Distr.
GENERAL

Working Paper
14 February 2014

ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE (ECE)
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION
AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

UNITED NATIONS
ECONOMIC AND SOCIAL COMMISSION
FOR ASIA AND THE PACIFIC (ESCAP)

Meeting on the Management of Statistical Information Systems (MSIS 2014)
(Dublin, Ireland and Manila, Philippines 14-16 April 2014)

Topic (i): How IT can contribute to changing organizational culture

Quality assurance -Population and Housing Census

Prepared by Alma Kondi, INSTAT, Albania

I. Introduction

Data collected by Population and Housing Census in Albania, provide information on people's age, sex, disabilities, level of education and conditions under which the citizens live. For this reasons census data facilitate the process of planning, management and evaluation of policy-making.
The main objective of this quality assurance is to provide valuable information for producer and the user of the data regarding the quality of census data. This paper describes all steps used before, during and after the census enumeration in order to increase data reliability.
The multivariate nature of Population and Housing Censuses may arise errors at any stage in the census operation. Census errors can occur because of many reasons such as poor planning, varying interpretation of questions by enumerators and/or respondents, data capture errors through scanning process.
Census accuracy is mainly measured by the errors found in the raw data and the editing and imputation process they went through. Accuracy is also measured with derived methodology by comparing census data with Post Enumeration Survey data. Using this method we evaluate the quality of Population and Housing Census in Albania by core demographic variable (age, sex, marital status). This analysis evaluate the quality of census data that are carried out by age groups, sex and marital status for all census topics included in the analysis to detect eventual inconsistencies of data.

II. Measures taken to ensure quality data collection

Before the census took place, a set of measures have been developed and introduced by INSTAT (Institute of Statistics) in order to guarantee the quality of the most important phase of the census: gathering complete and trustful data by all Albanian inhabitants.

A. Organizational measures

The Census in Albania followed the traditional method (door to door) of universal direct enumeration based on field operations with completion of questionnaires by enumerators. Due to the complexity of some questions and in order to achieve harmonized and coherent answers, allowing questionnaires being directly filled-in by the respondents would have increased the risk of incorrect responses and would have had as a consequence a loss of data quality. INSTAT staff defined specific selection criteria for the enumerators, the controllers as well as for the supervisors. These were based on the education level, experience, knowledge of local context etc. Trainings and training modules were given particular importance. It was understood that the performance of field work and the quality of data collection was depending from the quality and care provided to the whole training procedures.
Trainings were organized in a top-down cascade form. Trainings were conducted all over Albania and were all done with the same training materials, instruments and methodology; each training centre was equipped with projectors.
Although it was proposed in the instruction manuals that the enumerators make at least three (3) trials to get in contact with the inhabitants of a household, during the trainings it was recommended, if needed, to visit the households more times in case of absence. Also, before the census start, each enumerator visited its enumeration area to get acquainted with its geographical limits and to update the existing building and dwellings existing. During this preliminary visit the enumerators were given as a task to leave at each dwelling a notification letter informing the residents on the census process. During the census, if a dwelling was found empty, enumerators were instructed to leave a note with their contact telephone number asking the eventual inhabitant to contact them and agree about the timing for the interview.

B. Census data collection forms

The questionnaire for Census was designed by INSTAT with advice from international experts. The most delicate and important element to ensure census quality has been the proper design of the questionnaire. The highest care was taken to ensure that questions were well understood, that they were clear to the interviewed person and not ambiguous.
A questionnaire test was implemented at beginning 2010 and a pilot census was conducted in April 2010 in order to test some of the main aspects of the 2011 full census operation. The analysis of the pilot census showed the need to further improve census questionnaires and the overall census preparation. Additional consultative meetings with stakeholders were organised in 2010 and 2011, and additional tests of questionnaires were conducted. The highest care was taken to ensure that questions were well understood, that they were clear to the interviewed person and not ambiguous. Structure of the questionnaire and its design were also given the utmost attention in order to facilitate an easy and correct reporting of the information provided. Questions addressed only to some persons were clearly identified, for example questions on fertility that were to be asked only to women of a certain age. Numerous examples in the various instruction manuals (provided often in illustrative form easy to understand) also participated to increasing the quality of data collection on the field.

C. Census mapping

Albania lacks administrative data such as population registers and it has no physical address system, which would facilitate the identification of locations on the field. As a consequence INSTAT, since its first census preparations in 2007, started an ambitious project to establish a Geographic Information System (GIS). For CENSUS 2011, the 12,000 enumeration areas each contain around 100 dwellings: a GIS system has been developed internally which has mapped these areas. The map base is satellite images which show buildings: The original photographs

themselves were some three years out of date at the time of the Census but were updated at buildings level in 2010 and 2011 in the field and changes reported to the GIS database. Controllers visited the areas in advance of the enumerator to mark additional buildings and delete demolished or derelict buildings. Each residential building was sequentially numbered and new buildings were also numbered following on from the last marked number.
Each enumerator then had to investigate each building to identify dwellings units: the numeric coding of each unit included the building number, the building entrance number, the level within the building and the 'door' number on each level. Within any building the allocation of these numbers was defined by a strict procedure which could be understood and replicated by any controller or supervisor revisiting the building. Integrating the best experiences of other countries, Albanian census ensured a census organization likely to ensure a data-collection of the best possible quality.

III. Localization of errors and correction of the inconsistencies

The design, development and execution of the data editing and imputation procedure for a Population and Housing Census is a very crucial task, whose aim is to guarantee the highest quality level in validated data. This implies to define all the relevant edit rules enabling to localize errors in data, and to impute both erroneous and missing values trying to restore the "true" data that were lost during the various phases of the survey.
A number of processes were performed to turn the tick and text responses on the questionnaire into data that could be edited and imputed. Figure 1 summarizes the data processing adopted for the 2011 Albanian CENSUS.

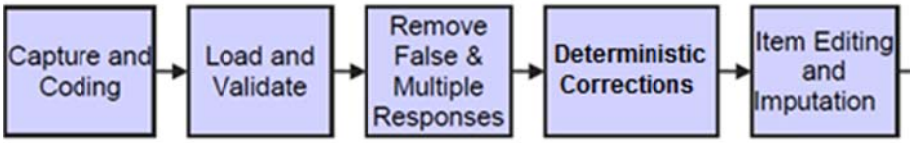

Figure 1 - Data processes for Census 2011
Firstly, at capture, questionnaires were scanned and complex coding was used to assign numerical values to written text and ticked boxes. This involved applying coding rules and standardized national coding frames, such as NACE rev2 (Statistical Classification of Economic Activities) and ISCO-88 (International Standard Classification of Occupation), which allow data to be easily compared between different sources. The data were then loaded into a dataset and validated to ensure that the values for each question were within the range specified in the relevant coding frame. Next, multiple responses and false persons were detected and removed.
Multiple responses occurred when a household completed more than one questionnaire or recorded the same person more than once, while a false person was where not enough information was recorded to identify them. Following this, deterministic corrections addressed inconsistencies in terms of errors clearly identifiable, mainly linked with the routing of the questionnaire.
The data were then ready for item editing and imputation. After it, all of the returned questionnaire records were complete and consistent.

A. The principles of data editing and imputation

The process of cleaning can be described as having two components:

- Editing, which consists in localize errors in collected data. Very seldom errors are self evident (it is only when values in a variable are out of the domain defined for that variable), to localize

them correctly it is necessary to define rules (logical, mathematical or statistical). Missing values when answers to questions in the questionnaire are due, are assimilated to errors: also for missing values, it is not always clear when they can be considered as errors.

- Imputation of the localized errors and missing values: this second level of editing is also considered as the "correction" step of the errors in the responses.

As for the definition of the edit rules, the editing rules have to be designed to detect data inconsistencies and errors, such as: i) unasked questions; ii) unrecorded answers; iii) inappropriate responses. In the case of categorical data (which is the case of mostly Population and Dwellings Census variables), edit rules are of the logical type.
Once logical edit rules have been explicitly defined by analyzing the questionnaire and also by making use of all the knowledge regarding the objects of the survey, they need to be controlled (to avoid inconsistencies and redundancies). It is also necessary to derive all edits that are implicitly contained in the initial ones: this is essential in order to correctly localize errors.
Once the errors are localized, the nearest neighbour hot-deck imputation technique will be used to correct the data. This approach guarantees that with the minimum set of changes in terms of variables corrected, the corrected data will satisfy simultaneously all the edits.
All the above operations can be executed by using the functions implemented in the software Concord (Scia module) implemented by the Italian Institute of Statistics (ISTAT), that adopted the probabilistic approach known as the Fellegi-Holt methodology. Concord is being used in ISTAT to handle "stochastic errors" (characterized by a random nature, i.e. no particular cause behind them can be detected) in all the most important current surveys regarding households and individuals.
In some cases, a different approach must be used: it is when a particular category of errors is present in data, the so called "systematic errors" that are characterized by detectable causes that produce them during the process of data collection and processing. These errors must be treated prior to the probabilistic step, by applying deterministic rules implemented in ad hoc programs.

B. Statistical evaluation

In the following table are shown the frequency of the errors detected by the edit rules applied to the Census data, before and after the deterministic corrections:

	Before deterministic correction			After deterministic correction		
	Dwellings	**Households**	**Individuals**	**Dwellings**	**Households**	**Individuals**
Number of Values	9,191,988	21,667,860	176,408,694	9,191,988	21,667,860	176,408,694
Number of Records	1,021,332	722,262	2,800,138	1,021,332	722,262	2,800,138
Number of Variables	9	30	63	9	30	63
Number of Errors	188,799	291,148	11,792,578	71,292	111,742	3,009,092
Inconsistencies	71,673	54,507	1,923,286	35,890	26,295	1,613,760
Non-responses	117,126	236,641	9,869,292	35,402	85,447	1,395,332
Number of Valid values	9,003,189	21,376,712	164,616,116	9,120,696	21,556,118	173,399,602
Error rate	2.05	1.34	6.68	0.78	0.52	1.71
%of **inconsistencies**	37.96	18.72	16.31	50.34	23.53	53.63
% of Non-response	62.04	81.28	83.69	49.66	76.47	46.37
Non-Error rate	97.95	98.66	93.32	99.22	99.48	98.29

Table 1 - Errors before and after deterministic corrections

As it is easy to understand from Table 1, the errors detected by the edit rules before and after deterministic corrections are quite different, especially for what concern the dataset of the individuals.

Indeed, in this dataset the errors before deterministic was 11,792,578 over a total of 164,616,116 total values (6.68 %) while after the deterministic correction this number reduced to 3,009,092 (1.71 %). As mentioned before, this situation underlines a relevant impact of the questionnaire design and of the accurateness of the enumerators on the number of errors in the census data. Also should be mentioned here that in a significant percentage of the errors (4.97 %) the deterministic corrections were able to correctly reconstruct the information, bypassing in this way the enumerator's mistakes.
In the next table, the impact of the deterministic corrections at the level of records is illustrated.

	Before deterministic corrections			After deterministic corrections		
No of Errors	**Dwelling Records**	**%**	**Cumulative%**	**Dwelling Records**	**%**	**Cumulative%**
0	876,469	85.82	85.82	984,669	96.41	96.41
1	113,012	11.07	96.88	15,843	1.55	97.96
2	27,060	2.65	99.53	13,899	1.36	99.32
3-5	3,445	0.34	99.87	5,583	0.55	99.87
6 +	1,346	0.13	100.00	1,338	0.13	100.00
No of Errors	**Household Records**	**%**	**Cumulative%**	**Household Records**	**%**	**Cumulative%**
0	606,546	83.98	83.98	664,944	92.06	92.06
1	78,059	10.81	94.79	26,323	3.64	95.71
2	8,884	1.23	96.02	4,201	0.58	96.29
3-5	6,371	0.88	96.90	25,523	3.53	99.82
6 +	22,402	3.10	100.00	1,271	0.18	100.00
No of Errors	**Individual Records**	**%**	**Cumulative%**	**Individual Records**	**%**	**Cumulative%**
0	1,352,952	48.32	48.32	1,642,454	58.66	58.66
1	631,195	22.54	70.86	579,597	20.70	79.36
2	327,370	11.69	82.55	265,494	9.48	88.84
3-5	335,886	12.00	94.55	192,778	6.88	95.72
6 +	152,735	5.45	100.00	119,815	4.28	100.00

Table 2 - Records by number of errors, before and after deterministic corrections

From Table 2, the impact of the deterministic corrections on the data it is quite evident: the numbers of records with zero errors pass from 85.82% to 96.41% for the dwelling records, from 83.98% to 92.06% for the household records and from 48.32% to 58.66% for the individual records. It is also interesting to underline how the number of record with errors smaller or equal than two does not change a lot before and after the deterministic corrections.

The set of indicators considered to perform the assessment of the effects of the cleaning procedure at aggregate level can be grouped in three different kinds:

- Indicators on the amount of data submitted to the imputation procedure, like Number of Records, Number of Variables, and Number of Variables subject to the Imputation procedure and Number of Total Values.
- Indicators for the evaluation of the overall effects of the imputation procedure, like Number of Valid values, Number of Imputed values, Number of Modifications, Number of Additions, Number of Eliminations, Imputation rate, Addition rate, Modification rate, Elimination rate;
- Synthetic indicators on the imputation rate by records, like for instance Number of Records with Imputation rate greater than 2% and Number of Records with Imputation rate greater than 5%.

The complete definition of these indicators is given in Table 3 are reported their values:

	Dwellings	Households	Individuals
Number of Total Values	**9,191,988**	**21,667,860**	**176,408,694**
Number of Records	1,021,332	722,262	2,800,138
Number of Variables subject to Imputation	9	30	63
Number of Imputed values	**71,292**	**111,742**	**3,009,092**
Number of modification	35,890	13,216	621,342
Number of Invalid routing	35,402	98,526	2,387,750
Number of additions	*35,402*	*85,447*	*1,395,332*
Number of eliminations	-	*13,079*	*992,418*
Number of valid values	**9,120,696**	**21,556,118**	**173,399,602**
Number of valid blank	1,341,348	13,604,626	79,588,293
Number of valid codes	7,779,348	7,951,492	93,811,309
Imputation rate	**0.78**	**0.52**	**1.71**
Non-Imputation rate	**99.22**	**99.48**	**98.29**
Modification rate	**50.34**	**11.83**	**20.65**
Addition rate	**49.66**	**76.47**	**46.37**
Elimination rate	**-**	**11.70**	**32.98**
Invalid routing Rate	**49.66**	**88.17**	**79.35**
Invalid codes Rate	**50.34**	**11.83**	**20.65**
Records with Imputation rate greater than 2%	**3.59**	**4.29**	**20.64**
Records with Imputation rate greater than 5%	**2.04**	**3.71**	**4.28**

Table 3 - Quality assessment indicators at aggregate level

From indicators shown in Table 3 it is evident the very low impact that imputation has had on observed data as it – is once and for all – well shown by the imputation rate which is always under 2% (Individuals) and frequently under 0.8% (Dwellings and Households). Even the synthetic indicators on the imputation rate by records confirm the good performance of the imputation process with a very low percentage of records having an imputation rate greater than 2% or 5% which is always under 4.3%.

C. Evaluation of the census quality based on Census/PES matched records

From a population perspective, it is interesting to analyse the differences in the reported information between the cleaned census individuals and the PES individuals that was possible to match. This kind

of analysis has been conduct for the three main variables available in both Census and PES: Sex, Civil status and Age.

	%	95% Confidence Interval	
Sex		Lower	Upper
Different	**1.3**	0.4	2.1
Equal	98.7	97.9	99.6
Civil Status			
Different	**3.8**	1.0	5.9
Equal	96.2	94.1	99.0
Age			
Different	**6.9**	1.4	9.8
Equal	93.1	90.2	98.6

Table 4 - Percentages of differences in the Census/PES matched records

As it is shown in Table 4, the estimated percentage of differences for Sex, Civil status and Age is respectively of 1.3%, 3.8% and 6.9%. It should be underlined here that not all the difference between Census and PES data should be automatically considered as a mistake since both the Census and PES operations are affected by errors. On the other hand PES was a smaller operation, the questionnaires were much easier and the enumerators were better trained. In this view, a conservative approach could be the assumption that each four differences three can be attributed to mistakes in the Census and only one to mistake in the PES. Under this assumption an indicator of the total error of the variables can be calculated as:

$$\text{TE} = \frac{3}{4}\ (\%\ \text{of differences})$$

While the total quality of the variables it is obviously expressed by:

$$\text{TQ} = 100 - \text{TE}$$

In the table below TE and TQ are calculated for the three variables object of the comparison between the Census/PES matched records.

	TE	TQ
Sex	1.0	99.0
Civil Status	2.8	97.2
Age	5.2	94.8

Table 5 - Indicators of Total error and Total quality for Sex, Civil status and Age

The results illustrated in Table 5 confirm also something that is well known by statisticians: more is complex the definition of the variable object of the study and more its quality will be affected. Indeed the lowest value of TE is in correspondence of the variable Sex (1.0%) that has only two modalities, it is bigger (2.8%) for Civil status that has five modalities and it is double of it (5.2%) for Age that is a numerical variable calculated starting from day, month and year of birth.

Finally, an analysis was conducted for what concern the geo-localization of the households. Indeed the geo-localization of the households inside the correct building and dwelling it is crucial for any survey based on the sample frame derived from the Census data. Unfortunately, for some methodological reasons, the code attributed to the dwelling in the PES was different from the one adopted for the Census. This fact limited the possibility of the comparison only to the geo-localization of the households inside the correct building.

Geo Localization	%	95% Confidence Interval	
		Lower	Upper
Different BL	16.49	6.9	25.8
Same BL	83.51	74.2	93.1

Table 6 - Percentages of differences in the geo-localization of the household comparing Census/PES matched records

Since the geo-localization code considered is composed by four distinct codes (District code, Municipality Code, EA code and Building code) it is more subjected to errors than the other variables considered before. In this case seems more adequate assuming that mistakes in PES and Census were equally distributed. Under this assumption the indicators TE and TQ are:

	TE	TQ
geo-localization	8.2	91.8

This result seems to be confirmed by the LSMS (Living Standard Measurement Survey) that was carried 10 months after the Census on the base of the new sample frame derived from the census. In that survey indeed, the percentage of no-contact was around 9.0%.

IV. Conclusion

The overall procedure of edit and imputation of the 2011 Population and Housing Census in Albania is a very complex one. It is composed by different sub-procedures, each dedicated to a given unit of analysis. In its development, the focus was on the maximization of the final quality level of the data.
In particular, the following desired requirements have been set:

- to harness all the available knowledge regarding the involved units of analysis (buildings, dwellings, households, individuals), and in particular the logical relationships among and inside each of them, in order to be able to localize the maximum amount of errors on the basis of the detected inconsistencies;
- in localizing the errors, to respect the principle of minimum change, in order to minimize the deviations from the original distributions of data;
- in the imputation of erroneous and missing values, to maximize the use of techniques (as the nearest neighbor donor) in order to respect the multivariate original distribution of data.

The analysis carried out, based on the comparison between raw and clean data, shows that the impact of the edit and imputation procedure on the data related to the different units has been in some cases relevant, in order to obtain final datasets that are complete and free from inconsistencies and, at the same time, the original distributions have been preserved. Over all analysis of census data suggests very good quality of data collected in the census. The results of the census is highly coherence with the reasonable pattern of the characteristics of population and consistent with other sources of data available in Albania.

V. References

1. A systematic approach to automatic edit and imputation. Fellegi, I.P., and Holt, D., Journal of the American Statistical Association, 71, 17-35 (1976)

2. Evaluating Censuses of Population and Housing Census. US Census Bureau, Statistical Training Document, ISP-TR-5, 1985
3. Two algorithms for Error Localization Problems. Scarnò M and Caramanna L., 45th Scientific Meeting of the Italian Statistical Society (2010