This project is funded by the European Union

INSTAT
Republic of Albania
Institute of Statistics

ALBANIA

QUALITY DIMENSIONS OF THE 2011 POPULATION AND HOUSING CENSUS

2011 CENSUS-AL

May, 2014

QUALITY DIMENSIONS OF THE 2011 POPULATION AND HOUSING CENSUS OF ALBANIA
MAY, 2014

Director of the Publication:
Gjergji FILIPI, PhD

INSTAT
Olgeta DHONO
Ogerta ELEZAJ
Alma KONDI
Blerina SUBASHI

EU TECHNICAL ASSISTANCE
Guido PIERACCINI
Bart DE BRUJIN

Printed with the support of the Swiss Agency for Development and Cooperation

INSTAT
INSTITUTI I STATISTIKAVE
Blv. "Zhan D'Ark" Nr. 3, Tiranë
Tel : + 355 4 2222411 / 2233356
Fax : + 355 4 2228300
E-mail : info@instat.gov.al

Printing house: Gent•grafik

Preface and Acknowledgment

The 2011 Population and Housing Census of Albania is the 11th census conducted in the history of Albania. The preparation and implementation of the census operations required a large amount of financial, human, and material resources, which were supported by the Albanian government, the European Union and international donors. The methodology was based on the EUROSTAT and UN recommendations for the 2010 round of Population and Housing Censuses, taking into consideration the specific needs of Albanian data users.

One of the innovative technical aspects of the Albanian 2011 Population and Housing Census was the implementation of a comprehensive system of quality assurance for such a big statistical operation like the Population and Housing Census. INSTAT has the pleasure to release for the first time a technical report on the quality dimensions of the Population and Housing Census of Albania.

INSTAT would like to express the gratitude and acknowledgments for their valuable contribution to: the EU through EUROSTAT and the EUD to Albania, the UN through One UN programme, the Swedish government through SIDA and the Swiss government through SDC.

It would not have been possible to produce this report without the valuable contribution of the expertise of the EU Technical Assistance Project team. In particular, special thanks goes to Giulio Barcaroli (ISTAT) and Marco Scarno' (CINECA) for their contribution on defining and implementing the data cleaning strategy, to Ed Swires-Hennessy (The Statistical Consultancy) and Ian White (ONS) for their contribution on implementation, monitoring and evaluation the PES results, to Giorgio della Rocca (ISTAT) for his contribution at different stages of the data processing and to Dennis Osterberg (Statistics Sweden) for his help developing and testing the scanning platform.

INSTAT also wishes to acknowledge the contributions of its staff from the IT sector, the Methodology of Social Survey sector and the Socio-Demographic and Gender Analysis sector, who showed high professionalism and dedication.

Special appreciation also goes out to all other INSTAT staff involved in the census operation, data processing and in the production and analysis of the data required for this report.

Gjergji FILIPI, PhD

Director General of INSTAT

Lista e publikimeve tematike të Censusit 2011, Maj 2014
List of 2011 Census thematic publications, May 2014

- Censusi i Popullsisë dhe Banesave 2011: karakteristikat ekonomike
- 2011 Population and Housing Census: Economic Characteristics

- Dimensionet e cilësisë së Censusit 2011
- Quality Dimensions of the 2011 Population and Housing Census of Albania

- Kushtet e banimit dhe të jetesës
- Dwelling and living conditions

- Migracioni në Shqipëri
- Migration in Albania

- Një klasifikim i ri urban - rural i popullsisë shqiptare
- A new urban - rural classification of Albanian population

- Popullsia dhe dinamikat e saj - horizonte të reja demografike?
- Population and population dynamics in Albania - New demographic horizons?

- Projeksionet e popullsisë, 2011-2031
- Population Projections, 2011-2031

- Shqipëria 2011 Censusi në harta
- Albania 2011 Census Atlas

- Tipologjia e komunave dhe bashkive
- Communes and Municipalities Typology

- Lëvizjet vajtje-ardhje për qëllime punësimi
- Commuting from home to work

- Dinamikat e tregut të punës, 2001-2011
- Labour market dynamics, 2001-2011

- Aplikimi INSTATGIS – hartat në web (www.instatgis.gov.al)
- INSTATGIS – Atlas web application (www.instatgis.gov.al)

TABLE OF CONTENTS

List of Tables

List of Figures

ABBREVIATIONS

ALUIZNI	Agency of Legalization, Urbanization and Integration of Informal Areas/Constructions
CES	Conference of European Statisticians
CINECA	Consortium of universities
EA	Enumeration Area
EA ID code	Enumeration Area Identification code
EU	European Union
EUROSTAT	Statistical Office of the European Union
GIS	Geographical Information System
HQ	Headquarter
ICR	Intelligent character recognition
ICSE	International Classification of Status in Employment
ILO	International Labor Organization
IM	Leti's Dissimilarity Index
IMR	Intelligent Mark Recognition
Ind	Individual
INSTAT	National Institute of Statistics of Albania
ISCED	International Standard Classification of Education
ISCO88	International Standard Classification of Occupations
ISSN	International Standard Serial Number
ISTAT	Italian National Institute of Statistics
KFI	Key for Image
KS	Kolmogorov-Smirnov Distance
LFS	Labor Force Survey
LinkKing	Public domain application for record linkage and un-duplication
LSMS	Living standard measurement survey
MS	Member States
NACE rev2	Statistical Classification of Economic Activities - revision 2
PES	Post Enumeration Survey
PHC	Population and Housing Census
SAS	Statistical Analysis System
SDMX	Statistical Data and Metadata exchange
SMS	Short Message Service
TA	Technical Assistance
TE	Total Errors
ToR	Terms of Reference
TQ	Total Quality
UNECE	The United Nations Economic Commission for Europe
UK ONS	Office for National Statistics (ONS) of the UK national statistical institute

INTRODUCTION

The Population and Housing Census (PHC) is the biggest and the most expensive statistical operation that a country may carry out. Census results allow creating a picture on time and space about the structure and characteristics of the entire population of the country and about the housing units, whose relevancy largely exceeds the simple knowledge of the demographic reality of a country. Indeed, the planning of the future deeply depends on the availability of accurate information about what we are, who we are and how we live.

In the framework of production of official statistics, quality of collected data is only one of the quality dimensions that are needed to consider, since reliable data depends also on reliable and accurate processes. Indeed, in the Census operation, the risk of error is sparse at various levels and at different stages. Therefore, a comprehensive system of quality assurance has to be designed and implemented together with the Census operation itself.

Census management requires input and support from all functional areas and it is within this context that a tradeoffs is necessary to ensure an appropriate balance between quality and concerns about costs, response burden and other relevant factors.

This is the first time that the Albanian Institute of Statistics (INSTAT) implemented a comprehensive system of quality assurance for such a big statistical operation like the PHC. With all the limits illustrated in this report, it is the opinion of both INSTAT and the EU Technical Assistance Project that the implementation of such a system had as a result an improvement of the quality of the overall process.

The present publication is divided in two main parts: a first one, more qualitative, that provides a detailed overview, together with some elements of evaluation, about the main measures implemented to ensure the quality of the Census process and a second one, more quantitative, focused on the quality of the data collected.

1. QUALITY ASSURANCE

1.1 DIMENSION OF QUALITY ASSURANCE

Quality assurance is relevant for each activity and each operations stage of statistical process. It should be understood as a multi-dimensional concept, including the following dimensions: [1]

I. Relevance
II. Completeness
III. Accuracy
IV. Comparability
V. Coherence
VI. Timeliness
VII. Accessibility

This means that Data Accuracy is only one, albeit important, dimension of overall data quality. No Census is able to achieve a perfect score on each of the dimensions of data quality, as one dimension tends to improve at the expense of others. For example, very high quality data require training of Census staff at each staff level for such an extensive period, field monitoring and supervision at such an intensity, and data editing at such a comprehensive detail, that the project will exceed any acceptable time and budget limit. Thus, any Census will have to find a practical and acceptable balance within the bounds of existing resources and constraints.

Table 1-1 gives a summary of the main measures taken to ensure the quality of the 2011 PHC for each of these dimensions. Each of the listed measures includes a reference to one or more elements of sections 1.2 to 1.4, which provide a more detailed descriptive overview of key activities and procedures that were relevant for data quality.

[1] See e.g. UNECE 2006, Recommendations for the 2010 Censuses of Population and Housing; United Nations 2008, Principles and Recommendations for Population and Housing Censuses, Rev. 2.

Table 1- 1 - Quality assurance dimensions and measures in the Albania 2011 PHC

Dimension	Description	Quality assurance measures
I. Relevance	The degree to which data serve to address the purposes for which they are produced and sought by data users. The value is further characterized by the merit of those purposes, in terms of the mandate of the agency.	• Conduction of the Census by INSTAT as the delegated authority by law (A) • Compliance with the information requirements of the Census law (A) • Series of consultations with data users in Albania on Census questionnaire contents (B) • Series of consultations with Census experts on Census questionnaire design (B) • Compliance with international recommendations with regard to inclusion of Census core topics (B) • Approval of Census questionnaire contents by Central Census Commission (B) • Series of consultations with users on the dissemination program (Q)
II. Completeness	The degree to which data cope with the necessity of data users as completely as possible, taking restricted resources into account.	• Meeting stakeholders' demands for information provided by the Census (A, B) • Design of a questionnaire that minimizes the risk of missing or mixing-up information (B) • Development and application of forms that increase the likelihood of full and timely coverage during the enumeration (EA map list, Map list summary, Daily summary forms, Notification letter) (C) • Development and application of tailored procedures and documents to cover the institutional population (questionnaire, manual, summary form, training) (B, C, E) • Production of up-to-date and high-quality field maps that increase the likelihood of full coverage during the enumeration (D) • Special attention paid to the enumeration of Roma households and households of other ethnic minorities (B, F, H, I) • Implementation of a tracking system for enumerator boxes to avoid loss of relevant Census materials (L) • Observation and checking by Census field staff (M) • Application of effective monitoring procedures for enumeration progress (M) • Conduction of a post-enumeration survey to check Census coverage (O)

(Continues)

Dimension	Description	Quality assurance measures
III. Accuracy	The degree to which the data correctly estimate or describe the quantities or characteristics that the survey was designed to measure.	• A series of consultations with Census experts on Census questionnaire design to assure correct conceptualization of the data to be collected (B) • Development and use of a comprehensive enumerator manual as to improve understanding of information to be collected and to adequately respond to data-collection challenges (E) • Development and use of detailed and advanced training materials for enumerator-, controller- and supervisor trainings as to improve understanding of information to be collected and to adequately respond to data-collection challenges (K) • Thorough testing of questionnaires, forms, and procedures in a field test and pilot Census (F) • Observation and checking by Census field staff (M) • Provision of field support by INSTAT staff and EU Technical Assistance Project experts (M) • Conduction of a post-enumeration survey to check age and sex distribution (O) • Development, test and application of the Census Control Package for monitoring and managing large scale scanning operations that was tailored to the quality of available operators (P) • A parallel process was in place for quality assurance during scanning operations (P) • Development and use of comprehensive operator's manuals for the different actors involved in the data capture phase (P) • Development and use of detailed and advanced training materials for scanner operators, tile operators, verifiers, quality assurance operators and codifiers as to improve a proper understanding of the work to be accomplished by each one of them (P) • Development and application of a large set of editing and imputation rules to identify and solve inconsistencies and out-of-range values (P) • Several indicators which take into account the changes carried out by the editing and imputation procedure in terms of their number and/or magnitude have been constructed to assess the quality of the whole operation (P)

(Continues)

Dimension	Description	Quality assurance measures
IV. Comparability	The degree to which statistics are comparable over space and time.	• Compliance with international recommendations on definitions, classifications and procedures applied in the Census (B) • Compliance with international recommendations on the inclusion of 'core topics' in the questionnaire (B) • Maintaining overall comparability with the 2001 Census (B) • Maintaining comparability in the conceptualization of specific statistics with targeted surveys and the 2012 Agricultural Census (B) • Compliance with international regulations on comparability of Census results between countries (Q)
V. Coherence	The degree to which data from a single statistical programme, and data brought together across statistical programmes, are logically connected.	• Inclusion of questions in the 2011 PHC that served as the basis for planning and implementing the 2012 Agricultural Census (B) • Census results, providing the sampling frame of subsequent household surveys. • Application of internationally recommended definitions, classifications and measurement (B).
VI. Timeliness	The delay between the period to which information pertains and the date on which the information becomes available.	• Development of Summary forms and associated rapid data processing procedures to compile timely provisional Census results (C) • Training of a pool of reserve field staff to avoid delays in enumeration due to drop-out (I) • Testing the Census procedures with all interconnected components in a comprehensive pilot operation (F) • Detailed logistic planning for document printing, material assembly, transport and storage, and field staff recruitment and training (G, I, J, L) • Application of effective monitoring procedures for enumeration progress (M) • Development of efficient and effective transport and storage procedures of completed questionnaires and related Census documents (N)
VII. Accessibility	The availability of information and the suitability of the form in which the information is available.	• Implementing a communication strategy to inform stakeholders and the general public about the procedures of the Census and dissemination of results (H) • Several meetings with donors and users were held to review and finalize the Census Dissemination Program (Q) • Implementation of a dedicated Census website providing timeless dissemination of Census results (Q) • Availability in the Census website of a sample of Census micro data representative of the level of Prefectures (Q)

1.2 CENSUS PREPARATION

A. CENSUS LAW

A component of the Relevance dimension of the quality is the reference to any acts, decrees and recommendations on which the compilation of statistics is based and which contribute to define the information content and the purpose of use of the statistics. To this extent, law No. 8669 dated 26 October 2000, "The general Census of population and dwellings", and its amendments[2] provide the legal basis of the 2011 PHC of Albania. Among other things, the laws stipulate the following:

- The aim of the Census:

  Providing the parliament, the government, the local authorities, the economic, scientific and cultural organizations, as well as the whole civil society, with reliable statistical information that is needed for planning and implementing general policies of development, for private and public analysis and decision making, for scientific research and, in general, for improving the citizens' knowledge and understanding of the demographic, economic and social reality of the country.

- The information to be provided by the Census:

  a. The figure and the geographical distribution of the resident population;

  b. The demographic structure and main characteristics of the population;

  c. The number, the geographical distribution and the structure of dwellings and buildings used for housing purposes;

  d. The housing conditions of the population.

- The authorized body to conduct the Census:

  The Census is organized and carried into execution by the Statistical Institute (INSTAT), under the supervision of the Central Commission for the Census and with the support of Census commissions in Prefectures and Census offices established in Communes and Municipalities.

The Census Law and amendments also specified the confidentiality of the individual information so as to help assure cooperation of the public to achieve full coverage. In this respect, the Law No. 10442 refused the use of Census data to "be used for any electoral list or to update any of civil registry and any other administrative register".

The 2011 Census complied with all these legal stipulations.

B. QUESTIONNAIRE DESIGN

With regards to the Relevance dimension, a key point is also the implementation of methods for hearing users and for monitoring the relevance and usefulness of the statistics concerned. To this respect a series of consultations with data users to identify the relevant information to be collected, as well as to ascertain the appropriate wording of questions and answers was held at different stages of the Census process. The consulted users included line ministries, civil society organizations, including representatives of ethnic minorities, and academic institutes. Meetings were conducted from 2009 to well into 2011. This resulted, among other, in an expanded battery of questions on migration and the inclusion of questions on disability and socio-cultural backgrounds.

The key concepts essential for the definition of the target population object of the study and the related classifications adopted are two aspects that have to do with the dimensions of Comparability and Coherence. Indeed, data are most useful when they enable reliable comparisons across space like countries or regions and are coherent as much as elementary concepts can be combined reliably in more complex ways. In this respect, two technical meetings with INSTAT staff, Technical Assistance staff and international Census experts to review and finalize the questionnaires were held in February and March 2011.

The International recommendations on data to be collected in Population and Housing Censuses was taken in high consideration, more specifically with a full compliance with regard to the inclusion of 'core topics' specified in the UNECE Recommendations for the 2010 Censuses of Population and Housing (UNECE 2006). Moreover, the international recommendations on definitions, classifications and procedures, such as definitions on dwellings and usual residence (following UNECE recommendations), status in employment (ICSE) and employment-unemployment (ILO), educational

[2] Law No. 10084, dated 23 February 2009; Normative Act No. 6, dated 30 September 2009; and Law No. 10442, dated 7 July 2011.

classification (ISCED) and measurement of disability, following the Washington Group recommendations were also fully endorsed.
From the point of view of the Coherence, the inclusion of questions that served as a base for the planning and the implementation of the 2012 Agricultural Census should be also noted.

Looking at the Comparability dimension, it is also essential to ensure a high consistence of the data with other sources on the same topic. In respect of this, differences between concepts were analyzed and their impacts assessed. More in details, the Albania 2001 PHC questionnaire and different surveys conducted by INSTAT, e.g. households, labor force characteristics and disability in such programmes as the LFS, LSMS and the Agricultural Census were taken in consideration.

Relevant to the dimension of Completeness is design questionnaires that minimize the risk of missing or mixing-up information. General good-practice considerations for questionnaire design, such as the optimal sequence of questions, explicitness and simplicity of wording, use of exhaustive and mutually exclusive answer categories, proper routing and routing instructions, user-friendly format and layout, expected competence of non-professional enumerators, interview burden and respondent fatigue were also taken in consideration.

The standard questionnaire consisted of a booklet to minimize the risk of questionnaire pages getting lost or mixed-up. The inclusion in the booklet of a household listing cross-referenced with the individual questionnaires was designed to reduce the risk of omitting household member's. Separate individual questionnaires were available for enumeration of households with more than six members.
In addition to the standard questionnaire booklet, a special questionnaire was developed to cover institutional households.

Revised questionnaires were first internally tested in INSTAT in April 2011 and subsequently in the field among 1,500 households, in June 2011. Approval of the questionnaires by the Central Census Commission was obtained in August 2011

C. DESIGN OF CENSUS FORMS

Besides the questionnaires, a set of Census forms was developed to increase the likelihood of full coverage during the enumeration. This action refers mainly to the Completeness dimension of the quality. The forms developed were:

- *EA map list*. Preceding actual enumeration, controllers, rather than enumerators, made a full inventory of the buildings and dwellings in each of the EAs in his/her controller area to update the EA maps. On the basis of provided EA maps and field observation, this inventory was recorded on the controllers' EA map list, which acted as guidance for enumerators and as a monitoring instrument for controllers. This procedure increased the likelihood that enumeration achieved full coverage of the respective EAs, because: i) controllers are better qualified staff than enumerators; ii) it increased the field knowledge of controllers; iii) it enhanced the monitoring opportunities for controllers; iv) it separated the responsibility of listing the dwelling to be enumerated and the actual enumeration.
- *Map list summary*. This document summarized the EA map lists of each controller and was used at Census headquarters as baseline input for monitoring enumeration progress in the field.
- *Collective living quarters form*. During the pre-enumeration inventory phase, controllers were required to record on this form any collective living quarters for being enumerated by specially assigned enumerators (e.g. hotels, orphanages, old people's homes, boarding schools, prisons, monasteries and convents, hospitals and other health-related and welfare institutions).
- *Daily summary forms for enumerators, controllers, supervisors and district coordinators*. The Daily summary forms recorded the daily completion score of enumerated dwellings, households, persons and refusals. This provided information to track enumeration progress against the inventory information recorded on the Map list summary at each level of the Census operation, from enumerator to Census headquarters.

Between the forms developed a special mention should have the *Summary forms for enumerators, special enumerators, controllers, supervisors and district coordinators*. The Summary forms recorded the final coverage score of key results of the Census for the different levels of the Census operation. For enumerators and controllers the summary form functioned as a tool for recording and monitoring daily tasks. Moreover, at the end of the enumeration the forms permitted to compile timely provisional results. Therefore, this action belongs to the quality dimension of Timeliness.

D. MAPPING

An innovation of the 2011 PHC was the use of Geographic Information System (GIS) technology at different stages of the Census process. INSTAT implemented a GIS national Census database containing a total number of 945,757 buildings: 295,057 in urban areas and 650,700 in rural areas and also 11,712 EAs from which 5,198 in urban areas and 6,514 in rural ones.

Census mapping operations started in 2009 with the implementation of a digital mapping infrastructure covering all the territory of Albania at building level using the most recent satellite images. This was a challenging process and requested important financial and human resource.

The first important step to create the GIS database in order to divide the country into small areas for statistical purposes was the map update in the field. For this purpose, the satellite images provided by ALUIZNI were used. These images were taken in 2007, therefore, it was necessary to update the maps mainly in urban areas. All the 74 cities of Albania were updated in the field using the 2007 satellite images. In the process of map update were included also some rural areas, mainly those in the outskirt areas of the principal cities of Albania, such as Tirana and Durres. The administrative boundaries defined in the GIS System for statistical purposes were: i) Geographic boundaries of districts; ii) Unique ID codes for buildings and EAs, iii) Street center lines with names, where available.

From 19,974 buildings under construction updated in 2010, 3,056 were inspected in order to verify if they were inhabited. Map updating in urban and peri-urban areas for buildings classified as "under construction" were concluded in July 2011. A total of 510 EAs was updated in urban/peri-urban areas. Furthermore, new detected buildings were included in the GIS Census database. In rural areas, an average of 30% of the buildings inspected was classified as not used for residential purposes.

Concerning the quality dimension of Completeness has to be underlined the production of up-to-date and high-quality field maps increased the likelihood of full coverage during the enumeration.

The digitalization and geo-coding of Census maps was concluded in July 2011, after the end of map updating activities. The total number of dwelling units estimated by the Cartography and GIS Unit at INSTAT was equal to 1,216,742 of which 553,664 in urban areas and 640,225 dwellings in rural areas.

The delineation of enumeration areas and the printing of Census maps were based on a semi-automatic procedure using the new geo-database. INSTAT developed a template for map layout in order to speed-up the printing process. Maps were prepared for each enumerator, controller and supervisor. Maps of municipalities and communes were also provided to INSTAT regional offices.

Printing of maps for the Census started at the end of July and ended at the beginning of September 2011. The Cartography and GIS sector printed all the enumeration areas maps in A3 format and about 3,000 maps in A1 format using large format printers available at INSTAT.

In order to facilitate the identification of the buildings with their unique respective codes on the map, a nominal scale between 1:1500 or larger was used for the layout maps in urban areas. In some cases, where an EA covered a large geographical area, the EA map included also a zoom that showed the details to support the orientation of the enumerators in the field. For this reason, the enumerators were provided with additional large-scale maps (map books) of one or more parts of the EA where the buildings were located. Almost all the Census maps were printed in A3 format in 2 copies: one copy for the controller and one copy for the enumerator. The use of the A3 format facilitated the use of maps in the field, and indirectly improved the Census coverage.

E. DEVELOPMENT OF FIELD MANUALS

A main component of the quality dimension of the Accuracy is the development and the adoption of comprehensive manuals as to improve understanding of the information to be collected and to adequately respond to data-collection challenges. Field manuals were developed for different levels of field staff and tailored to their specific tasks and duties: enumerators, enumerators for institutional households, controllers, supervisors and district coordinators. A choice was made to provide comprehensive guidelines that could be used as a reference guide during training and during fieldwork.

The Enumerator manual included definitions of Census-relevant concepts, detailed instructions for questionnaire completion, EA map use and implementation of various procedures preceding, during and following enumeration. Manuals for controllers, supervisors and district coordinators provided additional instructions for EA map updates, supervision, monitoring, recording and questionnaire handling.

Draft manuals for enumerators and supervisors were provided during the June pilot Census and updated on the basis of field experience and questionnaire adjustments.

F. FIELD TEST AND PILOT CENSUS

Another relevant component of the quality dimension of the Accuracy was the testing of questionnaires, forms, and procedures in a field test and pilot Census. INSTAT staff members tested the questionnaires that were drafted on the basis of Census users consultations and technical meetings in April 2011 in different field settings. Lessons learned fed into revised Census questionnaires. The revised questionnaires were subsequently tested in a pilot Census on June 2011, an almost complete rehearsal of the full Census.

In June and July 2011 the pilot Census process and results were evaluated and thereupon questionnaires, forms and procedures were revised when necessary.

G. PRINTING OF CENSUS DOCUMENTS

A detailed overview of materials to be printed for training and enumeration purposes was drafted in the pre-enumeration phase. The Census required the printing of 46 different documents. In total these amounted to around 43 million pages, the large majority of which were Census questionnaires. Besides questionnaires, manuals, summary forms, daily summary forms and other Census forms, the documents also included authorization letters, delivery notes, contracts and terms of references, badges, labels, publicity leaflets and maps. The bulk of the documents was printed by a printing house, for which a tendered contract was made.

An important component of the quality dimension of the Accuracy was the verification of the compliance of the printed questionnaires with the scanner requirements. The printing specifications indicated that, to comply with scanner requirement, the questionnaires had to be printed with high quality paper and specifics predetermined colors while others kind of materials had regular printing requirements, in order to reduce the costs. Daily samples of the printed questionnaires were taken by INSTAT to verify the compliance with the printing specifications.

1.3 DATA COLLECTION

H. PUBLICITY CAMPAIGN

A key point in the quality dimension of the Accessibility is the implementation of a communication strategy to inform the general public about the procedures of the Census and dissemination of results. The Census is an operation that is completely dependent for successful coverage upon the cooperation of the general public. Consequently, INSTAT, supported by the EU Technical Assistance Project, developed and implemented a communication strategy that targeted the public through various channels, including a Census website, radio, television, newspapers, press conferences, posters, leaflets and a Census call centre. Communication activities started in August 2011 and continued until the collection of Census materials from the field in November 2011. Key messages that were conveyed related to:

- The reasons for conducting the Census and the need for accurate and complete Census data;
- The objects and the timing of the Census, as well as the general enumeration procedure;
- The legal obligation to participate in the Census;
- The confidentiality of data collected in the Census and the strict segregation between information from the Census and civil registers;
- The expected timing of publishing provisional and final results;
- Issues that appeared to be sensitive in the context of Albania, such as questions on ethnicity, religion and mother tongue;
- Information portals where the public could obtain additional information about the Census or could address complaints.

In addition, INSTAT spoke persons were instructed how to respond to a variety of questions that could be raised about the Census. In response to appeals from sections of the Greek minority to boycott the Census, representatives from different levels in the INSTAT hierarchy established communication with the Greek community. Special attention was also paid to introduce the Census into the Roma community and to obtain a readiness to participate.

I. FIELD STAFF RECRUITMENT

For the conduction of the Census, over 14 thousand temporary field staff, enumerators, controllers and supervisors, were recruited.

A key strategy in the quality dimension of Timeliness, was the appointment of a suitable number of trained reserves to avoid delays due to possible drop-out of enumerators. As a rule, the number of reserve enumerators and controllers amounted to 5 percent of the required staff for each district. For supervisors, one person was assigned as reserve per district. Table 1-2 provides an overview of staff recruited per prefecture together with the number of staff that were trained and kept as a reserve to replenish drop-outs during the fieldwork phase.

Table 1-2 - Temporary Census field staff recruited, by selection status, staff type, and by prefecture

Prefecture	Selected			Reserve		
	Enumerators	Controllers	Supervisors	Enumerators	Controllers	Supervisors
Berat	666	132	6	32	6	3
Dibër	533	110	8	27	6	3
Durres	1,067	208	10	58	10	2
Elbasan	1,227	239	13	61	12	4
Fier	1,426	277	14	72	15	3
Gjirokastër	511	103	7	26	4	3
Korçë	1,141	224	13	56	12	4
Kukës	320	63	5	16	4	3
Lezhë	512	104	7	26	5	3
Shkodër	893	177	9	45	9	3
Tiranë	2,648	510	27	140	24	3
Vlorë	962	183	9	52	9	3
Total	**11,906**	**2,330**	**128**	**611**	**116**	**37**

The selection procedure for temporary field staff included the submission of Curriculum Vitae, and taking tests and interviews. The Terms of Reference for each type of field staff was developed by INSTAT and stipulated their role and duties in the enumeration process and the requirements to fulfill the respective jobs. For enumerators and controllers, key requirements included educational attainment, age and place of usual residence, in order to assure familiarity with the area of work and to reduce travel time and costs. In addition, attention was paid to recruit persons with Roma ethnicity in order to assure easy access to the Roma community. Census commissions at the municipality/commune and prefecture level made the final appointment of shortlisted candidates for, respectively, enumerator and controller positions.

Whereas recruitment of enumerators and controllers was done locally, recruitment of supervisors was centrally done in Tirana. INSTAT with EU Technical Assistance Project supported, developed and implemented interviews and rating procedures to draw up a shortlist of candidates, from which the Prefecture Census Commissions did the final selection and appointment.

In addition to temporary field staff, a large share of INSTAT staff in the central and district offices was assigned to tasks related to the preparation and implementation of the Census: as trainers, monitors or for other supporting activities, or in combinations of these.

J. CENSUS OFFICES AND TRAINING VENUES

For the specific purpose of the Census, the accommodations were secured for the Census offices in the municipalities and communes, and for the training of field staff. The Census offices for the local Census commissions, 401 in total, were established in all municipalities and communes in July and August 2011 to facilitate the processes of selecting enumerators, receiving, storing and dispatching the Census materials and monitoring the enumeration process. The Technical Assistance project supported INSTAT in securing Census offices with adequate access and safe storage capacity.

For the 12.517 enumerator trainees, in total 251 classes in 107 different training venues were arranged and for the 2.446 controller trainees 51 classes in 22 training venues across the country. These ratios amount to on average around 50 trainees per class. Supervisor training was centrally conducted in Durres. INSTAT district offices, supported by the EU Technical Assistance Project, identified and, if necessary, furnished training venues that were located in the respective locations and that could adequately accommodate selected trainees. Public institutes, universities, municipality/commune centers and cultural centers, provided the majority of venues. Catering and, where necessary, sleeping accommodations were also arranged.

K. FIELD STAFF TRAINING

For the 2011 Census training the methodology of master trainers for field staff was adopted. For this purpose INSTAT staff members from headquarters and district offices were trained as master trainers who, subsequently, trained supervisors (22 August to 3 September), controllers (6-10 September) and enumerators (22-26 September). Supervisors supported the master trainers in the training of controllers and enumerators.

The considerations for adopting this methodology instead of the originally planned cascade training included the following consideration: i) INSTAT master trainers have more understanding of and experience with statistical operations; ii) maximizing the consistency of training messages in the successive levels of staff training; iii) minimizing the accumulation of misconceptions during transfer of training input.

In the quality dimension of Accuracy an important role is played by the development and use of detailed and advanced training materials for enumerator's, controller's and supervisor's trainings as to improve understanding of information to be collected and to adequately respond to data-collection challenges.

INSTAT, together with the Technical Assistance developed the training programmes and detailed training materials. In total, 88 digital presentations were developed to address the variety of tasks and duties, documents, procedures and Census questions. These materials also included examples, exercises, role playing, mock interviews and daily tests that allowed assessment of the trainees' level of understanding and provided handles of the final selection of field staff. In addition to these presentations, a video was recorded and used to assure that the key messages in the various trainings were conveyed correctly and consistently.

The Technical Assistance further engaged in the training of supervisors, controllers and enumerators by monitoring the training logistics and performance and supporting trainers.

L. ASSEMBLY AND DISTRIBUTION OF FIELD MATERIALS

With respect to the quality dimension of Timeliness a detailed logistics planning for assembling materials and ensuing transportation of these materials to the training venues and to the Census offices across the country, was developed by INSTAT with the support of the Technical Assistance.

To better manage and organize the information essential for the logistic planning a Census Management Tool was developed and used for: i) store the number of questionnaires needed in each EA; ii) print the label of the enumerator boxes and the deliveries notes; iii) track the serial numbers of each questionnaire; iv) track the transportation of enumerator boxes from the assembly centre to the Census offices and from the Census offices to the storage centre; v) verify that all the materials in the box were received.

In the period 15 August to 30 September more than 41 thousand sets of materials, together containing 57 different items were assembled by a team of up to 40 assembly staff. The assembly was conducted in the premises of the printing house that was responsible for printing of the Census documents. This implied an easy supply of the bulk of materials to be packed. The field sets for enumerators, including the Census questionnaires, made up the bulk of the volume. Based on the preparatory mapping exercises, an estimated number of required questionnaires plus reserve was assigned to each enumerator box. Pre-printed labels specified the number of questionnaires for each enumerator box, together with EA ID codes and a scanning bar code that allowed tracking of each individual box.

For the transportation of materials, a detailed distribution plan was designed per assembly set. Distribution of training materials and field materials for controllers, supervisors and district coordinators was implemented by INSTAT, while the around 12 thousand enumerator boxes for the fieldwork were transported by the Albanian army under supervision of INSTAT staff. The distribution plan for the enumerator boxes specified the date of transportation, the exact number of boxes to be transported, the truck size required, the route to be followed, the names of the communes and municipalities to be visited, the address of the local Census offices and the number of boxes to be delivered there. Some days of margin were included and remote areas were serviced first as to maximize the likelihood of timely delivery. In the end, all enumerator boxes were received on time.

In the Census offices the local Census commissions were responsible for reception and checking the correct number of enumerator boxes. In order to track the transfer of enumeration materials in each step in the chain from the assembly centre to the enumerator in the field before the enumeration and upon completion of the enumeration from the enumerator to the store house, delivery notes were used that specified relevant information, such as the number of boxes, the ID codes of the boxes and completeness of materials.

M. FIELDWORK SUPERVISION, SUPPORT AND MONITORING

The Census data collection was scheduled for three weeks, from 1 to 21 October 2011. INSTAT extended this period to 31 October in order to complete the enumeration in areas where under-coverage was reported. Several parallel strategies, all belonging to the quality dimension of Completeness and Accuracy, were applied to supervise and monitor the enumeration process in the field.

The Census hierarchy of enumerators-controllers-supervisors-district coordinators included the standard mechanism of supervision in accordance with the tasks and duties of the staff above the level of enumerators. The tasks of controllers and supervisors related to quality assurance included: i) observing interviewers during enumeration; ii) checking households already enumerated; iii) checking coverage of the EAs; iv) checking completed questionnaires.

The local Census commissions were involved in monitoring the performance of the Census field staff in the respective communes and municipalities, and reported to the prefecture Census commissions, who, in turn, reported to the Central Census Commission. Staff from INSTAT headquarters was assigned to the districts for general support to and supervision of the enumeration process.

An SMS-based reporting system together with a web-GIS application was developed in order to monitor the Census process during the enumeration. The application was designed and implemented by INSTAT's Cartography Unit together with the IT Department, with the support of project experts. The Web-GIS application provided a valuable tool for the monitoring of the Census progress. Indeed, the application was updated daily with the data collected via SMS by the enumerators and controllers. During the enumeration period, a summary coverage report was prepared every day with the data received the previous day. The report showed data by district and by mini-municipality for the city of Tirana. The web-GIS application showed data by small area, allowing a real-time monitoring of the data collection process.

The data reported by SMS and uploaded on the web-GIS application was compared every day with the data reported to INSTAT HQ by INSTAT regional offices.

The Technical Assistance, represented by 2 Key Experts and a team of 19 national and international Short-Term Experts, supported and observed field operations, together with representatives from INSTAT regional offices and headquarters. The results of this observation are illustrated in section 1.5 of this chapter.

N. RETURN AND STORAGE OF CENSUS MATERIALS

In the frame of the quality dimension of Timeliness efficient and effective transport and storage procedures of completed questionnaires and related Census documents were developed.

Upon completion of the enumeration and questionnaire checking by controllers and supervisors, enumerator boxes with questionnaires and other Census materials were collected from the Census offices in the communes and municipalities. The Albanian army again transported the enumerator boxes on the basis of a detailed district-by-district transportation plan that specified the date of transport, the names of the communes and municipalities to be visited with the address of the Census office, the exact number of boxes to be collected, the truck size required and the route to be followed.

In the store house a detailed plan was laid out to efficiently handle and store incoming enumerator boxes. Specific actions were in place to retrieving documents that needed immediate data processing and recording the key results from the Summary forms.

Around 20 staff was involved in these tasks that were largely performed in parallel to the operation of collecting enumerator boxes from the field. The operation was carried out from 7 to 21 November 2011.

Belonging to the quality dimension of Accuracy, there should be noted that a tracking system of the EA's boxes, making use of the barcode on each box label, was developed to verify that all the boxes were received from the field.

O. POST-ENUMERATION SURVEY

Even though the word 'Census' implies a 100% count of the people and housing, it is rarely possible to achieve this result. Coverage errors arise from omissions or duplications of persons or housing units in the Census enumeration. The sources of coverage errors include incomplete or inaccurate maps or lists of enumeration areas, failure by enumerators

to canvass all of the units in their assignment areas, duplicate counting, omission of persons who are not willing to be enumerated, erroneous treatment of certain categories of persons such as visitors or travelers and loss of Census records after enumeration.

The Conference of European Statisticians (CES) agreed in their preparations for the 2010 and 2011 Censuses in Europe [3]that the Censuses should be evaluated to check on coverage and the quality of the information provided. Further the European legislation on the Censuses requires Member States to report on the quality of the Census results that they transmit to the European Commission (EUROSTAT) [4]. Consideration should be given to coverage and quality of information collected in the Census. INSTAT therefore planned a survey to measure the coverage and quality of the 2011 Census of Albania.

The essential features of the data defined by the CES (individual enumeration, simultaneity, universality within a defined territory, availability of small-area data and defined periodicity) were covered by the Census and the post enumeration survey met all except the availability of small area data: for this latter point, the sample was not large enough to break down the results even by prefecture.

For the first time in the history of Censuses in Albania, the Institute of Statistics of Albania (INSTAT) conducted a post enumeration survey to measure the effectiveness and reliability of the 2011 Census. This Census was by full enumeration of the population using some 12,500 specially recruited and trained enumerators. With such a large field force, a measure of the quality of the results was essential. This will enable a better understanding of the Census results and also provide additional information when using the results in other processes, such as forming a base for the projection of the population of Albania.

The post enumeration survey was designed[5] to provide an estimate of the Completeness of the Census process in the identification of both inhabited residential dwellings and the population.

The initial design was for a measure relating to the whole of Albania to produce an achieved sample of 4,000 dwellings. Information was available on the estimated numbers of buildings and dwelling units within each prefecture (Prefecture): each individual enumeration area was also designated as urban or rural according to the administrative division and the summary statistics also available with this categorization. Further, information on survey response rates by urban and rural areas within Prefecture and on the building/non-residential proportions in rural areas was available and used in the sample design.

The sample design was completed in July 2011 and proposed a sample of some 61 enumeration areas, specifying the number of areas to be chosen within each Prefecture by type of area. The required number of individual enumeration areas was then to be selected randomly within Prefecture / type of area by INSTAT.

During the Census operation, concern was expressed by INSTAT with regard to the refusal and non-contact rates in urban areas and a supplementary sample was suggested for urban enumeration areas. Consequently, an additional 44 areas were selected within five municipalities, randomly without replacement, in proportion to the populations in those urban areas.

The questionnaire for the survey was designed by INSTAT with advice from international experts. It was to be a short survey with the minimum number of questions to achieve the aims. Decisions as to what information was to be included, took into account the possibility of analyzing the resultant information.

The original design of the questionnaire was quite complex and designed to collect more information than was necessary: indeed, some of the information would be provided by so few people that the information could not be analyzed from the original sample. This would have increased the amount of time taken to complete the questionnaire in the field without any commensurate information gain. At a meeting in October 2011, the questionnaire was reviewed by the Technical Assistance team and only items that could be used in the matching process or comparative analysis were retained.

Four days of training for interviewers was held in November 2011 to ensure accurate understanding of the concepts, definitions and operational procedures of the survey. An extensive manual was given to each interviewer in case queries arose during the survey. Interviewers were mainly chosen from those applicants who had not worked on the Census to achieve independence. A small number, however, was recruited from those that had also undertaken the Census but were assigned to different areas from the ones assigned in the Census.

In addition to the extensive training, supervisors monitored progress and quality of completion on a daily basis in the urban areas and every two or three days in the rural areas. This monitoring occurred when interviewers returned completed questionnaires to the area offices.

[3] Conference of European Statisticians: Recommendations for the 2010 Censuses of Population and Housing (ISSN 0069-8458)
[4] EU legislation on the 2011 Population and Housing Censuses Explanatory Notes (ISSN 1977-0375)
[5] PES design was undertaken by Mr. Ed Swires-Hennessy, a Chartered Statistician of The Statistical Consultancy, UK, 9 Arlington Close, Newport, NP20 6QF, United Kingdom

The survey itself was undertaken in November 2011 over three weeks. This close proximity to the Census date meant that few changes would have taken place to both the numbers of inhabited dwellings and the population in them. The starting point for the interviewers was the same maps of the enumeration areas as used for the starting point in the Census with the enumeration area boundaries marked. This ensured that the interviewers did not assume anything about buildings and residential units, but investigated the whole of their area to find residential units.

The questionnaires for each enumeration area were collected together into a labeled box. Each interviewer reviewed their questionnaires for consistency and completeness. In the main Tirana office, many hundreds of questionnaires were examined during the three-week collection process by the supervisors: systematic errors were discussed with the interviewers or the supervisors as appropriate. This checking was also undertaken in a small number of other offices by members of the Technical Assistance team.

The Technical Assistance, represented by 2 Key Experts and a team of 2 international Short-Term Experts, supported and observed PES field operations, together with representatives from INSTAT headquarters. The results of this observation are illustrated in section 1.6 of this chapter.

After the return of the PES questionnaires, the following actions were implemented: i) development and implementation of data-capture and data-editing programmes; ii) development and implementation of the matching rules; iii) analysis of results and reporting. Section 2.5 of this publication provides detailed information on the PES-based quality assessment.

1.4 DATA PROCESSING AND DISSEMINATION

P. DATA PROCESSING

The decision to use automated entry versus manual entry was partly based on timetable requirements and budget. Other factors, such as whether it was feasible or possible to implement more sophisticated technology also were taken into consideration.

Image scanning technology is a system used to capture data from a questionnaire (form) with a limited amount of human intervention: once the questionnaire's forms are scanned, the images are saved and passed to an IMR/ICR subsystem that attempts to infer the contents of the answer. Depending on the confidence of the recognition process, the IMR/ICR system either accepts the inferred result or rejects it.

ReadSoft FORMS is a software application for automatically capturing and managing information from forms. ReadSoft scans, interprets, and verifies the forms, then transfers the data to a target system. INSTAT had a contract with the provider of ReadSoft that includes the development of a Census Control Package, with extended features for keeping track of the data quality and the flows within the scanning system.

With regards to the quality dimension of Accuracy, much effort was put into testing and fine-tuning the Census Control Package. The main task for 10 INSTAT employees from April to October was to test, troubleshoot and communicate any encountered problems to INSTAT and to ReadSoft developers. Despite some initial problems, when the scanning of the PHC materials started in November 2011 the Census Control Package were proved stable enough to be used.

Even with highly sophisticated interpretation engines a certain number of KFI (Key for Image) operators are needed to correct errors and, for obvious reasons, they cannot all be highly qualified. Consequently, there are two important requirements at this stage: verification should be simple and fast, and the KFI operators must not be able to insert additional errors. The solution adopted by INSTAT was the so-called 'Mass verification': the images of all interpreted characters were presented as a group according to their value. If a character appears in the wrong group, the verifier just selects it and he will be automatically taken to the field in which the character appears, to correct it. First, digits are mass verified, then letters and finally mark fields.

The next phase was the verification in which the field containing values selected in the tile phase or having a low confidence index of interpretation were manually verified with the images. In this phase also field that did not respect the range definitions were verified.

A parallel process, belonging to the dimension of Accuracy, was in place for quality assurance: a selected group of operators has checked for a second time all the individual questionnaires flagged with at least 4 inconsistence. A specific application was designed to show the questionnaires images and the value stored in the different fields, highlighting the values that were generating an inconsistence. Here, the objective was not to correct enumerator's mistakes, but being sure that the inconsistence was not ascribable to ICR/IMR misinterpretation. Another component available for quality assurance was the supervision, used for double-checking the verification process. One over twenty questionnaires was sent to supervision and the supervisor operators had to confirm all the verified fields.

The recruitment of the data processing operators was based on formal procedures and standards required for this process. The evaluation and selection of the candidatures were based on the following criteria: i) be an Albanian citizen, resident in Tirana; ii) posses the secondary diploma; iii) have good knowledge of Microsoft Office; iv) be 18-50 years old; v) have good communication skills; vi) having similar work experience was considered an advantage.

The announcement of the vacancy positions was published in two national daily newspapers for two days, respectively 5 and 7 November, and in INSTAT official website for five days starting from November 5, 2011. The ToR-s and respective job description were prepared by the INSTAT IT department.

The number of the applications was double the number of the operators requested. The selection was carried out from an evaluation committee established for this purpose and based on the criteria established by INSTAT, after an accurate evaluation of the documents attached to the applications submitted. Specifically, there were recruited a total of 120 operators.

The personnel started working on November 10th, 2011 and continued until April 24th, 2012, in two shifts. The activity of Data Entry Personnel originally forecast for 4 months was extended by 1.5 months. This modification has been done taking into consideration the time forecast for processing the remaining questionnaires of Census, the questionnaires of PES and the building lists.

The success of the Data Capture process depends to an important extent on the training of the staff involved in the capturing operations, again belonging to the quality dimension of Accuracy. Specific manuals and training sessions were organized for the different actors involved in the data capture phase: Scanner operators, tile operators, verifiers, quality assurance operators and codifiers.

A well-designed Census with minimal errors in the final product is an invaluable resource for a nation. To obtain such result, data must be free from errors and inconsistencies to the greatest extent possible. Census editing is the process of detecting errors that were made during and after data collection and capture, and then adjusting individual items to improve data quality. The process of cleaning can be described as having two components:

- *Editing* is the systematic inspection and correction of responses according to predetermined edit rules. Editing is done to ensure the validity and consistency of individual records and relationships among records in a household and to check the reasonableness of the aggregated data. Census publications will contain a certain amount of meaningless data if national statistical offices do not edit the Census or survey results. Editing reduces distorted estimates, facilitates processing, and increases user confidence. However, editing cannot correct all Census errors, including questionnaire responses that are internally consistent but are in fact instances of misreporting on the part of respondents or improper recording on the part of enumerators.
- *Imputation* is the process of resolving problems concerning missing, invalid, or inconsistent responses identified during editing. Imputation works by changing one or more of the responses or missing values in a record or several records being edited to ensure that plausible, internally coherent records result. Imputation values for erroneous or missing items are generated by using other entries from the housing unit, person, and other persons in the household, always in accordance with specified procedures. Records should satisfy all edits after imputation. Imputing a minimum number of variables is usually best, thereby preserving as much respondent data as possible.

The problem of assessing the impact of a data cleaning procedure consists in comparing the set of data after deterministic correction with the set of coherent clean data obtained applying the edit and imputation procedure. In order to allow this assessment, indicators at the aggregate level, which take into account the changes produced in terms of number and/or magnitude, have been considered. Indicators at variables level to underline the differences produced in the distributions of each variable were also considered. Section 2.4 of this report provides detailed information on the Accuracy of the editing and imputation procedure.

Q. DISSEMINATION

Several meetings with donors and users were held to collect their ideas regarding the 2011 Census Dissemination Program, according with the definition of quality Relevance. All comments and suggestions were collected and presented in an organized plan to the data users in a final workshop held in February 2012. Comments were carefully reviewed and informed INSTAT of users' needs and expectations. The input received during the consultations was important to improve the 2011 Census dissemination program.

INSTAT implemented a dedicated Census website providing timeless information about the Census conduction and the Census progress. A specific section of the website was dedicated to the communications with the citizenship during the enumeration. Another section contains the publication of the Census results and the access to the micro and macro data. All the final Census results were disseminated via the Census website as one of the main channel of dissemination. INSTAT

used PC-Axis as its strategic tool for data dissemination for a long time and all the Census results were made available through it on the Census website. For less skilled users the Census data were made available also in excel format.

INSTAT has made available on the Census website a 3% sample of Census micro data, useful to users who are doing research. Users can use micro-data to study relationships among Census variables not shown in existing Census tabulations. The data are anonymized and representative at prefecture level. While preserving confidentiality, these micro-data files permit users with special needs to prepare virtually any tabulation.

National Censuses are of greater value if their results can be compared between countries. This is why the EU is taking steps to harmonize Census outputs. The EU legislation on Population and Housing Censuses defined 60 mandatory aggregated hyper-cubes for member states. Since not a EU member country, Albania was not obliged to provide these hyper-cubes to EUROSTAT. Nevertheless, INSTAT took the decision to make available a huge subset of them (more than 50) by the end of 2014. The technical format that will be used for the transmission of data will be the Statistical Data and Metadata eXchange (SDMX) format.

The statistical data exchange will be completed with metadata that facilitate interpretation of the numerical data, including country-specific definitions plus metadata on the data sources and on methodological issues.

1.5 OBSERVATION OF THE CENSUS FIELDWORK[6]

As mentioned in section M, during the enumeration period of the 2011 PHC of Albania, a team of national and international experts was set up by the Technical Assistance Project to assist INSTAT in the field activities. In agreement with INSTAT, the primary aims of this activity were to support the coordination and monitoring of the field operations, to support enumerators, controllers and supervisors in the enumeration activities and, generally, to assure the proper implementation of the Census methodology as designed by INSTAT jointly with the EU Technical Assistance Project.

The team of experts consisted of 6 national and 13 international experts. Each expert was assigned to one prefecture and, in general, participated in the activity for 10 working days. In view of the large population concentration in Tirana, several experts were assigned to this prefecture. National experts were teamed-up with international experts to transfer expertise. The field assistance extended to all 36 districts of Albania. During the field support activity, the team of short-term experts was complemented by the EU Technical Assistance Project Team Leader and the Key Expert on Census Methodology and Logistics.

A summary of the fieldwork observation is presented below.

Training and Census materials

Almost without exception, the experts reported that the training received by supervisors and controllers was regarded by them as good to very good. However, project experts reported that some observed controllers may have needed more training and supervision. The most common reported issue was that controllers did not monitor and correct regularly Census forms made by enumerators. This largely also applied to the training of enumerators, although some cases were of less well-trained enumerators. This may have been related to staff replacement during the training or enumeration period. The only area for which the need for more extensive training was mentioned repeatedly was cartography and related topics.

The assessment of the logistic preparation of the fieldwork in terms of providing timely and complete field materials was very favorable. All Census offices and field staff had received their materials on time. There were no reports of Enumerator boxes being mixed-up, although a few had an incorrect number of questionnaires. In addition, about 5,000 questionnaires needed to be replaced due to misprinting, wrong stapling or damage. In the perspective of the overall numbers these issues can be considered negligible.

An additional supply or a redistribution of questionnaires has been necessary in several hundreds of EAs. In general, INSTAT managed to provide these new supplies in time, although there were a few reports of delays in enumeration areas due to a lack of questionnaires.

Field staff

Various reports were made about field staff dropping out: this was particularly prevalent during the training but also noticeable during the enumeration. In most cases, these drop-outs were replaced by persons from the reserve list but some were newly recruited staff. INSTAT District or HQ staff were able to effectively train additional staff in crash courses

[6] This section is based on the 19 reports of the national and international experts selected to support the coordination and monitoring of the field operations.

when necessary. In a few cases the fieldwork was delayed or interrupted by staff drop-out. Generally, these cases were isolated, but in some districts the issue was more structural: during the expert field visits very few of the recruited field staff were found to be lacking. Special attention by INSTAT HQ staff and experts' support managed to get the enumeration on track again in these areas.

The field support experts considered the large majority of the enumerators to be good and committed to their work. A small proportion performed in a substandard way and only a few were totally unacceptable. In general, the quality of enumerators in urban areas seemed to be somewhat better than in rural areas. With regard to the quality of controllers, supervisors and district coordinators, the evaluation was even better, some of them showing exceptionally high working standards, although even here some cases of poor performance were reported. Experts also frequently reported very good and constructive working relationships between enumerators, controllers, supervisors and district coordinators.

On a few districts, problems were reported about field staff who were unclear about payment conditions and whose payments were delayed. The latter was partly due to administrative information or documents not provided by field staff and partly by delays in payments. In a few cases this led to a temporary strike, which, however, did not severely affect the enumeration.

Census offices and office procedures

During the field support missions, over 150 of around 400 Census offices were visited. A great variation was found with regard to the quality of these offices. Whereas most were adequate for the purpose of the Census, quite a number had problems with respect to space, furniture, security and accessibility out of working hours. In several cases new or additional offices were secured or existing offices were upgraded.

The state of the Census office was regularly mentioned as a reason to keep completed questionnaires at the homes of enumerators or controllers. However, also in various other instances, this was standard practice during the first days of the enumeration, which was not in compliance with the procedures laid down by INSTAT. On various occasions the experts have been instrumental in conveying the message that completed questionnaires be returned to and stored in the Census offices. It was observed that subsequently this was effected almost everywhere in the course of the enumeration.

Methodology

The preparation of the Census maps was considered of high quality and very helpful in the field. There were only a very few cases where the borders of EAs raised problems. For specific areas with large new infrastructural constructions or newly built residential areas the maps required substantial updating.

In general, the EA inventory and map update were considered to be successful and effective. However, there were some reports that mentioned inconsistencies between the enumerator maps and the reality on the ground. Also, a number of controllers had misunderstood the updating principles, resulting in incorrect baseline information on the dwellings. This problem was concentrated in a few areas.

According to the field staff and experts, the different manuals were considered helpful and contained a good level of detail. However, it was found that actual consultation in the field was limited and the documentation was considered more training material than practical reference during the fieldwork.

In all districts, the enumeration of collective households only started in week 3 of the enumeration. Consequently, assisting and evaluating that part of the process was largely outside the reach of the experts. Also, during the enumeration period covered by the experts, no efforts were made to enumerate homeless people.

During the enumeration, as a general rule, the Census methodology was followed correctly. Interview technique and correct completion of questionnaires noticeably improved with the build-up of practical experience and upon feedback by controllers. Based on the examination of around two thousand questionnaires, the conclusion of the experts was that, overall, the completed forms were of a good standard. Although within the work of 12 thousand enumerators significant variation was observed, only a small minority was considered of unacceptable quality.

It should be noted that careful study of and compliance with the Enumerator manual would have prevented almost all these mistakes. This means that the controllers did not always refer to the manual when corrections to their method of form completion were required.

Supervision and reporting

In the large majority of areas, controllers were able to meet with their enumerators on a daily basis. In more remote areas this was not always possible, but nevertheless contact used to be on a regular basis. Regular face to face contact between controllers and supervisors was usually more difficult. In poorly accessible areas, maintaining regular contact amounted to large workloads for supervisors, which, in a few cases, exceeded manageable proportions. With respect to the communication in the field, the standard provision of SIM cards to all field staff was highly effective and highly valued, since it allowed free-of-charge communication between enumerators, controllers, supervisors, district coordinators and INSTAT staff from HQ.

The experts reported that the standard presence of INSTAT HQ in the districts had a favorable impact on the quality of the enumeration. On several occasions the HQ staff intervened in difficult cases (refusals, relations with authorities, advice to and training of enumerators, controllers and supervisors) and was an effective support to district coordinators.

It was observed that in the first half of the enumeration period in many areas, the focus of supervision was more on administrative issues than on quality checking. The backlog of questionnaire checking that was built up in such areas partly explains the delay in the completion of the enumeration procedures. With support of the experts, the attention of controllers and supervisors was redirected to the latter task where needed, also assisting in identifying inconsistencies that required more profound assessment of the questionnaires. However, it was noticed that most of the controllers did conscientiously check returned questionnaires on a daily basis and in accordance with the developed methodology.

The Census monitoring tool, based on establishing a baseline of dwellings to be enumerated immediately prior to the enumeration and a daily SMS reporting system for the number of dwellings, households and persons enumerated, as well as the number of refusals encountered, largely appeared to be informative and enabled INSTAT to keep track of the process.

In parallel to the SMS reporting, daily progress was communicated up the Census hierarchy using the daily summary forms and mobile communication. This system suffered from some of the above-mentioned problems as well. Results of both reporting systems showed a relatively small but significant inconsistency.

Relations and refusals

With a few exceptions, the Census publicity campaign lacked visibility: posters were not seen in the streets and often not even more around the Census offices. Nevertheless, the major part of the households was aware of the Census and was very willing to participate. Reportedly, most of the population was effectively reached by radio and TV messages, although the particulars about why the Census was taking place and what happened to the Census data was less well-known.

Support by local authorities, and in particular local Census Commissions, varied across the municipalities and communes. Mostly, authorities were supportive, ranging from passive support to active engagement and forward-thinking attitudes. Only in a few cases, experts received reports of lacking support.

As far as refusals are concerned, the review of the experts' assessments reveals no structural problems. In addition, common problems of lower response rates in urban areas occurred in some areas of Tirana. Census staff managed to resolve a good number of initial refusals and the opportunity offered to the public to be enumerated in the extension week apparently recorded another fair number.

Conclusions

According to the general opinion of the experts and as far as the field observations allow, the Albania 2011 PHC meets international standards and can be considered a success with regard to the implementation of the enumeration, which reflects also the preparatory work carried out by INSTAT with the support of the EU Technical Assistance Project.

Various field problems were identified, but, in relation to the number of staff involved and the overall scale of the operation, these are not expected to have any major implications for the Census results. Moreover, in most instances, INSTAT seemed to have been able to readily resolve such issues.

1.6 OBSERVATION OF THE PES FIELDWORK[7]

As mentioned in section O, during the enumeration period of the PES, a team of national and international experts was set up by the Technical Assistance Project to assist INSTAT in the field activities. In agreement with INSTAT, the primary aims of this activity were to support INSTAT to ensure the independence of the PES from the Census operation, support and monitor training activities, support and monitor PES data collection operations providing methodological inputs to field staff in order to ensure an high quality of the data collected by enumerators.

A summary of the PES fieldwork observation is presented below.

Training

The training of the interviewers and supervisors took place in Durres, from 11-15 November 2011. The training programme was well prepared and thought through, being an incremental progression of knowledge and application. The general opinion of the experts is that the training was generally effective and the delivery of information was appropriate to the individuals taking part, being a mix of older people and students.

A specific session was held for the supervisors and the INSTAT staff that was involved in the post enumeration survey. This was effective in summarizing the information required in the survey and delivering the requirements for the supervisors in assisting with map updating, building unit identification and general checking of the questionnaires when completed.

The closeness of the post enumeration survey to the Census and the use of the same staff for both have led on several occasions to last minute planning. This inevitably leads to a sub-optimal use of resources and sub-prime decision making as time availability forces instant decisions.

Some interviewers did not attend all of the training sessions, and thus may not be fully equipped to be effective in all situations. They do, however, have recourse to supervisors who only had three interviewers to manage.

Fieldwork

The field support experts considered that the post enumeration survey team at INSTAT were very dedicated and, as most had also worked on the Census, continued to give beyond what could normally be expected of staff despite not having fully been refreshed after the Census. Notwithstanding the short time available to prepare and print all questionnaires and the associated documents, INSTAT staff not only provided the materials in time for the training, but also the maps and forms were available for the start of the survey.

Any initial fears that the public would not co-operate with a second demand for information on the household was proved unfounded. The response has been higher than expected for an ad-hoc survey and provided a good basis for the comparison with the data in the Census. The standard of interviewing was high and none of the respondents had any issues with the fact that the data had already been collected in the Census. The reported refusals in the survey appear to be even less than in the Census.

Most of the surveys were conducted on the doorstep. This shortened interview time and allowed much more time for the interviewers to visit more units. It was also clear from discussions with the interviewers that the Interviewer Report Book did not reflect the survey on the ground in that many 'difficult to contact' units were visited many more than the three visits specified in the manual and allowed for in the Report Book.

During the field support missions many questionnaires were examined and only minor errors were identified. These were either going to be corrected prior to the close of the survey or via the data cleaning process. The majority of the data could be directly compared with the Census information for the enumeration areas chosen.

Buildings that had been built since the map photographs were taken were allocated numbers continuing the building sequence. However, as the additional buildings in the enumeration areas were allocated numbers by the interviewers as they came across them, they was not necessarily in the same sequence as in the Census: this entail some manual intervention during the match of the records.

The post enumeration survey also contacted people in many apparently non-residential buildings to identify whether any individuals lived on the premises. These people would probably have been missed in the Census as no strict monitoring of enumerators to investigate at non-residential premises was made in the Census.

The day to day oversight of the survey in the field seemed to be over-managed. Not only did each interviewer have a

[7] This section is based on "Report on the Post Enumeration Survey, 2011, Republic of Albania, Institute of Statistics" prepared by the international expert Ed Swires-Hennessy and on the mission report of the international expert Ian White, summarizing his observation of the PES fieldwork.

supervisor (basically responsible for three interviewers), the entire post enumeration survey team was on the road for the main two weeks of the survey. The effectiveness of this additional oversight may be questioned: indeed with such a short survey, which had an extensive training period for the interviewers, it was probably unnecessary.

Conclusions

According to the general opinion of the experts and as far as the field observations allow, the post enumeration survey had good preparation and, on the basis of observation, was generally executed effectively and efficiently, though late planning at various stages and the use of the same staff as in the Census were noted. The survey was well managed in the field with much attention paid to the detail of the operation by INSTAT staff and supervisors.

The survey was conducted to international standards and the processing supervised by the international experts. Some issues were identified through the analysis stages, but the additional work to investigate and come to a solution was undertaken.

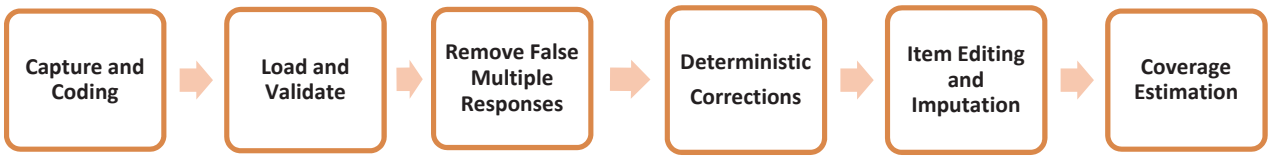
2. QUALITY OF CENSUS DATA

2.1 BACKGROUND

In October 2011, enumerators were sent out to every residential building in Albania to administer the Population and Housing Census (PHC) questionnaires to the usual resident population. The enumerators compiled the questionnaire in the paper form provided by INSTAT.

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 PHC.

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.

However, it is inevitable that some persons and households are missed in a Census, and these must be estimated to get the final Census population estimates. The coverage estimation and coverage adjustment processes estimated the missed persons.

2.2 PRELIMINARY ISSUES

When dealing with Census, some preliminary problems have to be taken into account with regard to people to be enumerated that do not belong to usual household: people living in Collectivity, in Jail or being visiting an ordinary household.

With regard to Collectivity (as, for instance, hospitals, therapeutic community, monastery, and so on) no particular problem was rising: the individuals have been recorded as belonging to the same institutional household.

With regard to individuals that were in jail during October 2011, a list was provided by the Ministry of Justice contain 4,576 prisoners in the whole Albania. The information provided was: name, surname, father's name, sex, year and place of birth. District, Commune/Municipality, enumeration area and building code, were derived from the GIS system, based on the prison's geographical location.

The remaining variables were imputed according to the following procedure. The Ministry of Justice provided aggregated information about the level of education of prisoners, which was used to select a sample of individuals having the same proportions for the different level of education. From this sample a random imputation to prisoners was made for the unavailable information.

All the prisoners were imputed as not currently attending formal education. Regarding the sensitive questions only the

mother language was imputed, while the information about religion and ethnic cultural group was treated as missing information. In Table 2-1, the distribution of prisoners, according to the level of education is given.

Table 2-1 - Distribution of prisoners according to the level of education

Without diploma	Primary education	Lower secondary (obligatory education)	Upper secondary	Tertiary
6.3	17.0	50.6	23.8	2.3

With regard to the problem given by temporary present persons it is worthwhile to notice that in the Census questionnaires visitors to the households had to be listed in list 3. In total, there were 20,844 visitors.

The dataset of visitors was linked with the individual dataset using the Link King, a SAS/AF application, to identify all the visitors already enumerated in their place of usual residence and not in the household visited. There were 4,691 visitors matching with the individual dataset that have been suppressed.

The 16,153 unmatched visitors were imputed in the data set making use of the information on first name, last name, sex, date of birth and place of usual residence available in list 3. The remaining information was treated with probabilistic imputation.

Regarding the sensitive questions only the mother language was imputed, while the information about religion and ethnic cultural group was treated as missing information.

The visitors were added randomly to households belonging to the same place of usual residence as reported in list 3. They were not imputed as new household not to modify the total number of households in Albania.

Finally the problem of duplication was approached. To detect duplications of Census records Link King has been used. In this software two features are included:

- Exact Record Linkage, which founds all the exact matches between two records
- Probabilistic Record Linkage, which matches records that are not exactly equal but have a high probability of being a duplicate

The fields used for detecting duplications, are those that uniquely identify persons: First Name of the individual, Father Name of the individual, Last name of the individual, Date of Birth, Gender and District.

After the process of scanning and data entry of Census questionnaires in the dataset, there were 2,802,082 records of individual data. This dataset was processed by Link King and the result was a dataset with 49,905 records matched. Exact matching has been deleted from the dataset.

For Probabilistic Record Linkage the result of the linking has been aggregated at the household level and only in the case in which the whole household members were marked as duplication the family was considered a duplicated record and was eliminated from individuals and household datasets. In total there were 4,995 households where all the members were linked by Link King, in total 18105 individuals. The dataset of household had 727,255 records and after deletion, the final household dataset has been 722,260 households, while the individual dataset had 278,3977 records. The results of the described operation are reported in Table 2-2.

Table 2-2 - Distribution of individuals after removal of multiple and false responses

Individuals	Collective	Prisoners	Visitors	Duplications	Total
2,786,491	11,023	4,576	16,153	- 18,105	2,800,138

2.3 EDITING AND IMPUTATION PROCEDURE

2.3.1 Editing principles

As with any other survey, also in the Albanian 2011 PHC it was common for respondents to make errors in providing answers and for enumerators on recording them, resulting in a certain amount of data that were not valid.

Invalid responses include missing, multi-ticks, out of range values and partially answered answers. Referred to as item non-response, this can be unintentional, for example where an enumerator missed to administer a question or tick more than one option where not allowed, or intentional where a respondent either does not know the answer or does not want to provide the answer. The extent of item non-response can vary greatly between questions. Items such as sex and civil status usually have few non-responses while the level of education may have higher non-responses.

It was also common for some in-range values to be considered invalid because they were inconsistent either with other values on the questionnaire, or with auxiliary information or definitions. Referred to as item inconsistency, these errors are detected by validating the data against a set of pre-defined edit rules. For example, the rule which stated that a person aged less than 16 years cannot have a university degree would flag a record where the age is five and the higher title of the study is university degree.

Non-response and inconsistency can lead to bias and inconsistencies in the analysis, so it is important to correct them before using data for further analysis. This is desirable because users do not necessarily have all the information required to be able to estimate non-response. It also prevents different methods of editing and imputation being applied, which could result in incomparable or contradictory estimates.

When a record[8] fails an edit rule it is considered erroneous, but this does not mean that all the items in it are wrong. The problem is to find the minimum number of items that has to be changed in order to ensure that the record satisfies all the edit rules at the same time. This problem is strictly related to the need of keeping as much as possible the information as it is given by the respondents.

Not valid values can be replaced through deterministic or probabilistic corrections. A deterministic correction is generated by constrains, which, if violated, lead to an error we are clearly able to identify. An example of deterministic edit is the case in which at the question "What is your place of birth?" the enumerator wrote "Tirana" in the appropriate box but forgot to tickle the answer "In Albania". In this case, we are clearly able to identify the error, because if a person is born in Tirana is also born in Albania. Deterministic errors can be corrected with a simple if-then procedure.

On the contrary, a probabilistic correction is generated by an edit rule that, if violated, lead to an error that we are not clearly able to identify. An example of probabilistic edit is "civil status is married imply that age is more than 14". When this rule is violated, we know that the situation is wrong, but we are not able to identify clearly where the error is: could be that the age is wrong or could be that the civil status is wrong. In these cases we referred to the approach based on the Fellegi-Holt method, which considers the possibility to find the minimum number of variables to change in order to let that the record satisfies all the edits at the same time. This problem is strictly related to the need to keep as much as possible the information given by the respondents. After the identification of the not corrected variables, their values are corrected by mean of imputation.

To avoid introducing bias, the method of imputing non-response must account for the question structure and the distributional properties of the observed data. It must also take into account the possibility that unrecorded data is not missing completely at random, and that some people are more or less likely to have a valid response. For example, in the Census some questions were more likely to be left blank for children, such as address one year ago, and the employment questions were most often left blank for those over the age of 65.

Donor-based methods are appropriate for this type of data because they can handle categorical and numeric variables simultaneously and if applied correctly, donor imputation will estimate the distributional properties of the data accurately.

In donor-based methods, records with invalid values (recipients) are matched to clean records (donors) based on characteristics observed in both records. Values from the donor record are given to the recipient in order to make it complete and correct. Donor methods easily extend to multivariate applications where several variables are imputed together, allowing the implementation of complex edit constraints. This accounts for the implicit and explicit routing on the questionnaire and ensures that the imputed values satisfy the edit rules.

Imputation has three desirable features. First, like weighting adjustments for total non-response, it aims to reduce biases in

[8] With record we intend here the collection of the different items corresponding to one unit of the target population (i.e. the answers provided by a respondent to the individual questionnaire).

the results arising from non-responses. Second, by assigning values at the micro-level, imputation makes analyses easier to conduct and results easier to present. Third, the results obtained from different analyses are bound to be consistent.

Imputation does, however, have its drawbacks. First, it does not necessarily lead to estimates that are less biased than those obtained from the incomplete data set. Second, there is the risk that analysts may treat the completed data set as if all the data were actual responses, thereby overstating the precision of the estimation. Analysts working with a data set containing imputed values should proceed with caution, and should be aware of the extent of imputation for the variables in their analyses as well as the details of the procedures used.

The multivariate nature of Population and Housing Censuses, with all variables potentially subject to missing data and inconsistencies, suggests the need of a general strategy for handling item non-responses. These techniques are not only applied to Censuses, they are also largely used to ensure consistency of data collected from sample surveys.

2.3.2 Editing and imputation strategy for the PHC

The primary objectives of the 2011 editing and imputation strategy were to produce a complete and consistent database. The following three key principles were adopted: i) impute all missing data (except the voluntary questions) to provide a complete and consistent database; ii) minimize the number of changes to the data; iii) ensure that all the changes made to the observed data do not affect their quality.

The Albanian 2011 PHC editing and imputation strategy included four separate processes:

- *Step 1 - Deterministic imputation:* A deterministic procedure was implemented for the correction of the systematic errors. The deterministic corrections were all applied in situations where the response was deemed to be correct with a high level of certainty, based on the information observed in the record. The main purpose of applying these edits was to improve the coherence of the data going into the further editing steps.
- *Step 2 - Inter-record imputation:* The relation with the reference person in a household it is complex. The answer to this question could be erroneous because missing or because not coherent with other collected information like age, sex and civil status. Moreover, detecting the inconsistencies between the relationship and the structural variables involve intra-record editing rules. For this reason it is not appropriate to deal with it using intra-record imputation or with plain deterministic rules. Indeed, imputation from the most "similar" corrected household it is more powerful in finding coherent values of the structural variables.
- *Step 3 - Intra-record imputation:* aimed to resolve non-responses and inconsistencies within the non-structural variables. The major advantage of this approach is that seeks to minimize the number of changes required to repair a record, thus minimizing changes to observed data.
- *Step 4 - Nuclei imputation:* there is another kind of editing activities that are related to the nuclei inside the households. Also, these relationships can be very complex, and the resolution of inconsistencies in this level requires again an Inter-record imputation. This operation was performed as a last stage so that the consistency of the other variables (especially the structural ones) was already ensured by step 2 and 3.

A good editing and imputation procedure is automatic, objective and reproducible, make an efficient use of the matching fields, ensure that imputed records are internally consistent and have an audit trail for evaluation purposes.

INSTAT staff was guided and trained in statistical methods and specialized software to perform data cleaning and data quality assessment. This exercise resulted in coherent and consistent final Census data, which is essential to the objectives of data relevance, accuracy and coherence, which are three of the main criteria for statistical data quality required by the European Statistics Code of Practice.

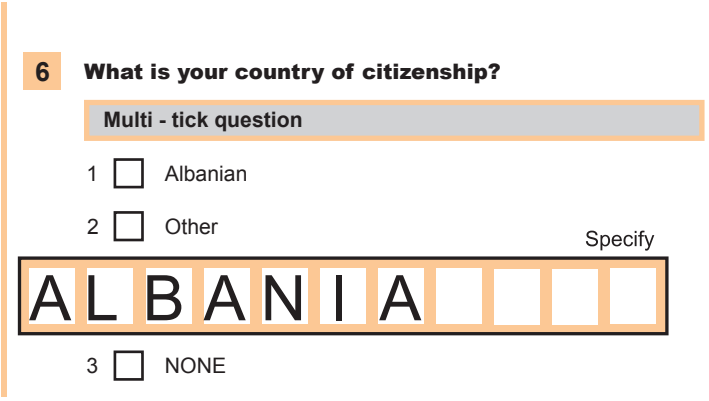
The whole procedure of editing was applied to the Individual dataset. Regarding the Household dataset only step 1 and step 3 were applied while for the Dwelling dataset only step 3 was applied.

2.3.3 Deterministic imputation

Experience from the questionnaire test highlighted a number of common response errors that could easily be addressed with deterministic edit rules. These kinds of errors are normally due to the mistakes of the enumerators that did not fill in the questionnaire properly, to the design of the questionnaire that in some parts resulted not clear enough or to mistakes of the respondent.

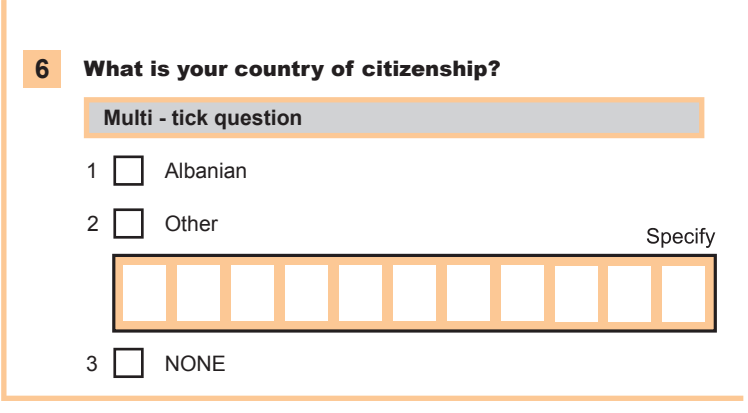
The deterministic edits were all applied in situations where the response was deemed to be correct with a high level of certainty, based on the information observed in the record. The following examples are enlightening in this respect:

Figure 2 - Deterministic errors on Country of Citizenship



As can be easily understood the error highlighted in Figure 2 is due to a mistake of the enumerator that, instead of tickling option 1, wrote in the specify of option 2 the country of citizenship. The instruction on how to fill-in this question was clearly stated in the enumerator's manual, but probably an alternative design of the question as the one illustrate in Figure 3, would have limited much more these situations:

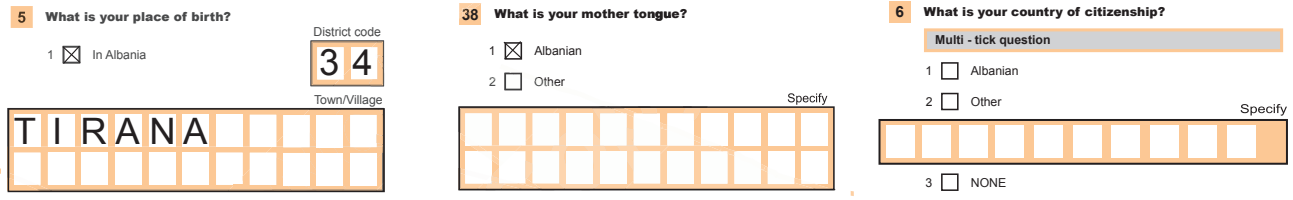
Figure 3 - Alternate design of the question on Country of Citizenship



Unfortunately, there were some constraints on the size of the boxes to be adopted for the OCR System that didn't allow to adopt such solution. The situation illustrated in Figure 2 was detected 22,429 in the Census data, and with more or less the same magnitude, in several other questions having a similar design.

Another relevant case was for the variable Citizenship in which a consistent number of respondents did not answer to the question. This situation is illustrated in Figure 4.

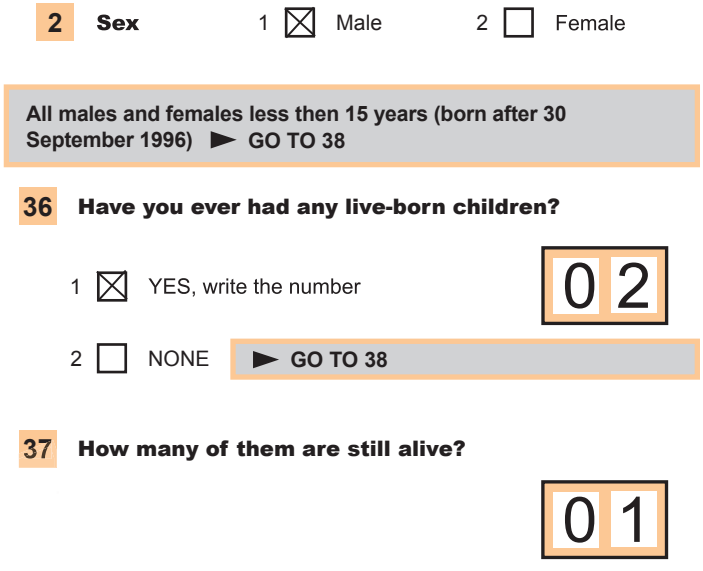
Figure 4 - Deterministic errors on questions regarding Citizenship



The solution adopted was to use, where available, the auxiliary information coming from the questions Place of Birth and Mother Tongue to determine if the person had an Albanian citizenship. The number of changes made by this correction was 283,780.

Another situation that was detected 113,166 times in the Census data was the following:

Figure 5 - Deterministic errors on questions regarding fertility

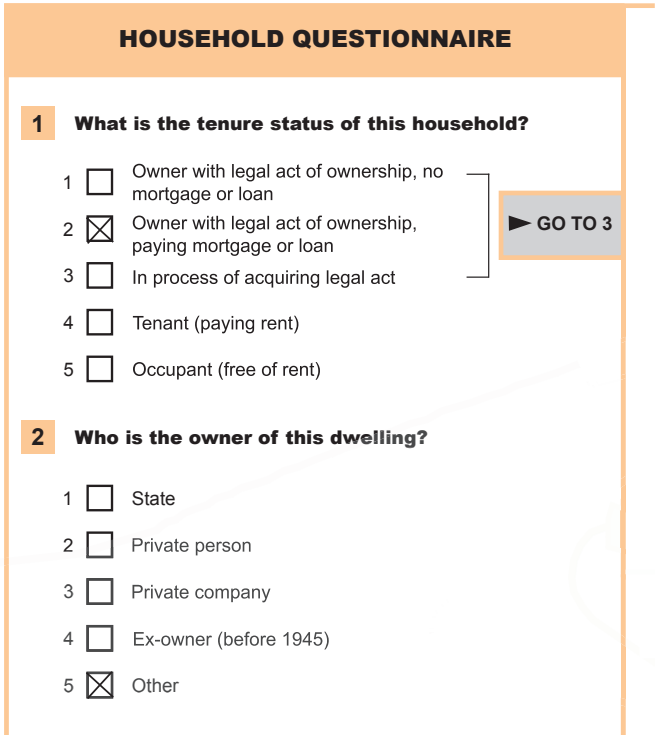


This situation happened when the enumerator reported option 1 (male) on the question about the sex of the respondent and then asked question 36 and 37 to that person. The enumerator's manual clearly stated that these questions should be answered only by Female aged 15 years and more, but probably the routing instruction on the questionnaire created some confusion. Indeed "All male and female less than 15 years" was sometimes intended as "male less than 15 years" and "female less than 15 years". These cases have also been easy to correct since, after having verified the name the sex of the respondent, the solution was just to blank the answer to question 36 and 37.

Concerning the Household questionnaire two were the main deterministic rules applied: one for the variable Ownership of the dwelling and the second for the variable Car or Minivan.

With respect to the Tenure Status in the case of ownership (answers 1, 2 and 3) the Ownership of the dwelling was not supposed to be specified. The situation is shown in Figure 6.

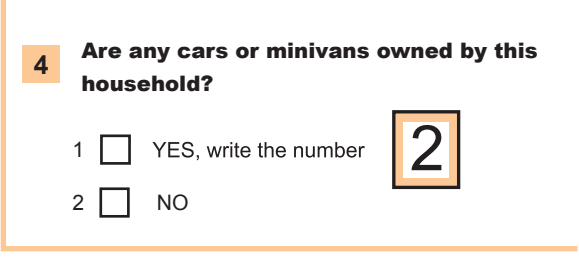
Figure 6 - Deterministic errors on questions regarding Ownership of the dwelling



The Census technical staffs decided to fix these situations blanking the value of the variable Ownership of the dwelling. The cases corrected with this procedure were 48,283.

With respect to the variable Car or Minivan, the situation most relevant is illustrated in Figure 7.

Figure 7 - Deterministic errors on questions regarding Car or Minivan



The deterministic correction applied to solve these situations assigned the value 1 to the variable Car or Minivan. The cases corrected with this procedure were 41,891.

In Annex 1 are listed all the deterministic corrections applied to the Census data.

2.3.4 Inter-record imputation

Relationship with the referent person was one of the most complex questions of the Census. The relationship question can be viewed on page 5 of the 2011 Population and Housing Census Questionnaire.

In the Albanian Population and Housing Census it was asked, to each member, the Relationship to the reference person. The answer to this question could be erroneous because missing answers or because not coherent with other collected information like age, sex and civil status.

This was a concern because relationship needed to be imputed with priority variables like age, sex and civil status. Indeed, high levels of inconsistency or non-response in relationship could affect the quality of the next steps of imputation by reducing the donor pool.

The type of these relations is such that it is not possible to deal with it using the intra-record imputation or using deterministic correction. For this reason, an inter-record imputation was applied.

It should be also noticed that the inter-record imputation can be more powerful in finding a coherent value for a missing field because it is based on the structure of the family. For example, knowing that an individual is the last son of the household can help in giving him an age less or equal to 4 when the one before him is 5 years old.

Unfortunately to implement the inter-record imputation, it was not possible to use any special purpose systems (like CANCEIS from Statistics Canada). Instead, a complex application was developed using the general purpose package ADaMSoft.

The application developed was a script that included three main steps:

- *Check of household coherences:* in this step the values of age, sex and civil status was verified to be coherent with the position of each individual in the household, i.e. the age of the son cannot be greater than the age of the parent minus 13 years;
- *Check of nucleus coherence:* in this step, for each family nucleus, was verified that the relation with the head of the household respected some pre-defined rules;
- *Imputation of the not valid values:* from the above steps was possible to define two kinds of data sets: one composed of households with all the information coherent and a second with inconsistencies in relationship with the referent person or in other structural variables. The individuals in this last dataset were then treated by identifying the most "similar" coherent household (considering the variables Relationship to the Reference Person, age, sex and marital status) and their erroneous values were replaced by the values of the correspondent individuals of the donor household.

For what concerns the coherence between *age, sex, civil status and relation with the reference person,* the process produced the following results.

Table 2-3 - Distribution of erroneous households by Type of errors

Type of error	No
Total households	56,028
A. More than one RP	2,482
B. RP not married with a married partner	7,056
C. Age of RP not valid because his/her parents	2,182
D. Age of RP not valid because his/her grandparents	803
E. RP married and husband/wife not	8,863
F. More than one husband/wife	2,354
G. Second wife married	1,564
H. Husband/wife with the same sex of RP	4,331
I. RP and partner with the same sex	2,060
L. Age of husband/wife not compatible	1,060
M. Son/daughter less than 13 and married	919
N. Age of son/daughter not valid with RP	7,797
O. Age of grandchild not valid with RP	1,076
P. Age of mother/father not compatible with grandparents	154
Q. Age of RP not compatible with son/daughter	1,410
R. Age of husband/wife not compatible with son/daughter	11,748
S. Age of mother/father not compatible with RP	169

As shown in table 2-3, there were 56,028 households with at least one not coherent value of age, sex, civil status and relation to the reference person that brings to a sub-population of 521,465 individuals.

In terms of changes in the variables of age, sex, civil status and relation to the reference person, the procedure localized the following errors.

Table 2-4 – Distribution of non-coherent values for age, sex, civil status and relation to the reference person in the given erroneous households

Variable	No
Age	40,066
Sex	39,282
Civil status	17,839
Relation with the RP	11,954

It should be noted that the greater sources of not valid values were for the variable age, followed by sex and civil status.

2.3.5 Intra-record imputation

A general approach for handling item non-response is generally based on three components: i) the definition of a system of consistencies constrains, called *edit rules*; ii) the verification of them on the given data, known as *error localization*; iii) the correction of values that violate the defined set of edit rules, referred as *imputation*.

Usually the edits rules are derived from the structure of the questionnaire (i.e. people over 15 years have to answer the employment questions) or using some external information (i.e. age should be between 0 and 110 years) or from a logical

relation between two or more variables (i.e. a youth of 12 years cannot have a bachelor degree) [9]. Once the edit rules are specified *(explicit rules)* others can be derived considering them altogether *(implicit rules)*[10]. The whole set of explicit and an implicit rule are applied to detect the erroneous data.

The list of the edit rules defined for Individual, Household and Dwelling datasets are available in *Annex 3*.

The editing methodology used to find the minimum number of items for each erroneous record that need to be corrected in order to ensure that the record itself satisfies all the edit rules at the same time is called Errors Localization Problem[11].

Once the errors are localized, the nearest neighbor hot-deck imputation[12] technique is used to correct the data. This is a method of imputation whereby the deck of donors is built from the set of records that shows no violations of the edit rules (correct records). To make the imputation procedure more precise, the method looks for the minimum distance between the erroneous record and records in the donor's deck. Values of variables from the deck of donors are then used to impute item non-responses for the erroneous records.

First a *restricted joint imputation* is attempted: the donor's deck is searched for a record having the same values of the exact variables of the record to be imputed. In the case in which there are several donors that satisfy this condition, a random selection of one of them is performed. If such donor is found all the erroneous values are imputed from it.

If it is not possible to find such a record in the donor's deck an *enlarged joint imputation* is tried: given the values of the exact variables in the record to be imputed a distance from the correspondent fields in each record of the deck of donors is calculated. On the bases of this distance, the method finds the nearest donor to the erroneous record. In the case in which there are several donors that have the same distance value from the record to be imputed, a random selection of one of them is performed. If such donor is found all the erroneous values are imputed from it.

Finally, if also this procedure fails, *a sequential imputation* is performed: the erroneous values are imputed one by one, each time searching for compatible record in the donor's deck.

2.3.6 Nuclei imputation

The Family nucleus is another complex question in the Census questionnaire. It was a matrix style question that collected the nucleus belonging to each member of the household. The enumerator had to decide by himself regarding this variable, taking in mind the concept of family nuclei. It is to be underlined that the concept of "nuclei" can be difficult to understand and this can lead to a high level of inconsistency or non-response in the answers.

For these reasons it was needed that after deterministic, Inter-record and Intra-record imputation a further imputation step had to be applied. To determine the incorrect cases, other variables like Relationship to the Reference Person, age, sex and marital status were also being taken in consideration. It is to be underlined that at this stage the consistency of these variables was already ensured by step 2 and 3.

Options where a group of persons constitutes a family nucleolus are:

- Couples (even in cohabitation) with one or more children that live in the same household and have no family of their own in the same household
- Couples (even in cohabitation) without children in the same household
- Single parents with one or more children having no family of their own in the same household

From the above rules, the valid prototypes of nuclei were derived. The list of the valid prototype of nucleus it is available in *Annex 2*. From this list were defined two kinds of data sets: one composed by households with all the members having a consistent value of the family nucleus and the other with household having not all the members with valid data for the family nucleus. The individuals having wrong nucleolus were then treated by identifying the most "similar" corrected household (considering the variables Relationship to the Reference Person, age, sex and marital status) and their erroneous values were imputed from the correspondent member of the corrected household.

The script implemented, corrected a considerable number of nucleolus: from 67,301 Not-Valid Nuclei the result after the imputation was 32,940 not valid nuclei.

[9] Edit rules includes: i) logical edits - ensure that two or more data items do not have contradictory values; ii) consistency edits - check to ensure that precise and correct arithmetic relationships exists between two or more data items; iii) range edits - identify whether or not a data item value falls inside a determined acceptable range.

[10] Let's consider the following explicit edit rules: i) the person listed as 1 should be the head of the household; ii) the head of the household should be more than 13 years old. An evident implicit rule that can be derived from these two rules is: iii) the person listed as 1 should be more than 13 years old.

[11] Fellegi and Holt (1976) discuss methods for data editing and imputation with large surveys like the Population and Housing Censuses.

[12] Nearest-Neighbor Hot-Deck Imputation define a distance measure between observations, and impute the value of a respondent who is "closest" to the person with the missing item, where closeness is defined using the distance function.

A final deterministic step was applied in order to fix a number of common response errors that could be easily addressed with deterministic edits. The result of deterministic corrections on nuclei was considerable, from 32,940 not valid family nuclei remained only 21,069 not valid nuclei.

In the main Census results these nuclei were aggregated under the category "Not valid nucleus".

2.4 ASSESSMENT OF THE EDITING AND IMPUTATION PROCEDURE

The problem of assessing the impact of a data cleaning procedure consists in comparing the initial set of data with the set of coherent, clean data obtained applying the whole edit and imputation procedure.

In order to allow this assessment, indicators at aggregate level, which take into account the changes produced in terms of number and/or magnitude, have been considered. Indicators at variables level to underline the differences produced in the distributions of each variable were also considered.

2.4.1 Data quality assessment at aggregated level

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

1. Indicators on the amount of data submitted to the imputation procedure, like *Number of Records, Number of Variables, Number of Variables subject to the Imputation procedure* and *Number of Total Values*.
2. Indicators for the evaluation of the overall effects of the imputation procedure, like: i) *Imputation rate*[13] (I): (Number of Imputed values/Number of Total values)*100; ii) *Addition rate* (Ia): (Number of Additions/ Number of Total values)*100; iii) *Modification rate* (Im): (Number of Modification/Number of Total values)*100; iv) *Elimination rate* (Ie): (Number of Eliminations/Number of Total values)*100.
3. 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%.

[13] As it is easy to understand, the Imputation Rate is the sum of the Addition, Modification and Elimination rates.

Table 2-5 - Quality assessment indicators at aggregate level

	Individual dataset	Household dataset	Dwelling dataset
Number of Records	2,800,138	722,262	1,021,332
Number of Variables	66	30	9
Number of Total values	184,809,108	21,667,860	9,191,988
Number of Valid values	**180,558,499**	**21,376,712**	**9,003,207**
Number of Valid blanks	79,399,119	13,568,117	1,334,076
Number of Valid non-blanks	101,159,380	7,808,595	7,669,131
Number of Imputed values	**4,250,609**	**291,148**	**188,781**
Number of Additions	2,180,747	236,641	117,126
Number of Eliminations	1,201,259	49,588	7,272
Number of Modification	868,603	4,919	64,383
Imputation rate (I)	**2.3**	**1.3**	**2.1**
Additions rate (Ia)	1.2	1.1	1.3
Elimination rate (Ie)	0.6	0.2	0.1
Modification rate (Im)	0.5	0	0.7
Non-Imputation rate	**97.7**	**98.7**	**98**
% of records with I greater than 2%	**30.9**	**16**	**14.2**
% of records with I greater than 5%	**11.5**	**5.2**	**14.2**

From the indicators shown in Table 2-5 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 2.3%. Synthetic indicators on the imputation rate by records also confirm the good performance of the imputation process with an acceptable percentage of records having an imputation rate greater than 5%, which is always less than 14.2%.

2.4.2 Data quality assessment at variable level

The indicators that we are going to present in this context should be necessarily linked to the different kind of variables considered. In particular, it is possible, to construct indexes to evaluate the following points:

- Rate of modification of the initial values;
- Degree of change induced by imputation on the marginal distributions.

For what concerns the rate of modification of the initial values, the four indicators already mentioned in section 2.4.1 were considered, but, this time, at the level of each variable.

To assess the degree of change induced by imputation on the marginal distribution of variables, two main indexes have been considered:

A. Leti's Dissimilarity Index (IM) which will be used for categorical, ordinal and non-ordinal variables:

$$IM = \frac{1}{2}\sum_{k=1}^{K}\left|f_{Y_k} - f_{Y_k}{}^{*}\right|$$

It assumes a 0 value when the two distributions before and after correction are equal (i.e. when the observed relative frequency before correction is equal to the frequency after correction) while it is greater than 0 when they are different; and reaches its maximum value of one when there is maximum dissimilarity between the two distributions (i.e. when both are concentrated in one point which is different for the two of them).

B. Kolmogorov-Smirnov Distance (KS) used for quantitative variables (discrete or continuous) which starting from the two cumulative distribution functions

$$F_{\bar{Y}_n}(t) = \frac{1}{n}\sum_{i=1}^{n} I(\bar{Y}_i \leq t)$$

$$F_{Y_n^*}(t) = \frac{1}{n}\sum_{i=1}^{n} I(Y_n^* \leq t)$$

Before and after correction computes the value

$$KS = max_t\left(\left|F_{\bar{Y}_n}(t) - F_{Y_n^*}(t)\right|\right)$$

Like the Leti's Dissimilarity Index also the Kolmogorof–Smirnov Distance assumes 0 when the two distributions before and after correction are equal while reaches its maximum value of 1 when there is a maximum dissimilarity between the two distributions.

A. INDIVIDUAL DATA

In the following table the rates of modification of the initial values together with the dissimilarity indexes are shown for what concern the Individual dataset for the variables having an Imputation Rate greater than 2%.

Table 2-6 - Variables with Imputation Rate greater than 2% for the Individual dataset

	Imputation Rates				Dissimilarity Indexes	
	I	Ia	Ie	Im	Type	Value
Citizenship: Albanian	**10.98**	10.96	0.01	0.00	IM	**0.110**
Had live-born children	**9.81**	5.47	4.19	0.15	IM	**0.013**
Family nucleus	**9.49**	3.80	0.00	5.69	IM	**0.058**
Completed years of education	**9.06**	1.05	2.24	5.76	KS	**0.012**
Highest completed level of education	**7.55**	1.13	2.05	4.37	IM	**0.022**
Searched for work during the month of September	**5.76**	1.70	3.56	0.50	IM	**0.019**
Place of residence in 2001	**5.72**	4.02	1.38	0.32	IM	**0.026**
Had a job last week of September	**5.31**	1.29	4.02	0.00	IM	**0.027**
Willingness to start job in two weeks	**4.92**	1.33	3.59	0.01	IM	**0.023**
Place of work: Type	**4.57**	3.78	0.29	0.50	IM	**0.035**
Place of residence was changed	**4.13**	2.15	0.14	1.84	IM	**0.020**
Have lived abroad	**4.10**	1.94	1.97	0.19	IM	**0.020**
Worked on last week of September	**4.08**	1.15	2.15	0.78	IM	**0.019**
Employment Status	**4.07**	2.47	0.00	1.60	IM	**0.015**
Currently attending formal education	**3.79**	2.99	0.45	0.34	IM	**0.028**
Place of work	**3.69**	3.23	0.33	0.12	IM	**0.029**
Main reason for not searching for work	**3.41**	0.21	3.18	0.02	IM	**0.030**
Place of stay on CRM	**3.08**	1.25	0.00	1.83	IM	**0.019**
Previous place of usual residence	**2.62**	2.11	0.21	0.30	IM	**0.019**
Civil status	**2.50**	2.12	0.00	0.38	IM	**0.005**
Place of birth	**2.49**	2.34	0.00	0.16	IM	**0.006**
Know to read/write	**2.03**	1.11	0.64	0.27	IM	**0.012**

The imputation made is sometimes higher than 5% and the higher values are in connection with Citizenship (10.98%), Children live born (9.8%), Family nucleus (9.49%), Completed years of education (9.06%) and Highest completed level of education (7.55%).

Concerning the variable Citizenship the imputation rate was high because external factors lead a relevant number of respondents not to answer this question. As mentioned in section 2.3.3, the technical staffs decided to impute this variable using the auxiliary information coming from the questions Place of Birth and Mother Tongue, where available. These situations had a relevant impact also to the Dissimilarity Index that for this question shows its higher value.

The second higher variable imputed was Live Born Children because, as also observed during the fieldwork, some enumerators did not respect to skip the question in case of males respondent. It should be noted that, even if the imputation rate is high, the Dissimilarity Index is having a low value, meaning that the imputation process slightly affected the character's distribution.

Concerning the Family Nucleus it is to be underlined that this question can be very complex and the resolution of the inconsistencies at this level may be difficult. Indeed, as explained in chapter 2.3.6, a high level of wrong nucleus was present in the Census data. Nevertheless, the Dissimilarity Index for this question is of moderate magnitude underlining that the imputation procedure did not change significantly the distribution of the character.

The higher values for education are in the variables Completed Years of Education and Highest Completed Level of Education, which is frequently the case in surveys since the strong correlation between the two characters.

Other moderate levels of imputations are those related to work: Search of work in the month of September (5.76%), Have a job last week of September (5.31%) and Willingness to Start Work in Two Weeks (4.92%). It should be noted that in all of these cases the Dissimilarity Index shows a moderate impact of the imputation on the marginal distribution of the characters.

Concerning the remaining variables all of them present an imputation rate less than 2.00% and a dissimilarity index less than 0.016, underlining very small changes both in the magnitude of changes and in the marginal distributions of the variables.

In the following table the total imputation rate is split into its two components: the deterministic imputation rate (step 1) and the probabilistic imputation rate (step 2-5).

Table 2-7 – Composition of the Imputation Rate: deterministic and probabilistic imputation rate

	Total Imputation	Deterministic imputation	Probabilistic imputation	Deterministic imputation %	Probabilistic imputation %
Citizenship: Albanian	**11.0**	10.4	0.6	94.4	5.6
Had live-born children	**9.8**	3.8	6.1	38.3	61.7
Family nucleus	**9.5**	0.0	9.5	0.0	100.0
Completed years of education	**9.1**	0.0	9.0	0.2	99.8
Highest completed level of education	**7.6**	0.0	7.6	0.0	100.0
Searched for work during September	**5.8**	0.0	5.8	0.0	100.0
Place of residence in 2001	**5.7**	3.2	2.5	56.6	43.4
Had a job last week of September	**5.3**	0.2	5.1	3.7	96.3
Willingness to start job in two weeks	**4.9**	0.0	4.9	0.0	100.0
Place of work: Type	**4.6**	0.1	4.5	1.6	98.4
Place of residence was changed	**4.1**	3.0	1.1	73.4	26.6
Have lived abroad	**4.1**	1.7	2.4	41.5	58.5
Worked on last week of September	**4.1**	0.9	3.2	21.1	78.9
Employment Status	**4.1**	0.9	3.2	21.8	78.2
Currently attending formal education	**3.8**	0.2	3.6	5.5	94.4
Place of work	**3.7**	1.6	2.1	42.8	57.2
Main reason for not searching for work	**3.4**	0.0	3.4	0.0	100.0
Place of stay on CRM	**3.1**	0.1	3.0	3.0	97.0
Previous place of usual residence	**2.6**	2.2	0.4	83.9	16.1
Civil status	**2.5**	0.8	1.7	31.6	68.4
Place of birth	**2.5**	1.4	1.1	57.8	42.2
Know to read/write	**2.0**	0.0	2.0	0.0	100.0

Concerning the variable Citizenship the deterministic imputation rate was about 94.4% of the total imputation rate and only 5.6% of imputation was performed during the probabilistic step. The imputations of Citizenship by deterministic edits were applied in situations where the response was deemed to be correct with a high level of certainty.

The second higher variable imputed in the deterministic step was Previous Place of Usual Residence about 83.9% of imputation rate. The technical staffs decided to impute this variable using the auxiliary information coming from the answer to districts, town/village or country on the same question, where available.

Concerning the variable Place of Residence, 73.4% of imputation rate were coming from the deterministic step. This rate was high because some enumerators did not respect the skip to the question in case of persons less than 1 year old.

For the remaining variables most of the imputation was performed during the probabilistic step. Indeed, this is the

normal way of performing imputation for cases in which the variables are not affected by systematic errors. As already mentioned, using probabilistic imputation ensure that the distribution of the characters would not be significantly altered. This is also confirmed by the values of the Dissimilarity Index in table 2-7.

In Annex 4 are listed all the variables of the Individual datasets and for each of them is reported the Imputation Rate and the Dissimilarity index.

B. HOUSEHOLD DATA

In the following table the degree of change of the initial values with regard to the rate of modification are shown for what concern the Household dataset.

Table 2-8- Imputation Rate and Dissimilarity Index for the Household dataset

	Imputation Rate				Dissimilarity Indexes	
	I	Ia	Ie	Im	Type	Value
Ownership of the dwelling	**7.30**	0.60	6.70	0.00	Im1	**0.060**
Car or Minivans	**5.90**	5.80	0.00	0.10	Im1	**0.012**
Kitchen garden larger than 200 m2	**4.80**	4.80	0.00	0.00	Im1	**0.007**
Tenure status	**4.70**	4.10	0.00	0.60	Im1	**0.010**
None of the previous	**3.90**	3.80	0.10	0.00	Im1	**0.037**
Use of agricultural land	**3.10**	3.10	0.00	0.00	Im1	**0.031**
Livestock or bees	**3.00**	3.00	0.00	0.00	Im1	**0.030**
Income: Paid work or self-employment	**0.90**	0.90	0.00	0.00	Im1	**0.009**
Income: Property or other investments	**0.90**	0.90	0.00	0.00	Im1	**0.009**
Income: Remittances	**0.80**	0.80	0.00	0.00	Im1	**0.008**
Income: Other sources	**0.70**	0.70	0.00	0.00	Im1	**0.007**
Income: Support by another person	**0.70**	0.70	0.00	0.00	Im1	**0.007**
Income: Social assistance and benefits	**0.70**	0.70	0.00	0.00	Im1	**0.007**
Income: Pensions of any type	**0.70**	0.70	0.00	0.00	Im1	**0.007**
Number of Cars or Minivans	**0.60**	0.50	0.10	0.00	KS	**0.005**
Refrigerator	**0.20**	0.20	0.00	0.00	Im1	**0.002**
TV	**0.20**	0.20	0.00	0.00	Im1	**0.002**
Mobile Telephone	**0.20**	0.20	0.00	0.00	Im1	**0.002**
Washing Machine	**0.20**	0.20	0.00	0.00	Im1	**0.002**
Boiler	**0.10**	0.10	0.00	0.00	Im1	**0.001**
Deep Freezer	**0.10**	0.10	0.00	0.00	Im1	**0.001**
TV Decoder	**0.10**	0.10	0.00	0.00	Im1	**0.001**
Computer	**0.10**	0.10	0.00	0.00	Im1	**0.001**
Internet Connection	**0.10**	0.10	0.00	0.00	Im1	**0.001**
Microwave Oven	**0.10**	0.10	0.00	0.00	Im1	**0.001**
Fixed Telephone	**0.10**	0.10	0.00	0.00	Im1	**0.001**
Air Conditioner	**0.10**	0.10	0.00	0.00	Im1	**0.001**
Dress Dryer	**0.00**	0.00	0.00	0.00	Im1	**0.000**
Solar Panel	**0.00**	0.00	0.00	0.00	Im1	**0.000**
Dish Washer	**0.00**	0.00	0.00	0.00	Im1	**0.000**

Here again, the bigger percentage of imputation is linked to the most complicated answers: Ownership of the dwelling

(7.3 %), Car or Minivans (5.9 %), Kitchen garden larger than 200 m2 (4.8 %), Tenure Status (4.7 %), None of the previous (3.9%), Use of agricultural land (3.1%) and Livestock or bees (3.0%).

It should be emphasized that with the exception of Ownership of the dwelling (0.06), the dissimilarity index is always very low, showing that few changes have been introduced in the distribution of the phenomenon under exam.

Table 2-9 – Composition of the Imputation Rate: deterministic and probabilistic imputation rate

	Total Imputation	Deterministic imputation	Probabilistic imputation	Deterministic imputation %	Probabilistic imputation %
Ownership of the dwelling	7.3	6.7	0.6	91.4	8.6
Car or Minivans	5.9	5.8	0.1	100.0	0.0
Kitchen garden larger than 200 m2	4.8	0.0	4.8	0.0	100.0
Tenure status	4.7	0.0	4.7	0.0	100.0
None of the previous	3.9	3.9	0.0	100.0	0.0
Use of agricultural land	3.1	0.0	3.1	0.0	100.0
Livestock or bees	3.0	0.0	3.0	0.0	100.0
Income: Paid work or self-employment	0.9	0.0	0.9	0.0	100.0
Income: Property or other investments	0.9	0.0	0.9	0.0	100.0
Income: Remittances	0.8	0.0	0.8	0.0	100.0
Income: Other sources	0.7	0.0	0.7	0.0	100.0
Income: Support by another person	0.7	0.0	0.7	0.0	100.0
Income: Social assistance and benefits	0.7	0.0	0.7	0.0	100.0
Income: Pensions of any type	0.7	0.0	0.7	0.0	100.0
Number of Cars or Minivans	0.6	0.1	0.5	18.8	81.2
Refrigerator	0.2	0.0	0.2	0.0	100.0
TV	0.2	0.0	0.2	0.1	99.9
Mobile Telephone	0.2	0.0	0.2	0.1	99.9
Washing Machine	0.2	0.0	0.2	0.0	100.0
Boiler	0.1	0.0	0.1	0.0	100.0
Deep Freezer	0.1	0.0	0.1	0.0	100.0
TV Decoder	0.1	0.0	0.1	0.0	100.0
Computer	0.1	0.0	0.1	0.0	100.0
Internet Connection	0.1	0.0	0.1	0.0	100.0
Microwave Oven	0.1	0.0	0.1	0.0	100.0
Fixed Telephone	0.1	0.0	0.1	0.0	100.0
Air Conditioner	0.1	0.0	0.1	0.2	99.8
Dress Dryer	0.0	0.0	0.0	0.0	100.0
Solar Panel	0.0	0.0	0.0	0.0	100.0
Dish Washer	0.0	0.0	0.0	0.0	100.0

Concerning the variable Car or Minivans and None of the previous it should be underlined that all the changes were made during the deterministic imputation. The second higher variable imputed during the deterministic step was Ownership of the dwelling for which 91.4 % and of the imputation rate were performed during the deterministic step.

C. DWELLING DATA

In the following table the degree of change of the initial values with regard to the rate of modification are shown for what concern the Dwelling dataset.

Table 2-10 - Imputation Rate and Dissimilarity Index for the Dwelling dataset

	Imputation Rate				Dissimilarity Indexes	
	I	Ia	Ie	Im	Type	Value
Occupancy status of the dwelling	**7.44**	3.73	0.39	3.32	IM	**0.037**
Type of dwelling	**3.72**	3.00	0.00	0.73	IM	**0.005**
Main type of heating	**1.80**	0.83	0.05	0.92	IM	**0.011**
Have a fixed bath or shower	**1.26**	1.26	0.00	0.00	IM	**0.013**
Total surface of these rooms	**1.18**	0.59	0.11	0.47	KS	**0.007**
Main type of energy used for heating	**1.01**	0.27	0.12	0.62	IM	**0.005**
Type of water supply system available	**0.98**	0.98	0.00	0.00	IM	**0.010**
Total number of rooms used for living	**0.70**	0.42	0.04	0.24	KS	**0.005**
Type of toilet is in use	**0.41**	0.41	0.00	0.00	IM	**0.004**

All the Imputation rates in Table 2-10 are very low except for Occupancy status of the dwelling and for Type of dwelling for which they are respectively 7.44% and 3.72 %, mostly because the complexity of both of the questions. Moreover, the Dissimilarity Indexes are very low, indicating a low difference between the distributions before and after the imputation.

2.5 CENSUS EVALUATION: THE POST ENUMERATION SURVEY

Even though the word 'Census' implies a 100 percent count of the people and housing, it is rarely possible to achieve this. The Conference of European Statisticians (CES) agreed in their preparations for the 2010 and 2011 Censuses in Europe:

1. That the Censuses should be evaluated to check on coverage and the quality of the information provided. Further the European legislation on the Censuses requires Member States to report on the quality of the Census results that they transmit to the European Commission (EUROSTAT)
2. Consideration should be given to coverage and quality of information collected in the Census. INSTAT therefore planned a survey to measure the coverage and quality of the 2011 Census of Albania.

The essential features of the data defined by the CES (individual enumeration, simultaneity, universality within a defined territory, availability of small-area data and defined periodicity) were covered by the Census and the post enumeration survey met all except the availability of small area data: for this latter point, the sample was not large enough to break down the results even by region.

For the first time in the history of Censuses in Albania, the Institute of Statistics of Albania (INSTAT) conducted a post enumeration survey to measure the effectiveness and reliability of the 2011 Census. This Census was by full enumeration of the population using some 12,500 specially recruited and trained enumerators. With such a large field force, a measure of the quality of the results was essential. This will enable a better understanding of the Census results and also provide additional information when using the results in other processes, such as forming a base for the projection of the population of Albania. The post enumeration survey was designed to provide an estimate of the efficiency of the Census process in the identification of both inhabited residential dwellings and the population.

The initial design was for a measure relating to the whole of Albania to produce an achieved sample of 4,000 dwellings. Information was available on the estimated numbers of buildings and dwelling units within each prefecture (Prefecture): each individual enumeration area was also designated as urban or rural according to the administrative division and the summary statistics also available with this categorization. Further, information on survey response rates by urban and rural areas within Prefecture and on the building/non-residential proportions in rural areas was available and used in the sample design.

The sample design was completed in July 2011 and proposed a sample of some 61 enumeration areas, specifying the number of areas to be chosen within each Prefecture by type of area. The required number of individual enumeration areas was then to be selected randomly within Prefecture / type of area by INSTAT.

During the Census operation, concern was expressed by INSTAT with regard to the refusal and non- contact rates in urban areas and a supplementary sample was suggested for urban enumeration areas. Consequently, an additional 44 areas were selected within five municipalities, randomly without replacement, in proportion to the populations in those urban areas.

The questionnaire for the survey was designed by INSTAT with advice from international experts. It was to be a short survey with the minimum number of questions to achieve the aims.

Decisions as to what information was to be included, took into account the possibility of analyzing the resultant information. Hence questions that could not be reliably analyzed from the information collected were not included: for example, the origin of returning migrants to Albania.

The survey itself was undertaken in November 2011 during three weeks. This close proximity to the Census date meant that few changes would have taken place to both the numbers of inhabited dwellings and the population in them.

The starting point for the interviewers was the same maps of the enumeration areas as used for the starting point in the Census with the enumeration area boundaries marked. This ensured that the interviewers did not assume anything about buildings and residential units, but investigated the whole of their area to find residential units.

One difference from the Census dwelling identification procedure was that forms had to be completed for each entrance into a building, even if it appeared to be non-residential. For non-residential units, the interviewers had to ascertain whether such units did contain any residential accommodation as part of the unit or if someone was living there. If nobody was living in the unit, just a brief summary needed to be completed. Supervisors followed up refusals and non–contacts to seek to improve response.

The questionnaires for each enumeration area were collected together into a labeled box. Each interviewer reviewed their questionnaires for consistency and completeness. In the main Tirana office, many hundreds of questionnaires were examined during the three-week collection process by the supervisors: systematic errors were discussed with the interviewers or the supervisors as appropriate. This checking was also undertaken in a small number of other offices by members of the Technical Assistance team.

The forms were optically scanned and a database produced. The data were checked for visual credibility and cleaned using a system that allowed sight of the scanned questionnaire and data entry.

The requirements for credibility checking of the questionnaires were specified by international experts in June 2012 and applied by INSTAT. The cleaned database was available in September.

The matching process was specified in June 2012 and the actual matching of the Census and PES records was undertaken in September. Obviously, in Albania, the matching was made a little more difficult as the dwellings did not have unique explicit addresses. Match codes were applied to the data according to a specification prepared by international experts. An explanation of the matching process adopted can be found in Annex 5.

Tables were then produced giving information for each enumeration area. The examination of the results from the matching exercise gave rise to a few queries. These have now been examined and an issue with the classification of dwellings between no contact and not inhabited was identified, particularly with the Census records in a small number of enumeration areas. As a result, the visits of enumerators to dwellings classified as no contacts were excluded from both data sets and included in the categories of not-inhabited dwellings. The same applied to the Census data.

The examination of the datasets for the selected enumeration areas was in three parts: a comparison of the identified inhabited dwelling units; a comparison of the populations found and a comparison of the age/sex breakdown of the population from both sources.

2.5.1 Estimation of the under-coverage[14]

For each enumeration area, a ratio was calculated of the number of identified dwelling units in PES divided by those identified in the Census. Inhabited dwelling units included those where information was collected and refusals: no contacts were excluded, as it was clear in the analysis that some interviewers had classified empty properties incorrectly as non-contacts. The weighted ratio for Albania was 0.96, which means that the Census identified these units better.

[14] This section is based on "Report on the Post Enumeration Survey, 2011, Republic of Albania, Institute of Statistics" prepared by the international expert Ed Swires-Hennessy

Some 54 per cent of the ratios, however, were over 1.00 indicating that, in over half of the area, the PES identified more units: it is estimated that the under-coverage of the Census was 3.7 per cent in terms of inhabited residential units.

Table 2-11 - Percentage undercount of inhabited dwelling units

Albania	3.7
In urban areas	4.7
In rural areas	2.3

From a population perspective, ratios were calculated for each enumeration area: the total of the populations from the matched records together with the unmatched populations, in PES divided by the total identified in Census.

Table 2-12 - Percentage undercount of population

Albania	3.0
In urban areas	4.0
In rural areas	1.8

The estimate of under-coverage in terms of population was 3.0%, a little lower than that in inhabited dwelling units. However, when calculating the total population estimate, it is necessary to add an estimate for the refusals.

The PES data cannot give any indication about refusals, but estimation of them can be derived from the Census data. Indeed, in case of refusals the enumerators were instructed to compile the questionnaire with the identification codes of the dwelling, tickling option 4 (refusal) in the field "Questionnaire is completed". The total amount of dwellings from the Census database with this characteristic is 6,765 and, on this base, the number of persons who refused to participate in the Census is estimated at 21,839. If considered together with the estimations coming from PES, the estimated undercount of the population could be considered 3.8%.

Table 2-13 - Percentage undercount of population, including refusals

Population	3.8

The measure of the undercount of population is acceptable and, when taken with the non-response to the Census, is within limits achieved by developed countries. These results should be taken into account in the estimation of the population on a year-to-year basis.

2.5.2 Evaluation of the Census quality based on Census/PES matched records

From a population perspective, it is interesting to analyze the differences in the reported information between the cleaned Census individuals and the PES individuals that were possible to match. This kind of analysis has been conducted for the three main variables available in both Census and PES: Sex, Civil status and Age.

Table 2-14 - Percentages of differences in the Census/PES matched records

	Different %	**Equal %**
Sex	1.3	98.7
Civil Status	3.8	96.2
Age	6.9	93.1

As it is shown in Table 16, 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. In this view, it seems adequate assuming that mistakes in PES and Census were equally distributed. Under this assumption an indicator of the total error of the variables can be calculated as:

$$TE = 1/2 \text{ (\% of differences)}$$

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

$$TQ = 100 - TE$$

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

Table 2-15 - Indicators of Total error and Total quality for Sex, Civil status and Age

	TE	TQ
Sex	0.7	99.3
Civil Status	1.9	98.1
Age	3.5	96.5

The results illustrated in Table 2-15 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 the 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.

Table 2-16- Percentages of differences in the geo-localization of the household comparing Census/PES matched records

Geo Localization	%
Different BL	16.49
Same BL	83.51

The indicators TE and TQ related to the Geo Localization are shown in Table 2-16 under the usual assumption that mistakes in PES and Census were equally distributed.,

Table 2-17 - Indicators of Total error and Total quality in the geo-localization of the household comparing Census/PES matched records

	TE	TQ
Geo-localization	8.2	91.8

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

3. CONCLUSIONS

Despite the fact that this is the first time that INSTAT implemented a comprehensive system of quality assurance for such a big statistical operation like the PHC and also considered all the limits underlined in this report, the implementation of such a system of quality assurance significantly improved the total quality of the whole Census process. All the critical milestone activities were completed in time, thanks also to improved coordination and cooperation among the different entities involved in the Census.

According to the general opinion of the EU technical assistance experts, the 2011 Population and Housing Census of Albania meets the international standards and can be considered successful with regard to the implementation of the enumeration, which reflects also the preparatory work carried out. Various field problems were identified, but, in relation to the number of staff involved and the overall scale of the operation, these had not any major implications on the Census results. Moreover, in most instances, INSTAT seemed to have been able to readily resolve such issues.

It is also a general opinion of the experts that the post enumeration survey had good preparation and was generally executed effectively and efficiently, though late planning at various stages and the use of the same staff as in the Census were noted. The survey was well managed in the field with much attention paid to the detail of the operation by INSTAT staff and supervisors.

The survey was conducted to international standards and the processing supervised by the international experts. Some issues were identified through the analysis stages, but the additional work to investigate and come to a solution was undertaken. Moreover, the measure of the undercount of population is acceptable and, when taken with the non-response to the Census, is within limits achieved by developed countries.

The overall procedure of edit and imputation of the 2011 Population and Housing Census in Albania is a 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. Overall item editing and imputation was successful in meeting the main objectives and aims outlined in the strategy: a complete and consistent database was achieved with few issues identified.

A full evaluation of item editing and imputation is being undertaken and published. The analysis carried out, based on the comparison between raw and clean data, shows a general low impact of the edit and imputation procedure on the data related to the different items. In a few cases a more relevant impact was observed, but at the same time, the original distributions have been always preserved.

ANNEX 1: DETERMINISTIC CORRECTIONS

A. INDIVIDUAL DATASET

Description	Conditions	Action
In cases in which there was a missing value for the marital status for persons less than 15 years old, according to the Albanian law the status of these individuals was corrected to "Single".	ind_04=. and age<15 and age ne .	ind_04=1
There were cases in which the answer to place of birth was missing, but in the same question the answer was given by mistake as "Albania" in the "other specifies" question for country. In those cases "In Albania" should be chosen as the place of birth and the code of the country should be left blank.	ind_05_05_Code=8	ind_05_01=1; ind_05_05_Code=.
In cases when the place of birth was missing, but the code of "District" and "Town/Village" was not empty and the code on the country is empty, then the alternative 'In Albania" should be chosen as the place of birth.	ind_05_01=. and ind_05_02 ne . and ind_05_03_code ne . and ind_05_05_code=.	ind_05_01=1;
In cases when the place of birth was missing and there was not value in "District" and "Town/Village" and there was an answer in "Country", then the place of birth should be "Abroad".	ind_05_01=. and ind_05_02=. and ind_05_03_code=. and ind_05_05_code ne .	ind_05_01=2
In the question for citizenship there were cases in which there was an answer as "Albanian" in the country specification. The alternative "Albanian" should be chosen and the alternative "Other" and the code of the country of citizenship should be empty.	ind_06_02_Code=8	ind_06_01=1; ind_06_01_2=.; ind_06_02_Code=.;
There were a lot of cases in which the answers to questions of citizenship were missing. If for those individuals the place of birth was Albania then the citizenship was supposed "Albanian".	ind_06_01=. and ind_06_02_Code=. and ind_06_01_3=. and ind_05_01=1	ind_06_01=1;
If in the question of "Place of residence at the Moment of Census" the answer to the specification of the country was "Albania" and the code of the District was not empty than the alternative "Elsewhere in Albania" was supposed as the right answer.	ind_07_05_Code=8 and ind_07_02 ne .	ind_07_05_Code=.; ind_07_01=2;

Description	Conditions	Action
If in the question of "Place of residence at the Moment of Census" the answer to the specification of the country was "Albania" and the code of the District was empty than the alternative "The same town/village" was supposed as the right answer.	ind_07_05_Code=8 and ind_07_02= .	ind_07_05_Code=.; ind_07_01=1;
If the answer to question "Place of residence at the Moment of Census" is missing and the code of the District and Town/village is not missing and the code of the Country is missing then the alternative "Elsewhere in Albania" was supposed as the right answer.	ind_07_01=. and ind_07_02 ne . and ind_07_03_Code ne . and ind_07_05_ Code=.	ind_07_01=2;
If the answer to question "Place of residence at the Moment of Census" is missing and the code of the District and Town/village is missing and the code of the Country is not missing then the alternative "Abroad" was supposed as the right answer.	ind_07_01=. and ind_07_02=. and ind_07_03_Code=. and ind_07_05_ Code ne .	ind_07_01=3;
There were cases, in which in the question of "Place of residence in Census 2001" the specification for the country was "Albania". In these cases the country code should be empty and the alternative "In Albania" should be chosen.	ind_09_04_Code=8	ind_09_01=1; ind_09_04_Code=.;
If the answer to question "Place of residence in Census 2001" is missing and the code of the District and Town/village is not missing and the code of the Country is missing then the alternative "In Albania" was supposed as the right answer.	ind_09_01=. and ind_09_02 ne . and ind_09_03_Code ne . and ind_09_04_ Code=.	ind_09_01=1;
If the answer to question "Place of residence in Census 2001" is missing and the code of the District and Town/village is missing and the code of the Country is not missing then the alternative "Abroad" was supposed as the right answer.	ind_09_01=. And ind_09_02=. and ind_09_03_Code=. and ind_09_04_ Code ne .	ind_09_01=2;
If the question of "Last place of usual residence" is missing and the codes of the District and Town/Village are not missing, then the alternative "Somewhere else in Albania" is supposed as the right answer.	ind_11_01=. And ind_11_02 ne . and ind_11_03_Code ne .	ind_11_01=1;

Description	Conditions	Action
If the person has moved more than once in the question of "Last place of usual residence" should be specified the last movement. If the movement abroad (year in ind_16) is done after the movement specified in ind_12 then the answer to question "Last place of usual residence" should be "Abroad" and all the other questions till ind_15 should be empty.	ind_16>ind_12 and ind_16 ne . and ind_12 ne .	ind_11_01=2; ind_11_02=.; ind_12=.; ind_13=.; ind_14=.;
If the year to the last arrival in Albania is more than 2001 and the answer to the question ind_10 about changes of place of usual residence since 2001 is "No" then the answer to this question should be "Yes".	ind_16>=2001 and ind_10=2	ind_10=1;
If the part of internal movement is missing, then the question referring the change of usual residence since 2001 should be "No".	ind_10=. and ind_11_01=. and ind_11_02=. and ind_12=. and ind_13=. and age>0	ind_10=2;
If the part of international movement is missing, then the question referring to ever lived abroad should be "No".	ind_14=. and ind_15_02_Code=. and ind_16=. and ind_17=. and age>0	ind_14=2;
If the answer to question Ind_19 is currently attending school, but the highest completed level of education is "Without diploma" and age is greater 18 years, then the right answer should be "Attended school in the past".	ind_19=1 and ind_20=1 and age>18	ind_19=2;
If the person has answered that is currently attending school, but the highest completed level of education is "Primary education" and age is greater 22 years, then the right answer should be "Attended school in the past".	ind_19=1 and ind_20=2 and age>22	ind_19=2;
If the person has answered that is currently attending school, but the highest completed level of education is "Lower secondary" and age is greater 30 years, then the right answer should be "Attended school in the past".	ind_19=1 and ind_20=3 and age>30	ind_19=2;
If for the question of the place of work in the specification of the Country is written Albania, then the alternative "In Albania" should be chosen and the code of the country should be empty.	ind_32_06_Code=8	ind_32_01=1; ind_32_06_Code=.;

Description	Conditions	Action
There are cases when the place of work is not specified if it is in “Albania” or “Abroad” and it is not specified if it is fixed, at home or not fixed. If the District, Town/Village is specified, then the Place of Work is supposed as "Fixed" and "In Albania".	ind_32_01=. and ind_32_03=. and ind_32_02 ne . and ind_32_04_Code ne . and ind_32_06_Code=.	ind_32_01=1; ind_32_03=1;
There are cases when the place of work is not specified if it is in “Albania” or “Abroad” and it is not specified if is fixed, at home or not fixed. If the District, Town/Village are missing, but the code of the country is not missing, then the place of work is supposed "Abroad".	ind_32_01=. and ind_32_03=. and ind_32_02=. and ind_32_04_Code=. and ind_32_06_Code ne.	ind_32_01=2;
If the place of work is not specified if it is in “Albania” or “Abroad” but it is not specified whether it is fixed, at home or not fixed and the code of the country is empty then the Place of Work is supposed "In Albania".	ind_32_01=. and ind_32_03 ne . and ind_32_06_Code=.	ind_32_01=1;
If the enumerated person is a male he should not have answered the questions referring to the number of children.	IND_02=1 and (IND_36_01 ne . or IND_36_02 ne . or IND_37 ne .)	I N D _ 3 6 _ 0 1 = . ; I N D _ 3 6 _ 0 2 = . ; IND_37=.;
If the person has children still alive, but the answer to the question of number of live-born children is missing, then this number is supposed the same as children still alive.	ind_36_02=. and ind_37 ne .	ind_36_02=ind_37;
There were cases when the number of still alive children was higher than the number of live-born children. The number of live-born children is supposed the same as children still alive.	ind_37>ind_36_02	ind_37=ind_36_02;

B. HOUSEHOLD DATASET

Description	Conditions	Action
Referring to the household questionnaire design, If the tenure status of the household is "owner"(3 first alternatives of Q1) then question 2 should not be asked.	hh_01 in(1,2,3) and hh_02 ne .	hh_02=.
In question 3, if all the household durables were not specified than the alternative" None of these" should be fulfilled if is not.	hh_03_1=. and hh_03_2=. and hh_03_3=. and hh_03_4=. and hh_03_5=. and hh_03_6=. and hh_03_7=. and hh_03_8=. and hh_03_9=. and hh_03_10=. and hh_03_11=. and hh_03_12=. and hh_03_13=. and hh_03_14=. and hh_03_15=. and hh_03_16=.	hh_03_16=16
Question 4 is composed by two answers. If household says that own any car or minivan and the number of cars specified is 0 or missing then is supposed that the household does not own any car.	(hh_04_2=0 or hh_04_2=.) and hh_04_1=1	hh_04_1=2; hh_04_2=.;
If the number of cars given in question 4 (hh_04_2) are greater than 0 then the answer to question hh_04_1 should be "Yes" if there is no answer.	hh_04_2>0 and hh_04_1=.	hh_04_1=1
In cases when there is no answer in the car ownership questions, was supposed that the household does not own any car.	hh_04_2=. and hh_04_1=.	hh_04_1=2

ANNEX 2: LIST OF VALID PROTOTYPES OF NUCLEI

In the following table, each row represents a combination of relation with the reference person pertaining to a valid nucleus, i.e. first row 1 2 4M: it is considered valid a nucleus where there is a person with Relation to the head 1 (Head of household), a person with Relation to head 2 (Husband or wife) and one or many persons with Relation to the head 4 (Son/daughter). The list of the codes for the relation with the reference person is available on the Census Questionnaire.

Person 1	Person 2	Person 3	Person 4
1	2	4M	
1	2		
1	3	4M	
1	3		
4	12		
3	4M		
1	4M		
4	5M		
12	5M		
4	12	5M	
1	6	6	7M
1	6	7M	
6	6	7M	
6	7M		
1	6	6	
1	6		
7	9	8	
7	8		
7	9		
9	8		
13	13	9	
13	13		
13	9		
11	11	6	14
11	6	14	
1	6	10	7M
1	6	10	
1	10	7M	
6	10	7M	
10	7M		
16	15M		
4	16	5M	
4	15	15M	
14	16	15M	
14	15M		
15M			
16M			

ANNEX 3: LIST OF EDIT RULES ADOPTED FOR THE INTRA-RECORD IMPUTATION

A. INDIVIDUAL DATASET

- The first person listed in the household members list should be the "Reference person"
- The reference person of the household should be listed the first in the household members list
- The age of "Reference person" should not be less than 14
- Individuals less than 15 years old, cannot be married, separated, divorced or widowed
- Husband of the wife of the referent person should be married
- If the place of birth is "In Albania", the district code must be specified
- If the place of birth is not "In Albania "the district code shouldn't be specified
- If the place of birth is not "In Albania", Town/Village shouldn't be specified
- If the place of birth is "In Albania" shouldn't be any country specified
- If the place of birth is "Abroad" the Country must be specified
- The individuals cannot have "Albanian" citizenship and "No citizenship" at the same time
- The individuals cannot have "Other" citizenship and "No citizenship" at the same time
- The individuals should answer the question of the citizenship
- If "Other" is specified as citizenship, the specification should be fulfilled
- If "Other" is not specified as citizenship, the specification shouldn't be fulfilled
- If the person in the Moment of Census has been "Elsewhere in Albania", the District Code must be specified
- If the person in the Moment of Census has not been "Elsewhere in Albania", the District Code shouldn't be specified
- If the person in the Moment of Census has not been "Elsewhere in Albania", the Town/Village shouldn't be specified
- If the person in the Moment of Census has been "Abroad", the country must be specified
- If the person in the Moment of Census has not been "Abroad", the country shouldn't be specified
- If the person in the Moment of Census has been in "The same town/village", shouldn't be any answer for the reason of absence from the place of usual residence.
- If the place of usual residence in Census 2001 is "In Albania", the district code must be specified
- If the place of usual residence in Census 2001 is not "In Albania", the district code shouldn't be specified
- If the place of usual residence in Census 2001 is "Abroad" the country must be specified
- If the age of the individual is less than 10 years old the usual residence in Census 2001 shouldn't be fulfilled
- If the age of the individual is 10 years old and the month of birth is from April to December 2001, the place of usual residence in Census 2001 must be specified
- If the age of the individual is 10 years old and the month of birth is from January to March 2001, the place of usual residence in Census 2001 shouldn't be specified
- If the age is more than 10 years old the place of usual residence is Census 2001 must be specified
- If age is 0 years old, questions on migration, education and employment shouldn't be fulfilled
- If age is more than 0 years old the question regarding changing the usual residence since Census 2001 must be fulfilled
- If the individual has not changed the place of residence since Census 2001, questions regarding internal migration shouldn't be fulfilled
- If the individual has not changed the place of residence since Census 2001, question 14 about ever lived abroad for a continuous year or more, must be specified

- If the individual has changed the place of usual residence since Census 2001, the place of the usual residence before coming to the actual one, must be specified.
- If the place of previous usual residence is "In Albania" the District Code must be specified
- If the place of previous usual residence is "Abroad", the District Code shouldn't be specified
- If the District Code is specified the Town/Village must be specified
- If the place of previous usual residence is "In Albania", the year of arriving at the current usual residence must be specified
- If the place of previous usual residence is "In Albania", the reason for changing the usual residence must be specified
- If the place of previous usual residence is "In Albania", question if ever lived abroad for a continuous year or more, must be specified
- If the place of previous usual residence is "Abroad", questions on internal migration shouldn't be fulfilled
- If the place of previous usual residence is "Abroad", questions on emigration must be specified
- If the persons has never emigrated more 12 months, other questions on emigration shouldn't be fulfilled
- If the individual has emigrated more than 12 months, questions on emigration must be specified
- If the individual is less than 6 years old, questions on education and employment shouldn't be specified
- If the age is more than 5 years old question of literacy must be specified
- If the age is more than 5 years old question on formal school attendance must be specified
- If the individual does not know how to read and write the highest completed level of education can be "Primary education"
- If the person never attended school question on highest completed level of education and completed years of education shouldn't be specified
- If the individual is attending school or has attended in the past, he should be more than 10 years old and question on highest completed level of education and completed years of education must be specified
- If the individual is less than 10 years old question on completed level of education, completed years of education and employment shouldn't be specified
- If the individual is less than 15 years old, questions regarding employment shouldn't be specified
- If the individual is more than 15 years old, questions regarding employment during the reference week must be specified
- If the individual has not worked during the reference week, he should state whether he/she is temporarily absent from a job
- If the individual has worked during the reference week, he should state the total number of working hours during this week
- If the individual has specified the number of working hours, questions regarding temporary absence, searching for work, reasons for not searching work and availability to start a work within two weeks, shouldn't be specified.
- If the individual has not worked during the reference week he should answer if he was temporarily absent.
- If the individual is temporarily absent from the job, he should not answer the questions for searching for a work, the reason for not searching for a work and the availability to start a work.
- If the individual is not temporarily absent from a job, he should be asked if he/she was searching for a work during the last month.
- If the individual is temporarily absent from the job, he should be asked for the position in the main job.
- If the individual has not searched a work during the last month, the reason for not searching must be fulfilled
- If the individual has answered the question for searching work during the last month, the position in the job, occupation, main activity and organization name and place of work should be stated.

- If there is an answer to the question regarding availability to start a job, questions for the position in the job, occupation, main activity and organization name and means of transport shouldn't be stated.
- If there is an answer for the hours of work during the reference week, the position in the main job should be stated;
- If the individual is temporarily absent from work, the position in the main job should be stated
- If the position in the main job is fulfilled, the place of work should be stated
- If the position in the main job is not fulfilled, the place of work should not be stated
- If the place of work is "In Albania" the District Code should be fulfilled
- If the place of work is "Abroad" the District Code shouldn't be fulfilled
- If the place of work is "In Albania" and "Fixed work place" the District Code should be fulfilled
- If the place of work is "In Albania" and "Mainly at home" or "No fixed", the District Code should not be fulfilled
- If the place of work is "In Albania" and "Mainly at home" or "No fixed", the Town/Village should not be fulfilled
- If the place of work is "Abroad", the country should be stated
- If the place of work is "In Albania", the country should not be fulfilled
- If the place of work is "In Albania" and "Mainly at home", travel and transport questions should not be stated
- If the place of work is "Fixed workplace, away from home", traveling question should be fulfilled
- If the place of work is "Fixed workplace, away from home", means of transportation for traveling should be fulfilled
- If the individual is male he should not answer the question regarding fertility
- If the individual is less than 15 years old, he should not answer the question regarding fertility
- If the individual is more than 15 years old and it is a female, questions regarding fertility should be fulfilled
- If the individual has children, the number should be stated
- If the individual has not children, the number should not be stated
- If the individual has not children, question about children still alive should not be fulfilled
- If the individual has children, question about children still alive should be fulfilled
- If the individual is "Without diploma", the number of completed years of education should not be more than 5
- If the individual has "Primary education", the number of completed years of education should be more than 3
- If the individual has "Primary education", the number of completed years of education should be less than 9
- If the person has "Lower secondary (obligatory education)" diploma, the number of completed years of education should be more than 6
- If the person has "Lower secondary (obligatory education)" diploma, the number of completed years of education should be less than 13
- If the person has Lower secondary vocational (2-3 years) diploma, the number of completed years of education should be more than 8
- If the person has Lower secondary vocational (2-3 years) diploma, the number of completed years of education should be less than 13
- If the person has Upper secondary (general) diploma, the number of completed years of education should be more than 10
- If the person has Upper secondary (general) diploma, the number of completed years of education should be less than 15
- If the person has Upper secondary technical (4-5 years) diploma, the number of completed years of education should be more than 10
- If the person has Upper secondary technical (4-5 years) diploma, the number of completed years of education should be less than 15
- If the person has Tertiary (BA) diploma, the number of completed years of education should be more than 13

- If the person has Tertiary (BA) diploma, the number of completed years of education should be less than 17
- If the person has Tertiary (BAMA) diploma, the number of completed years of education should be more than 14
- If the person has Tertiary (BAMA) diploma, the number of completed years of education should be less than 18
- If the person has Tertiary (old system before Bologna)diploma, the number of completed years of education should be more than 12
- If the person has Tertiary (old system before Bologna)diploma, the number of completed years of education should be less than 19
- If the person has Post-graduate/Master, the number of completed years of education should be more than 15
- If the person has a Doctorate / PhD, the number of completed years of education should be more than 17
- If the highest completed level of education is "Primary education", the age should be more than 9
- If the highest completed level of education is "Lower secondary", the age should be more than 12
- If the highest completed level of education is "Lower secondary vocational", the age should be more than 14
- If the highest completed level of education is "Upper secondary", the age should be more than 16
- If the highest completed level of education is "Upper secondary technical", the age should be more than 16
- If the highest completed level of education is "Tertiary (BA)", the age should be more than 19
- If the highest completed level of education is "Tertiary (BAMA)", the age should be more than 20
- If the highest completed level of education is "Tertiary (old system before Bologna)", the age should be more than 18
- If the highest completed level of education is "Post-graduate/Master", the age should be more than 21
- If the highest completed level of education is "Doctorate/PhD", the age should be more than 23
- If the individual is 10 years old the number of completed years of education should be less than 5 years
- If the individual is 11 years old the number of completed years of education should be less than 6 years
- If the individual is 12 years old the number of completed years of education should be less than 7 years
- If the individual is 13 years old the number of completed years of education should be less than 8 years
- If the individual is 14 years old the number of completed years of education should be less than 9 years
- If the individual is 15 years old the number of completed years of education should be less than 10 years
- If the individual is 16 years old the number of completed years of education should be less than 11 years
- If the individual is 17 years old the number of completed years of education should be less than 12 years
- If the individual is 18 years old the number of completed years of education should be less than 13 years
- If the individual is 19 years old the number of completed years of education should be less than 14 years
- If the individual is 20 years old the number of completed years of education should be less than 15 years
- If the individual is 21 years old the number of completed years of education should be less than 16 years
- If the individual is 22 years old the number of completed years of education should be less than 17 years
- If the individual is 23 years old the number of completed years of education should be less than 18 years
- If the individual is 24 years old the number of completed years of education should be less than 19 years
- If the individual is 25 years old the number of completed years of education should be less than 20 years
- If the individual is 26 years old the number of completed years of education should be less than 21 years
- If the individual is 27 years old the number of completed years of education should be less than 22 years
- If the individual is 28 years old the number of completed years of education should be less than 23 years
- If the individual is 29 years old the number of completed years of education should be less than 24 years

- If the individual is 30 years old the number of completed years of education should be less than 25 years
- If the individual is 13 years the number of live-born children should be at maximum 1
- If the individual is 14 years old the number of live-born children should be at maximum 1
- If the individual is 15 years old the number of live-born children should be at maximum 2
- If the individual is 16 or 17 years old the number of live-born children should be at maximum 3
- If the individual is 18 years old the number of live-born children should be at maximum 4
- If the individual is 19 or 20 years old the number of live-born children should be at maximum 5
- If the individual is 21 years old the number of live-born children should be at maximum 6
- If the individual is 22 or 23 years old the number of live-born children should be at maximum 7
- If the individual is 24 years old the number of live-born children should be at maximum 8
- If the individual is 25 or 26 years old the number of live-born children should be at maximum 9
- If the individual is 27 years old the number of live-born children should be at maximum 10
- If the individual is 28 or 29 years old the number of live-born children should be at maximum 11
- If the individual is 30 years old the number of live-born children should be at maximum 12
- If the individual is 31 or 32 years old the number of live-born children should be at maximum 13
- If the individual is 33 years old the number of live-born children should be at maximum 14
- If the individual is 34 or 35 years old the number of live-born children should be at maximum 15
- If the individual is 36 years old the number of live-born children should be at maximum 16
- If the individual is 37 or 38 years old the number of live-born children should be at maximum 17
- If the individual is 39 years old the number of live-born children should be at maximum 18
- If the individual is 40 or 41 years old the number of live-born children should be at maximum 19
- If the individual is 13 years old the number of children still alive should be at maximum 1
- If the individual is 14 years old the number of children still alive should be at maximum 1
- If the individual is 15 years old the number of children still alive should be at maximum 2
- If the individual is 16 or 17 years old the number of children still alive should be at maximum 3
- If the individual is 18 years old the number of children still alive should be at maximum 4
- If the individual is 19 or 20 years old the number of children still alive should be at maximum 5
- If the individual is 21 years old the number of children still alive should be at maximum 6
- If the individual is 22 or 23 years old the number of children still alive should be at maximum 7
- If the individual is 24 years old the number of children still alive should be at maximum 8
- If the individual is 25 or 26 years old the number of children still alive should be at maximum 9
- If the individual is 27 years old the number of children still alive should be at maximum 10
- If the individual is 28 or 29 years old the number of children still alive should be at maximum 11
- If the individual is 30 years old the number of children still alive should be at maximum 12
- If the individual is 31 or 32 years old the number of children still alive should be at maximum 13
- If the individual is 33 years old the number of children still alive should be at maximum 14
- If the individual is 34 or 35 years old the number of children still alive should be at maximum 15
- If the individual is 36 years old the number of children still alive should be at maximum 16

- If the individual is 37 or 38 years old the number of children still alive should be at maximum 17
- If the individual is 39 years old the number of children still alive should be at maximum 18
- If the individual is 40 or 41 years old the number of children still alive should be at maximum 19
- If the individual is less than 14 years old the main reason of absence from the usual residence shouldn't be "Employment"
- If the individual is more than 74 years old the main reason of absence from the usual residence shouldn't be "Employment"
- If the individual is less than 5 years old the main reason of absence from the usual residence shouldn't be "Study"
- If the individual is more than 61 years old the main reason of absence from the usual residence shouldn't be "Study"
- If the individual is less than 14 years old the main reason of changing the usual residence shouldn't be "Employment"
- If the individual is more than 74 years old the main reason of changing the usual residence shouldn't be "Employment"
- If the individual is less than 5 years old the main reason of changing the usual residence shouldn't be "Study"
- If the individual is more than 61 years old the main reason of changing the usual residence shouldn't be "Study"
- If the individual is more than 74 years old the main reason for coming to live in Albania shouldn't be "Employment opportunities in Albania"
- If the individual is more than 74 years old the main reason for coming to live in Albania shouldn't be "Finished job abroad"
- If the individual is more than 45 and less than 105 years old the main reason for coming to live in Albania shouldn't be "Study opportunities in Albania"
- If the individual is more than 45 and less than 105 years old the main reason for coming to live in Albania shouldn't be "Finished study abroad"
- If the individual is more than 61 and less than 105 the main reason for not searching for work or trying to start a business shouldn't be "Student/Pupil"
- If the individual is more than 74 and less than 105 years old the main reason for not searching for work or trying to start a business shouldn't be "No work available"
- If the person is currently attending formal education should not be more than 45 years old.

B. HOUSEHOLD DATASET

- If the tenure status of the household is "Owner with a legal act of ownership, no mortgage" or "Owner with legal act, paying mortgage" or "In process of acquiring legal act", the owner of the dwelling shouldn't be specified
- If the tenure status of the household is "Tenant" or "Occupant (free of rent", the owner of the dwelling must be specified
- If the household has "Refrigerator" as any of the amenities the alternative "None" shouldn't be chosen
- If the household has "Deep freezer" as any of the amenities the alternative "None" shouldn't be chosen
- If the household has "Washing machine" as any of the amenities the alternative "None" shouldn't be chosen
- If the household has "Dress dryer" as any of the amenities the alternative "None" shouldn't be chosen
- If the household has "Dish washer" as any of the amenities the alternative "None" shouldn't be choose
- If the household has "Boiler" as any of the amenities the alternative "None" shouldn't be chosen
- If the household has "Microwave oven" as any of the amenities the alternative "None" shouldn't be chosen
- If the household has "TV" as any of the amenities the alternative "None" shouldn't be chosen
- If the household has "TV decoder" as any of the amenities the alternative "None" shouldn't be chosen

- If the household has "Fixed telephone" as any of the amenities the alternative "None" shouldn't be chosen
- If the household has "Any mobile telephone" as any of the amenities the alternative "None" shouldn't be chosen
- If the household has "Computer" as any of the amenities the alternative "None" shouldn't be chosen
- If the household has "Internet connection" as any of the amenities the alternative "None" shouldn't be chosen
- If the household has "Solar panel" as any of the amenities the alternative "None" shouldn't be chosen
- If the household has "Air conditioner" as any of the amenities the alternative "None" shouldn't be chosen
- The household must have at least one of the amenities, otherwise the alternative "None" must be specified
- If the household has any car or minivans owned, their number must be specified
- If the household has no car or minivans owned, no number should be specified
- The household must specify at least one source of income

C. DWELLING DATASET

- If the type of dwelling is "Conventional dwelling", the occupancy status must be specified
- If the type of dwelling is "Non-conventional dwelling", the occupancy status shouldn't be specified
- For conventional dwellings inhabited by usual residents or persons not object of the Census, the dwelling questionnaire must be fulfilled
- If there is a main type of heating specified for the dwelling, the main type of energy used for heating must be specified
- If there is no heating in the dwelling, main type of energy used shouldn't be specified
- If the main type of heating is "Common heating in the building", the main type of energy used for heating cannot be "Electricity from the grid"
- If the main type of heating is "Separate central heating in the dwelling", the main type of energy used for heating cannot be "Wood", "Coal" of "Other"
- If the main type of heating is "Fireplace", the main type of energy used for heating must be "Wood"
- If the main type of heating is " Electric heater" or "Air conditioner" the main type of energy used for heating must be "Electricity"
- If the main type of heating is "Stove", the main type of energy used for heating cannot be "Coal", "Oil" or "Other type"
- If in the dwelling there are usual residents the occupancy status of the dwelling must be "conventional dwelling inhabited by one or more persons, usually resident"
- If in the dwelling there are no usual residents, the occupancy status shouldn't be "conventional dwelling inhabited by one or more persons, usually resident"
- If in the dwelling, live usual residents, the dwelling questionnaire must be fulfilled

ANNEX 4: IMPUTATION RATE AND DISSIMILARITY INDEX FOR THE INDIVIDUAL DATASETS

	Total Imputation				Deterministic Imputation				Probabilistic Imputation				Dissimilarity Indexes	
Citizenship: Albanian	**10.98**	**10.96**	**0.01**	**0.00**	10.36	10.36	0.00	0.00	0.61	0.60	0.01	0.00	IM	0.110
Had live-born children	**9.81**	**5.47**	**4.19**	**0.15**	3.76	0.00	3.76	0.00	6.06	5.47	0.44	0.15	IM	0.013
Family nucleus	**9.49**	**3.80**	**0.00**	**5.69**	0.00	0.00	0.00	0.00	9.49	3.80	0.00	5.69	IM	0.058
Completed years of education	**9.06**	**1.05**	**2.24**	**5.76**	0.02	0.00	0.00	0.02	9.04	1.05	2.24	5.75	KS	0.012
Highest completed level of education	**7.55**	**1.13**	**2.05**	**4.37**	0.00	0.00	0.00	0.00	7.55	1.13	2.05	4.37	IM	0.022
Searched for work during the month of September	**5.76**	**1.70**	**3.56**	**0.50**	0.00	0.00	0.00	0.00	5.76	1.70	3.56	0.50	IM	0.019
Place of residence in 2001	**5.72**	**4.02**	**1.38**	**0.32**	3.24	3.24	0.00	0.00	2.48	0.78	1.38	0.32	IM	0.026
Had a job last week of September	**5.31**	**1.29**	**4.02**	**0.00**	0.20	0.19	0.00	0.00	5.11	1.10	4.02	0.00	IM	0.027
Willingness to start job in two weeks	**4.92**	**1.33**	**3.59**	**0.01**	0.00	0.00	0.00	0.00	4.92	1.33	3.59	0.01	IM	0.023
Place of work: Type	**4.57**	**3.78**	**0.29**	**0.50**	0.07	0.04	0.00	0.03	4.50	3.73	0.29	0.47	IM	0.035
Place of residence was changed	**4.13**	**2.15**	**0.14**	**1.84**	3.03	1.57	0.00	1.45	1.10	0.57	0.14	0.38	IM	0.020
Have lived abroad	**4.10**	**1.94**	**1.97**	**0.19**	1.70	1.50	0.20	0.00	2.40	0.44	1.77	0.19	IM	0.020
Worked on last week of September	**4.08**	**1.15**	**2.15**	**0.78**	0.86	0.15	0.00	0.71	3.22	1.01	2.15	0.07	IM	0.019
Employment Status	**4.07**	**2.47**	**0.00**	**1.60**	0.89	0.74	0.00	0.15	3.18	1.73	0.00	1.46	IM	0.015
Currently attending formal education	**3.79**	**2.99**	**0.45**	**0.34**	0.21	0.00	0.00	0.21	3.58	2.99	0.45	0.13	IM	0.028
Place of work	**3.69**	**3.23**	**0.33**	**0.12**	1.58	1.58	0.00	0.00	2.11	1.66	0.33	0.12	IM	0.029
Main reason for not searching for work	**3.41**	**0.21**	**3.18**	**0.02**	0.00	0.00	0.00	0.00	3.41	0.21	3.18	0.02	IM	0.030
Place of stay on CRM	**3.08**	**1.25**	**0.00**	**1.83**	0.09	0.09	0.00	0.00	2.99	1.17	0.00	1.82	IM	0.019
Previous place of usual residence	**2.62**	**2.11**	**0.21**	**0.30**	2.19	2.01	0.00	0.19	0.42	0.10	0.21	0.11	IM	0.019
Civil status	**2.50**	**2.12**	**0.00**	**0.38**	0.79	0.79	0.00	0.00	1.71	1.33	0.00	0.38	IM	0.005
Place of birth	**2.49**	**2.34**	**0.00**	**0.16**	1.44	1.44	0.00	0.00	1.05	0.89	0.00	0.16	IM	0.006
Know to read/write	**2.03**	**1.11**	**0.64**	**0.27**	0.00	0.00	0.00	0.00	2.03	1.11	0.64	0.27	IM	0.012
Number of hours worked in the last week of September	**1.89**	**1.51**	**0.17**	**0.20**	0.00	0.00	0.00	0.00	1.89	1.51	0.17	0.20	KS	0.015
Job position in main job	**1.70**	**1.46**	**0.19**	**0.05**	0.00	0.00	0.00	0.00	1.70	1.46	0.19	0.05	IM	0.013

(continues)

	Total Imputation				Deterministic Imputation				Probabilistic Imputation				Dissimilarity Indexes	
Age	**1.69**	**0.63**	**0.00**	**1.06**	0.04	0.00	0.00	0.04	1.65	0.63	0.00	1.02	KS	0.006
Mother tongue	**1.61**	**1.61**	**0.00**	**0.00**	0.00	0.00	0.00	0.00	1.61	1.61	0.00	0.00	IM	0.000
Place of birth: Country	**1.60**	**0.43**	**1.17**	**0.00**	0.10	0.00	0.10	0.00	1.50	0.43	1.07	0.00	IM	0.008
Gender	**1.60**	**1.37**	**0.00**	**0.22**	0.00	0.00	0.00	0.00	1.60	1.37	0.00	0.22	IM	0.001
Number of children's still alive	**1.58**	**0.09**	**1.33**	**0.16**	1.38	0.00	1.22	0.16	0.20	0.09	0.11	0.01	KS	0.012
Citizenship: Other Specify	**1.58**	**0.68**	**0.86**	**0.04**	0.80	0.00	0.80	0.00	0.77	0.68	0.05	0.04	IM	0.006
Number of live-born children's	**1.51**	**0.04**	**1.31**	**0.16**	1.33	0.04	1.19	0.10	0.18	0.00	0.12	0.06	KS	0.013
Place of residence in 2001: District	**1.39**	**0.21**	**1.18**	**0.00**	0.00	0.00	0.00	0.00	1.39	0.21	1.18	0.00	IM	0.010
Citizenship: Other	**1.34**	**0.04**	**1.29**	**0.00**	0.02	0.00	0.02	0.00	1.31	0.04	1.27	0.00	IM	0.012
Difficulties: communication	**1.25**	**1.25**	**0.00**	**0.00**	0.00	0.00	0.00	0.00	1.25	1.25	0.00	0.00	KS	0.012
Difficulties: remembering	**1.19**	**1.19**	**0.00**	**0.00**	0.00	0.00	0.00	0.00	1.19	1.19	0.00	0.00	KS	0.012
Difficulties: daily self-care	**1.16**	**1.16**	**0.00**	**0.00**	0.00	0.00	0.00	0.00	1.16	1.16	0.00	0.00	KS	0.012
Difficulties: walking	**1.12**	**1.12**	**0.00**	**0.00**	0.00	0.00	0.00	0.00	1.12	1.12	0.00	0.00	KS	0.011
Place of residence in 2001: Town/Village	**1.03**	**0.00**	**1.03**	**0.00**	0.00	0.00	0.00	0.00	1.03	0.00	1.03	0.00	IM	0.016
Reason for changing the usual residence	**1.03**	**0.45**	**0.51**	**0.07**	0.19	0.00	0.19	0.00	0.83	0.45	0.31	0.07	IM	0.002
Year the current residence started	**1.02**	**0.25**	**0.51**	**0.26**	0.51	0.00	0.51	0.00	0.51	0.25	0.00	0.26	KS	0.003
Means used for travel to work	**1.02**	**0.74**	**0.27**	**0.01**	0.00	0.00	0.00	0.00	1.02	0.74	0.27	0.01	IM	0.005
Frequency of travel to work	**1.01**	**0.83**	**0.18**	**0.00**	0.00	0.00	0.00	0.00	1.01	0.83	0.18	0.00	IM	0.007
Difficulties: hearing	**0.98**	**0.98**	**0.00**	**0.00**	0.00	0.00	0.00	0.00	0.98	0.98	0.00	0.00	KS	0.010
Relationship to the Reference person	**0.95**	**0.33**	**0.00**	**0.62**	0.00	0.00	0.00	0.00	0.95	0.33	0.00	0.62	IM	0.005
Difficulties: seeing	**0.90**	**0.90**	**0.00**	**0.00**	0.00	0.00	0.00	0.00	0.90	0.90	0.00	0.00	KS	0.009
Place of work: District	**0.87**	**0.53**	**0.33**	**0.00**	0.00	0.00	0.00	0.00	0.87	0.53	0.33	0.00	IM	0.003
Place of work: Country	**0.82**	**0.76**	**0.06**	**0.00**	0.01	0.00	0.01	0.00	0.81	0.76	0.05	0.00	IM	0.007
Place of birth: District	**0.75**	**0.74**	**0.00**	**0.00**	0.00	0.00	0.00	0.00	0.75	0.74	0.00	0.00	IM	0.007
Main reason for coming to Albania	**0.66**	**0.17**	**0.08**	**0.40**	0.00	0.00	0.00	0.00	0.65	0.17	0.08	0.40	IM	0.004
Previous place of usual residence: District	**0.55**	**0.18**	**0.37**	**0.00**	0.19	0.00	0.19	0.00	0.36	0.18	0.18	0.00	IM	0.002
Place of stay on CRM: District	**0.52**	**0.03**	**0.49**	**0.00**	0.00	0.00	0.00	0.00	0.52	0.03	0.49	0.00	IM	0.005

(continues)

	Total Imputation				Deterministic Imputation				Probabilistic Imputation				Dissimilarity Indexes	
Reason for absence from place of usual residence	**0.48**	**0.00**	**0.48**	**0.00**	0.00	0.00	0.00	0.00	0.48	0.00	0.48	0.00	IM	0.005
Previous place of usual residence: Town/Village	**0.26**	**0.00**	**0.26**	**0.00**	0.00	0.00	0.00	0.00	0.26	0.00	0.26	0.00	IM	0.001
Country of last usual residence before coming to Albania	**0.23**	**0.15**	**0.08**	**0.00**	0.00	0.00	0.00	0.00	0.23	0.15	0.08	0.00	IM	0.001
Place of work: Town Village	**0.20**	**0.00**	**0.20**	**0.00**	0.00	0.00	0.00	0.00	0.20	0.00	0.20	0.00	IM	0.001
Year of last arrival to Albania	**0.20**	**0.11**	**0.03**	**0.06**	0.04	0.00	0.03	0.02	0.15	0.11	0.00	0.05	KS	0.001
Citizenship: No citizenship	**0.12**	**0.00**	**0.12**	**0.00**	0.00	0.00	0.00	0.00	0.12	0.00	0.12	0.00	IM	0.001
Place of residence in 2001: Country	**0.09**	**0.00**	**0.09**	**0.00**	0.07	0.00	0.07	0.00	0.03	0.00	0.03	0.00	IM	0.000
Place of stay on CRM: Country	**0.04**	**0.01**	**0.03**	**0.00**	0.01	0.00	0.01	0.00	0.04	0.01	0.03	0.00	IM	0.000
Place of birth: Town/Village	**0.00**	**0.00**	**0.00**	**0.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	IM	0.000
Place of stay on CRM: Town/Village	**0.00**	**0.00**	**0.00**	**0.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	IM	0.000
Ethno-cultural group	**0.00**	**0.00**	**0.00**	**0.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	IM	0.000
Religion	**0.00**	**0.00**	**0.00**	**0.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	IM	0.000

ANNEX 5: MATCHING PHILOSOPHY BETWEEN CENSUS AND PES DATA

For the analysis of the matched data it is necessary to insert Match codes into the abbreviated Census record and the post enumeration survey record. Additionally a Census record number needs to be added to the post enumeration survey record for matching at the later tabulation stage. Where the same dwelling is matched, the same code is added to both records together with the Census record identifier. Where the dwelling is not matched, a Match code is put against the visible record: for example, if a Census record cannot be matched, code 8 or 9 is put just against the Census record; where no matching record is found for a post enumeration survey record, Match code 7 is added to the post enumeration survey record, regardless of the outcome in post enumeration survey. The coding frame is set out below.

Coding Frame Census – post enumeration survey

Census completed [1]	PES completed [2]	1
	PES refusal/no contact [3]	2
	PES – dwelling space now empty [4]	3
Census refusal or non-contact [5]	PES completed [2]	4
	PES refusal/no contact [3]	5
	PES – dwelling space now empty [4]	6
Census completed, refusal or non-contact	PES record not found (and thus not matched)	7
Census –dwelling not identified (and thus not matched)	PES completed [2]	8
	PES refusal/no contact [3]	9
Census completed [6]	PES completed [7]	10

Initial computer match:

One major difference in the forms between Census and PES is that, in Census, a unit that is not a dwelling unit does not have a form completed for it, just a line in the summary booklet. In PES, every entrance and every door from the outside or a landing has a form completed for it. This was to ensure that questions were asked even in shops to ascertain whether anybody lived in the unit. Before attempting the computer match, it would be appropriate to remove these units from the file. They are identified as: Building questionnaire Q2=2.

Similarly, those buildings/dwellings identified as not habitable, ruined or under construction are not relevant to the match, they will not have any people and thus do not (yet) contain any residential units.

1 Census 'Questionnaire is completed', code 1 or 2 and HH code>0 (and not blank) and people in dwelling unit match with those in the completed matched PES dwelling unit
2 PES questionnaire, 'Interview status' code 1 or 2 and Building questionnaire Q4=1 and people in dwelling unit match with those in the completed matched Census dwelling unit
3 PES questionnaire, 'Interview status' code 3 or 4 and Building questionnaire Q4=1 4 PES questionnaire, 'Interview status' code 1 or 2 and Building questionnaire Q3=4 5 Census 'Questionnaire is completed', code 3 or 4
6 Census 'Questionnaire is completed', code 1 or 2 and HH code>0 (and not blank) and people in dwelling unit do not match with those in the completed matched PES dwelling unit (that is people in dwelling unit at Census moment moved out and another set of people moved in)
7 PES questionnaire, 'Interview status' code 1 or 2 and Building questionnaire Q4=1 and people in dwelling unit do not match with those in the completed matched Census dwelling unit (that is people in dwelling unit at Census moment moved out and another set of people moved in).

Within each EA:

1. For Census records with Household number >0 (that is an occupied dwelling) and 'Questionnaire completed' = 1 or 2 (Fully or partially completed):

   a. Seek match on Building number/floor number/name & surname of oldest person in household at Census moment/ age and sex of oldest person in household at Census moment/number of persons resident at Census moment, Match code 1 above. [Does the matched proportion improve significantly if the age is +/- 5? If father & son with the same name/surname they will be much more than 5 years apart in age.]

Manual match:

The majority of unmatched dwellings should be from the following scenarios:

a. Either Census or PES not completed with person information;

b. New dwellings not on the maps at the start of Census or PES which are then assigned building numbers in sequence, but not the same sequence in both Census and PES;

c. Different spellings of the name and/or surname of the oldest person;

d. Very different age/sex given for the oldest person.

As the completion rate for both the Census and PES is over 95 per cent, the proportion of dwellings to be manually matched is expected to be less than 50 per cent.

Any full matches identified in the manual match where both Census and PES are completed with some of the same people present at Census moment, that is having people in a residential unit, should be coded with Match code=1. Where both Census and PES are completed with none of the same people present at Census moment, that is both have people in a residential unit, records should be coded with Match code=10.

From the remaining records in Census file, exclude the non-inhabited dwelling units.

From the remaining records PES, exclude those non-inhabited dwelling units at the Census moment (building questionnaire Q3=4 and Q4=4) and any remaining non-residential units (building questionnaire Q2=2 or 3).

Then, with Census and PES remaining records sorted by building, level and unit number, allocate the match codes sequentially across the level. As seen in the examples, it is not necessary to have the 'unit' or 'door' numbers identical for the matching. Indeed, the entrance numbers will not match either, though both entrance number and 'door' numbers were allocated sequentially starting from the left. Hence, if the records are sorted by building number/entrance number/ level / 'door' number, it is expected that the units will match, though it is not guaranteed. Should any records not be matched in this way, codes are available in the table for allocation.