WP. 40
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Oslo, Norway, 24-26 September 2012)

Topic (vii): Editing and imputation of census data

Editing of Multiple Source Data in the Case of the Slovenian Agricultural Census 2010

Prepared by Aleš Krajnc[1], Rudi Seljak[2], Statistical Office of the Republic of Slovenia

I. Introduction

1. The Agricultural Census (AC) is a statistical survey used to collect exhaustive information on all the agricultural holdings (AH) in the county, which fulfil the certain criteria stated in the EU regulation. Observation units are hence agricultural holdings (family farms or agricultural enterprises) satisfying the criteria of the EU comparable threshold. The Agricultural Census was conducted as a Farm Structure Survey (FSS), which is one of the basic statistical surveys in the field of agriculture statistics. In accordance with the EU regulation it is conducted as a census every 10 years, while between censuses it can be conducted as a sample survey. In the period between two agricultural censuses, AC2000 and AC2010, sample surveys were carried out in 2003, 2005 and 2007. For the future, sample surveys are foreseen to be carried out in 2013, 2016 and later on again as a census in 2020.

2. The main goal of the AC is to obtain reliable data (at regular intervals) that cover certain topics of agricultural activities and are obligatory for all Member States (EU 27). These data cover information on:

(a) location of production
(b) organic farming
(c) land use
(d) livestock breeding
(e) labour force involved in the farm work
(f) other gainful activities on agricultural holdings
(g) rural development measures

3. The Agricultural Census was conducted in European Union (EU) Member States, Norway, Switzerland, Croatia and Montenegro in 2010, only some of them collected their data in 2009 (Greece, Portugal and Spain). The AC is based on Council and EP Regulation (EEC) No 1166/2008[3]. The main goal of applying the obligatory regulation is to get for the first time the comparable data on agricultural indicators based on the same methodology where all 27 countries followed the same rules and guidelines.

[1] Aleš Krajnc, Statistical Office of the Republic of Slovenia, Litostrojska 54, 1000 Ljubljana, Slovenia, rudi.seljak@gov.si, phone: +386 1
[2] Rudi Seljak, Statistical Office of the Republic of Slovenia, Litostrojska 54, 1000 Ljubljana, Slovenia, rudi.seljak@gov.si, phone: +386 1 24 15 294
[3] Regulation (EC) No 1166/2008 of the European Parliament and of the Council of 19 November 2008 on farm structure surveys and the survey on agricultural production methods and repealing Council Regulation (EEC) No 571/88

4. The Slovenian AC2010 was carried out by the Statistical Office of the Republic of Slovenia (SORS) in June-July 2010. One part of the data was collected with the field survey carried out between 1 June and 15 July 2010 when the interviewers surveyed around 95,000 agricultural holdings by using the CAPI collection mode. Additionally, a large part of the required data was obtained from different administrative sources. If such usage of multiple sources certainly brings many advantages in the collection phase, it can also bring some disadvantages, especially for the data editing phase. In the first part of the paper we will shortly describe the main features of the Slovenian AC2010 and editing procedures and applications used in the different stages of the statistical process. Then we will focus on the discussion what the usage of the multiple sources means from the editing point of view. We will point out the main advantages and disadvantages of such an approach an draw some conclusions at the end.

II. Slovenian AC2010

5. From the general point of view the Slovenian AC2010 certainly met its goals and purposes. As the consequence of good preparatory work, farmers were mostly well prepared for the census, they took it very seriously and thus greatly facilitated the work of fieldwork interviewers. The field data collection was in fact carried out by the external contractor, who hired its own interviewers to collect the necessary information in the field, while SORS prepared the methodological instructions and the required quality standards for them. There were 94,686 agricultural holdings visited in the field and after all the abandoned agricultural holdings and holdings under the threshold (by "European comparable agricultural criteria" - ECA) were removed, there were 74.646 AH left. ECA criteria is different in each country and removes the smallest agricultural holdings which together contribute 2% or less of the total utilised agricultural area excluding common land and 2 % or less of the total number of farm livestock units (see Regulation 1166/2008).

6. European comparable agricultural holdings in Slovenia are those having:
   (a) at least one hectare of utilised agricultural area, or
   (b) less than 1 hectare of utilised agricultural area, but:
      - at least 0.1 hectare of utilised agricultural area and 0.9 hectare of forest, or
      - at least 0.3 hectares of vineyards and/or orchards, or
      - two or more livestock units (LSU), or
      - 0.15 to 0.3 hectare of vineyards/orchards and 1 or 2 LSU, or
      - more than 50 beehives, or
      - are market producers of vegetables, herbs, strawberries, mushrooms, flowers or ornamental plants.

7. The "field part" of the AC2010 was conducted by using the computer-assisted personal interviewing (CAPI). The data entry program was made by the outsourced company, but all the instructions and rules were provided by SORS's staff. Fieldwork was conducted by about 600 interviewers, started on 1 June and finished on 15 July. After the completion of the fieldwork, telephone interviewing of some agricultural holdings continued until 25 July. The purpose of telephone interviewing was to check the correctness of entered data. In this way we checked the work done by fieldwork interviewers and the correctness of data entered into the computer application. All the remaining data editing work was done later when the field surveys were integrated with the administrative data and was performed inside SORS using a SAS-based application, which will be described later.

8. In the AC 2010 data from administrative sources were used to a greater extent than in the previous surveys. Some data from the statistical survey were supplemented with the data from the following administrative sources:

   (a) Register of Agricultural Holdings at the Ministry of Agriculture and the Environment (MAE). This register includes:
      - general data on agricultural holdings – address, location, etc.
      - data on holder and other family members

- data on permanent crops – intensive orchard plantations, olive groves, area with hops, vineyard producers, fruit producers in extensive and/or meadow orchards
- data on common land

(b) Subsidies applications at the Agency of the Republic of Slovenia for Agricultural Markets and Rural Development (AAMRD) – data on crops on arable land
(c) Data on Support for Rural Development (AAMRD)
(d) Register of Vineyards (MAE)
(e) Organic Farming Register (MAE)
(f) Register of Supplementary Activities (MAE) – data on gainful activities on agricultural holdings
(g) Register of Bovine Animals (MAE)
(h) Register of Beehives (MAE)
(i) Register of Sheep and Goats (MAE)
(j) Register of Pigs (MAE)

Some administrative sources were complete (questions were not asked on the field) and some were used for complementing the field data (pre-prepared filter questions in the questionnaire).

III. Statistical data processing in the AC2010

A. Organization of the data

9. The first challenge in the data processing phase was the integration of the data from different sources. Key for data linkage of field survey data and the different administrative sources was unique identification number (ID) of agricultural holding established by MAE. Also each agricultural holding in the Statistical Farm Register (SFR), which is the basis for the field survey, also has the same ID number. Preparation of the SFR took some time to refresh-update with all possible sources and to remove all duplicates to get a clear list of agricultural holdings, but nevertheless the data linkage of different administrative sources was quite an undemanding job and we did not encounter any major problems at this stage.

10. The data were organized in the ORACLE database. To facilitate maintenance of the data, we created quite a lot of tables. For instance, each of the different administrative sources was put in the separate table and also the field data were separated into the different tables according to the sets of related questions. What follows is the summary of the different objects (tables/views) in our database.

(a) Pre-prepared list of agricultural holdings (2 tables). The first table is the initial list of agricultural holdings prepared prior to the fieldwork. The second table consists of the personal data updated with the information from the field.
(b) The "raw" data[4] of information collected in the field (14 tables)
(c) "Raw" administrative data (14 tables)
(d) Tables with derived variables (5 tables). Two tables contain derived size classes, two tables contain derived variables connected to labour force data and one table is the so-called EUROFARM table, which is sent to Eurostat as the final microdata.
(e) Tables with code list (24 tables)
(f) Tables with the records for which any data have been changed during the process (28 tables). Each "raw data" table has one associated table, the so-called "EDI table" where all the records changed during the editing process are inserted.
(g) Tables with statuses of variables. For the raw data, the status shows how the variable was obtained (field data, administrative data, and derived data). For the edited data, the status shows

[4] The raw data for the statistical processing. As mentioned before, some data were at this stage already corrected according to the telephone re-contact.

in which step of the process the data was changed and what method was used to perform the change.

(h) Views where the data were arranged to be queried by the final user (56 views). The first set of views shows all the versions of the record (raw and all corrections) in one table. The second view shows for each record the latest version from the process. These views show the users the data in the form prepared for the final tabulation.

11. All together we had 199 tables and views in our database and all together 9,583 variables in these tables and views. The large amount of these objects, together with the large amount of the records in the database made the data processing a demanding and challenging task requiring thoughtful and precise planning and implementation.

B. Application for statistical data processing

12. The statistical data processing was performed by using general metadata driven application. The application was already described in papers presented at the previous sessions (see Seljak, 2009 and Seljak at al., 2011), hence we here give just the rough overview of the main characteristics:

(a) Application is metadata driven, meaning that the "core part" is generic and can be used for a wide range of different surveys. The parameterization for the particular survey is then always given through the specific set of "processing metadata", provided outside the programming code. It means that when the application is used for different surveys there is no need to change the core SAS code.

(b) The core part of the application, which performs the data processing, is built in SAS (with additional Banff procedures), while the processing metadata are provided through the MS Access tables.

(c) The creation of the metadata, which determines the editing process, is predominantly in the hands of subject matter specialists. IT and general methodology specialists only provide occasional support.

(d) The execution of the particular parts of the process is also carried out through the SAS system. For this purpose the stored-process user interfaces, which are part of the SAS Enterprise Guide, are used.

(e) The application consists of three "building blocks", each of them taking care of one of the following processes:
- **Logical checks.** Any version of data could be checked by using the set of edit rules. The set of edit rules is stored in the metadata database in the form which is recognizable for the SAS program code.
- **Deterministic corrections.** Data are corrected by providing the corrected values for the particular unit and the particular variable (individual corrections) or by providing the deterministic rules (e.g. IF X<Y then X=Y) by which the variables for the particular set of units are corrected. Corrections can be performed in several consecutive steps, where each can take into account already corrected data from the previous step.
- **Imputations.** More than 20 methods are on disposal to create the estimated values which replace the missing or incorrect data values. Also here, the process could be performed in several consecutive steps.

(f) The use of this generic application ensures two very important features: traceability and repeatability:
- Traceability means that all the changes of the data performed during the statistical process are transparently and clearly recorded.
- Repeatability means that each repetition of the process, under the assumption of unchanged input data, will result in the same output data.

(g) To enable the above mentioned features, the application follows the following rules:
- If any data in any record are changed during the process, the changed data never overwrite the original, but 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.

- For each variable, which could be a subject of change, the special variable, called also the status of the variable, exists. This status contains information about whether the data were corrected through the statistical process and in case of change, which method was used to perform the change. The values of the status are assigned according to the standard 4-digit classification used at SORS. A simple rule is: if the data change, the associated status codes also change in accordance with the given code list.

C. Specific features of the statistical processing in the AC2010

13. Although from the general point of view the AC2010 could be treated as any other statistical survey, it also has several specific features that make the statistical process a bit more complicated than in the case of most of the "regular" surveys. The whole statistical process consists of the following steps, each of them created the new version of the processed data:

(a) **Insertion of the new units into the database**. For some units that were according to the field survey data not agricultural holdings and were therefore not included in the database, the administrative records indicated that they in fact are agricultural holdings. These units were then inserted into the database with the available administrative data. The missing data were imputed later in the process. For this part of the process, the custom made SAS program was created and integrated into the general application.

(b) **Replacement of the whole set of the field data.** For some units for which the initial data checking showed that they are really of a bad quality, the repeated telephone interview was performed and then the record with the newly gathered data was created. Also for this part of the process, the custom made SAS program was created.

(c) **Individual and systematic data corrections.** Corrections of the data for which the set of predefined logical checks indicated that they are erroneous. Most of these corrections were performed by using the systematic deterministic rules, which corrected the whole set of the units which failed a certain logical check. In some cases, corrections had to be performed on the individual level, aiming at correcting particular data for the particular unit. This part of the process was carried out by using the general application for automatic data editing, which was complemented with a few functionalities. At this stage most of the records were supplemented/complemented with administrative sources.

(d) **Data imputation.** Missing and in some cases also inconsistent data were estimated by using the set of the imputation methods. The general application metadata driven application was used to carry out this part of the process. Most of the imputation methods that were required were already included in the general application. Only few additional parameterizations of the already existing methods were needed to fulfil the needs of the AC2010 imputation process. The following two (groups of) methods were by far the most used methods:

- **Hot deck imputation.** Different versions of the donor based imputations were used for most of the categorical and also for few numerical variables. Most of the hot deck imputations were performed in a way that the same donor was chosen for the whole set of the related missing variables. Many versions of the hot deck method were created, each of them based on different imputation cells and on the different matching variable(s).
- **Structure imputation.** Usage of the two different sources many times led to the situation when we had data for a variable which could be expressed as a sum of other variables, but the data for these summands were partly or completely missing. In such cases we had to estimate the structure (shares of each of the missing summand) of the sum by using the imputation procedure. There are several versions of the structure imputation method available in the general imputations, but in most of our cases we used the version where the shares are taken from the properly selected donor. To enable consistency of the final sum, the same donor, of course, had to be taken for all the missing summands. Even so, the rounding of the summands causes that the imputed

values did not sum up to the desirable value. In such cases additional adjustments were needed to ensure the consistency.

(e) **Final corrections.** Because of the huge number of tables, variables and consequently large numbers of logical checks, corrections and imputations, it is difficult to enable the "cleanliness" of all of the processed data. Therefore, we added another step in the process where we dealt with all the remaining inconsistencies and eventually still missing data. Also for this step the general application was used.

14. As already mentioned, the complexity of the data processing in the case of the AC2010 required that a few additional functionalities which have not been on disposal before had to be added to the general application. One example of such functionality is the possibility to dynamically join the tables in the process[5]. If for instance, for imputation of the variables in the table A, we also need the data from tables B, C and D, the metadata record also contained the following information:

Process	Table	Associated_Tables
Imputations	A	B*C*D

15. One of the features of the AC2010 data is also the fact that it refers to two different unit levels. For instance, the first one is the level of the agricultural holding and the other (sub-level) is the level of the persons working on the agricultural holding. This fact made the whole process even more demanding. We had to make a special procedure to the general application, which enabled calculation of derived variable on the level of the holding. We called these variables "aggregated derived variables". For illustration, we present (part of) the metadata table, where the rule for the calculation of one of the aggregated derived variables is given:

TABLE	Der_Var	AGR_ID	Expression	AGR
G_D	DELO_GOZD_G	ZAP_ST	Gx_7*WORKING_HOUR*(Gx_10*0.01)	SUM

Table	- name of the table with all the variables needed
Der_Var	- name of the derived variable
AGR_ID	- identification variable, which determines the unit of aggregation; in this case ZAP_ST denotes agricultural holdings, while data in table G_D are given at the level of each person
Expression	- expression which will be aggregated to a higher level
AGR	- Type of aggregation (e.g. sum, mean, count, etc.)

16. Special types of the derived variables were the size classes. A large number of different size classes had to be derived in the case of AC2010 data processing and many times their derivation required several variables from several different tables. Therefore, we created a special metadata driven process for the calculation of these variables. In contrast to the other derived variables, where metadata rules for the calculation could be quite long and complex, here only the lower limits of the size classes had to be provided. This facilitated the whole procedure significantly and made the process of adding new size classes quite straightforward. Once the metadata rules for the calculations of the size classes were given, they could be derived for any version of the data and automatically put into the database.

IV. Main challenges encountered

17. The Agriculture Census is already by its nature a very complex and demanding survey. Especially its exhaustiveness, resulting in large amount of gathered data, makes the data processing a very complex task. The combination of different data sources, as it was in the case of the Slovenian AC2010, makes the job even more challenging. What follows is the summary of the main challenges in

[5] This functionality was added mainly because of the large number of tables that had to be managed through the process.

our statistical process as seen now, when the job has been more or less done and some kind of "healthy analytical distance" already exists.

18. A large number of tables, variables, relations between variables and consequently a large number of edits and constraints were the main factors of complexity. Since the amount of the data editing work was so large, it had to be spread among several subject matter specialists, each of them covering one of the areas. In addition to methodologists, a person with "database management skills" prepared the database and adjusted it through the whole process of data editing, hence all new specifics in data editing influenced the build-up of the database. Also adjustment of application for statistical data editing had to be made during the whole process due to new and special cases. Each person in the process knew his work, but it is also important to have one person that is in charge of the overall coordination and guidance of the whole process, who knows how to integrate each part into the whole.

19. A large number of variables and relations among them caused that each part of the process consisted of many different steps. This was especially true in the case of the imputations, when the "filling up" of the data has to be carefully designed in consecutive steps. For instance, many times a certain variable which was imputed in one step served as the matching variable in the next step. Or the "sum variable" was imputed first and the summands were estimated in the next steps. So, for one of the tables the imputation process was performed in as many as 16 consecutive steps. Managing of the process consisting of so many steps was sometimes also very a demanding task.

20. The special issue was the integration of two different data sources. Although such practice certainly brings many advantages, especially in the area of response burden and costs reduction, we have to have in mind that it also brings some disadvantages, especially the significantly increased workload at the data editing stage. For some variables we had more than one source; this led to a quandary which source to use, hence all of them were of equal quality. There was an example for variables on organic farming, for which we had special administrative data, and beside that we also had administrative data on the whole area of utilised agricultural land, used by agricultural holding. For some of them we also got data from the field (even though they reported the area to administrative services, they reported the area also in the AC2010). So for some agricultural holdings we had three different sources, and of course not many of them had the same figures for the same agricultural holding. At that stage we had to decide which source to use and discover why there are differences (critical date of reporting, wrong entry of the data, mistake of the farmer, etc.). This kind of work took quite a lot of time, since there were many different situations to resolve. If we would have only field data survey, there would be no additional work behind it, because we would not have any other sources to compare the data with. Certainly if many source data are used, the quality of the data on individual agricultural holding increases.

21. As described in the article "*Use of Administrative Data Sources in the 2010 Agriculture Census" (Kutin B.2009),* we managed to reduce the reporting burden at variable level by about 14% because of the use of administrative data and at the level of agricultural holdings on average by about 11%. The reduction of the response burden did not have a big influence on the total cost of the AC2010, because there was still field data survey and we know that such surveys are the most expensive ones. If a different type of data collection was used (telephone, through postage, etc.), then also the usage of administrative sources would have a greater impact on the total cost of the AC.

IV. Conclusions

22. Agriculture census is by the definition exhaustive and complex survey, requiring a lot of financial as well as human resources. On the other hand it is also quite burdensome for the reporting units, since there is a large amount of information which is required to be provided by each of the observed units. One way to decrease the large cost and burden factor is the usage of the already available administrative sources. Such usage leads to the so called "statistical surveys with combined data sources" and these types of surveys are more and more popular in the modern official statistics.

23. In the case of Slovenian AC2010 large amount of the data were gathered from the different administrative sources. In the first part of the paper we described the main characteristics of the AC2010, focusing mostly on those that influenced our data editing process. In the second part we then described the data editing process and discussed it main problems and challenges.

24. The main features of the AC2010 editing process determined the fact that we had to deal with large amount of tables and variables, coming from several different sources. In the implementation phase this fact resulted in the process with many different parts of the process, each of them consisted with many different steps. We tried to use the already existing applications whenever it was possible, but for some parts of the process we still had to create a few custom made computer procedures. For the already existing generic application we took advantage of demands coming out from the complexity of the process to perform some upgrades and improvements.

25. The usage of administrative data undoubtedly decreased the response burden and (to a lesser extent) also decreased the costs of the survey. It is also undoubtedly true that the usage of administrative data significantly increased the amount of editing work that had to be performed inside the office and that the whole processing would be finished a lot earlier if all the data would be collected in the field. But on the other hand, we also strongly believe that the usage of administrative data and the editing work, performed due to the usage of administrative data, significantly improved the quality of our data, especially their accuracy and reliability.

26. One of the key approaches in the modern data editing practice is the approach of selective data editing, when we concentrate only on errors which could influence our statistical results. With such an approach we can undoubtedly reduce the data editing costs and significantly improve its efficiency. This approach becomes somehow questionable in the case of the AC2010 (as in the case of more and more surveys), due to the fact that also microdata have to be provided to Eurostat as well as these data are given on disposal to different researchers. Can we still afford to leave some inconsistencies in the data, even if we know that they could not significantly influence the statistical results?

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., Krek, M., Seljak, R. (2011), “Editing Process in the Case of Slovenian Register-based Census”, paper presented at the UNECE Work Session on Statistical Data Editing, Slovenia (Ljubljana).

3. Kutin B., Krajnc A., Stele A. (2009) , “Use of administrative data sources in the 2010 agriculture census”, paper presented at the Statistical days, Slovenia (Radenci).

4. 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).