.
3. Dolenc, D. (2009), “Register based Census 2011 - a New Challenge for the Slovenian National Statistics”, paper presented at the 19th Statistical days 2009, Radenci, Slovenia
4. Register-based census 2011 - new achievement of Slovenian statistics. Statistical Office of the Republic of Slovenia.(http://www.stat.si/eng/novica_prikazi.aspx?id=3669)
5. Seljak, R. (2009), “New Application for the Slovenian EU-SILC Data Editing”, paper presented at the UNECE Work Session on Statistical Data Editing, Switzerland (Neuchâtel).