United Nations ECE/CES/GE.41/2018/13

Economic and Social Council

Distr.: General
13 July 2018

Original: English

Economic Commission for Europe

Conference of European Statisticians

Group of Experts on Population and Housing Censuses

Twentieth Meeting
Geneva, 26–28 September 2018
Item 3 of the provisional agenda
Measurement of the quality of administrative sources for use in censuses

Development and implementation of census evaluation techniques integrating Administrative Registers

Note by the Colombian National Statistic Department (DANE)*

Summary

The processing of the 2018 National Housing and Population Census database takes into account administrative data for its thematic complementation, its support in the post-census tasks and monitoring in the quality and coverage of the operation both in the virtual and in the face-to-face collection. These processes are important to guarantee consistency and reliability in the variables of identification, location, and indirect recollection through Administrative Registers.

Within the quality measures adopted, we highlight the construction of statistical files that contain all the rules implemented by the competent authorities at the national level regarding personal identification. For example, the adoption of procedures for the detection of duplicate and inconsistent data based on the basic variables of names, surnames, type of identification, identification number, sex and date of birth. Additionally, information-matching measures are used among the various Administrative Registers. This is evidenced by the construction of a Population Registry at the national level through the integration of Administrative Registers of health, education, among others.

The foregoing given that, in the previous testing processes, it was identified that the application of these techniques allows the generation of information with greater precision that will facilitate the continuous characterization of the population in both the census and post-census periods. Keywords. Administrative Registers, matching measures, integration of Administrative Registers, consistency and reliability in the variables.

* Prepared by Andryu Mendoza Beltran (e-mail: aemendozab@dane.gov.co) and Mariana Francisca Ospina Bohorquez (e-mail: mfospinab@dane.gov.co).

GE.18-11666(E)
1811666

I. Introduction

1. The National Population and Housing Census collect basic information on population, households, and housing. Some of the information collected is also found in the Administrative Registers the country has, for example, the affiliation of each person to social security or the characteristics of the victims in the country. In order to optimize resources and ensure quality in such information, DANE has chosen to take the information from the registry and not include it in the census form.

2. Access to Administrative Registers is permitted in Decree 1743 of 2016, where DANE can request administrative registers for the production of official statistics or the strengthening of their quality and coherence. Through access, and knowing the administrative veracity in the administrative registers, DANE makes use of these data to expand the temporal scope of the census operation.

3. The integration of the administrative registers to the census is done through the direct pairing of the identification variables of the people, which are contained in the data tables. This pairing is done by type and identification number, as well as by names and date of birth to ensure the effective and consistent crossing of people. Once the pairing is done, the administrative variables of the registry are included in the census database for the generation of final results.

4. Given that the process of integrating Administrative Registers with the base of the census is based on the identification of people, it is essential to monitor the quality of the information in the identification variables (Names, Surnames, Date of birth, Sex, Type and Document number). In addition, the quality of the variables in the Administrative Registers attached to the census that will include the census database and are specific to each of the administrative registers is monitored.

5. This document shows the quality analysis performed on the identification variables and the variables used in the Administrative Registers for the use of them in the Census. For this, the document is divided into three chapters: the first one is the methodologies used for the quality analysis, which is divided into two parts, the quality of the information in the registry and the coherence of the information when compared with other registers. The second chapter shows results of the quality analysis and the experience obtained in this aspect. The third chapter will be the conclusions.

II. Methodology

6. The quality analysis of the registers is divided into two parts: the first is based on the quality of the data according to the rules and parameters of each variable that constitutes the registers. In this analysis, the amount of data that is empty, inconsistent and information that doubled to people in the Registry is taken into account as a descriptive statistical analysis, which is synthesized in a "Statistical registers".

7. The second part of the analysis is based on the integration of the Administrative Registers to see the response agreement of a person in registers that have the same and close temporality. In this case, we have the construction of a Statistical Base Population Registry (chapter 2.2.1, Wallgren c2012). This is because most of the Administrative Registers contemplate the identification information of the people.

A. Quality record

8. The quality record is a table that contains the absolute frequencies of empty, inconsistent fields and duplicates of the identification variables of the Administrative Record, as shown in Table 1.

Table 1
Statistical quality table

		Data presence		*Validation registers with data*		*Data duplicated by variable*	
No.	*Variable name*	*Empty records*	*With data*	*Records with inconsistent*	*Records consisting*	*Duplicated records by identification number*	*Duplicated records by names and birthdate*
1	ID_type_identification						
2	Number_identification						
3	First_surname						
4	Second_surname						
5	First_name						
6	Second_surname						
7	Birthdate						
8	ID_sex						

9. As the name indicates, the number of cells with "Empty records" corresponds to the totality of people who do not have information in this variable. Inconsistent registers meet the standards mentioned in Table 2. Duplicate records can have two considerations: be duplicate persons by type and identification number or be duplicated by names and date of birth.

Table 2
Tables with normal quality of the identification variables

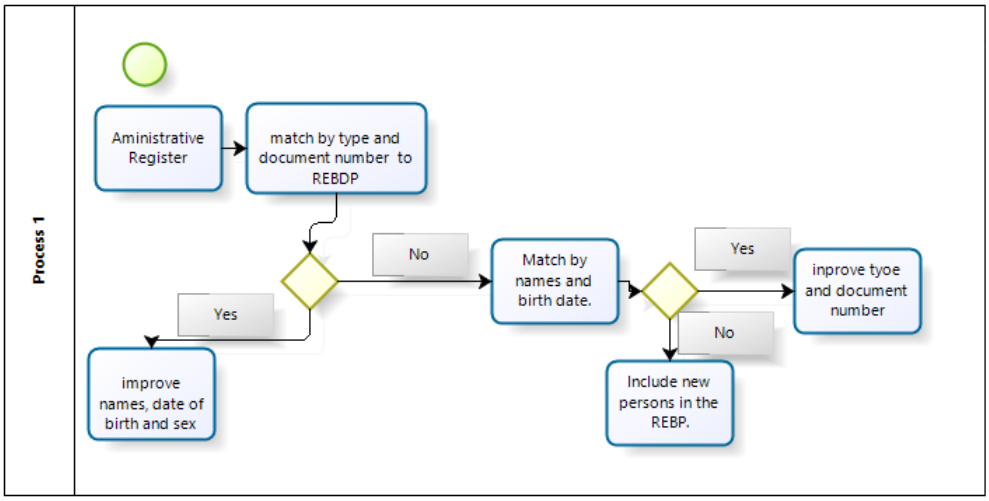
Variables	*Categories*	*Validation rules*
Type of identification	RC (Civil Register)	Between zero (0) to less than 7 years old
	TI (Identity Card)	Between 7 and 18 years old
	CC (Citizenship card)	More than 18 years old
	CE (Foreigner card)	Foreigners over 18 years old
	PA(Passport)	Foreigners under 18 years old
Identification number	RC (Civil Register)	8 or 11 numeric digits, 10 alphanumeric digits
	TI (Identity Card)	10 or 11 numeric digits
	CC (Citizenship card)	From 3 to 8, or 10 numeric digits
	CE (Foreigner card)	9 to 11 numeric
	PA(Passport)	Under 17 alphanumeric digits
Names and Surnames	Not numeric, greater than two characters, not containing strings such as NN or special characters	
Birth date	Date less than that of the registry, when calculating the age less than 110 years	
Age	Under 110 years old	

B. Quality Parameters with the construction of a Statistical Base Population Registers

10. One way to analyse the consistency and quality of the information of the Administrative Registers is through the integration between them. This integration can be done through the construction of a Statistical Base Population Registry (REBP), in such a way that we can see the agreement between the variables, and the improvement of the identification variables during the integration.

11. The construction of the REBP begins with the planning of an initial registry and the subsequent integration of each registry into it. Pairing is done by the type and number of the document, individuals who do not cross are paired by the phonetic fields of the names with date of birth. During the process, the fields of the Registry under construction are improved, as well as including the new individuals that were not. Graph I shows the integration process.

Graph I
The integration process in a REBP



12. From the integration process, we can obtain statistics of quality and concordance of the identification variables of the people in each one of them. In a complementary way, the construction of the REBP improves the data as the integration that is being carried out.

III. Results

13. Such as we showed in the methodology, the results are present in two parts: the first of them relate the results of the statistical card, and the second part shows the quality analysis under the integration.

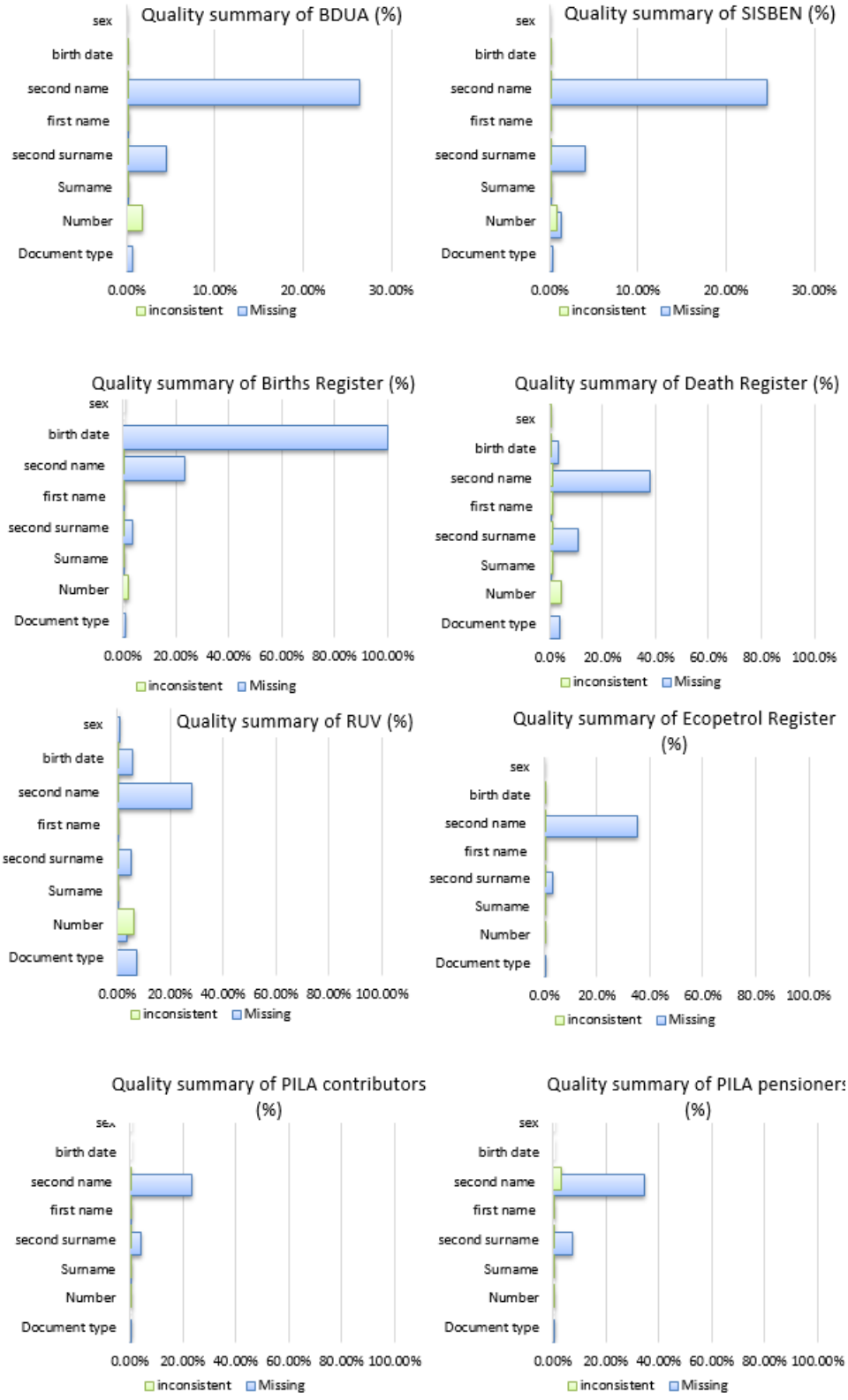
14. Taking into account the health record *Base Unica de Afiliados (BDUA)*, the register of beneficiaries *Sistema de Selección de Beneficiarios Para Programas Sociales (SISBEN)*, the register of victims *Registro Unico de Victimas (RUV)*, the register of *Ecopetrol*, the register health *Planilla de Afiliados cotizantes a pensiones (PILA)* and the health register *Planilla de Afiliados de pensiones*. Each of these registers has a quality card that allows to identify the total of empty, inconsistent and duplicate fields of the identification variables.

15. Graph II shows the percentage of inconsistent and empty registers in each of the Administrative Registers for each identification variables. We can observe a high number of empty fields in the reports of the second names and surnames. With relations to inconsistence information, we note a slight occurrence in the identification number, as observed in the RUV register. These results give a consistency in the use of the matching by alphabetical variables with the birth date.

16. Graph III shows the percentage of duplicated registers by the type and number of the document, and the duplicates by names and birth date. There is a high occurrence of the phenomenon in the SISBEN and RUV registers. We use this analysis by the integration of registers.

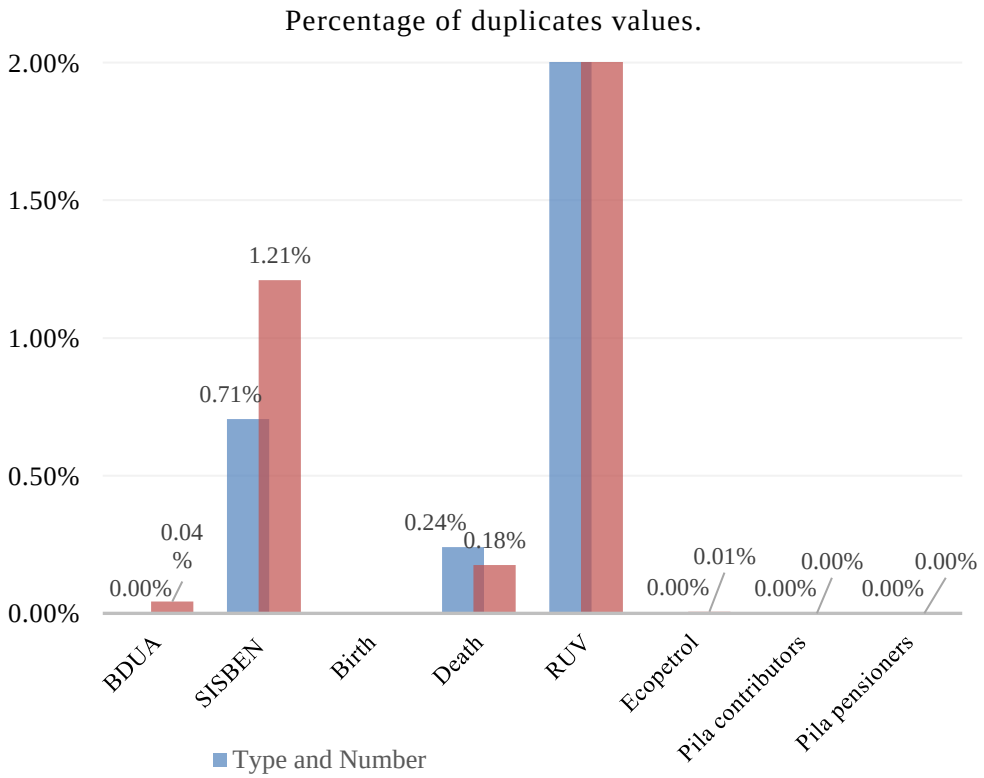
Graph II

Percentage of empty and inconsistent values in each of the administrative record by each identification variables



Graph III

Percentage of duplicate values by type and number of document, and for name and birth date in each Administrative record



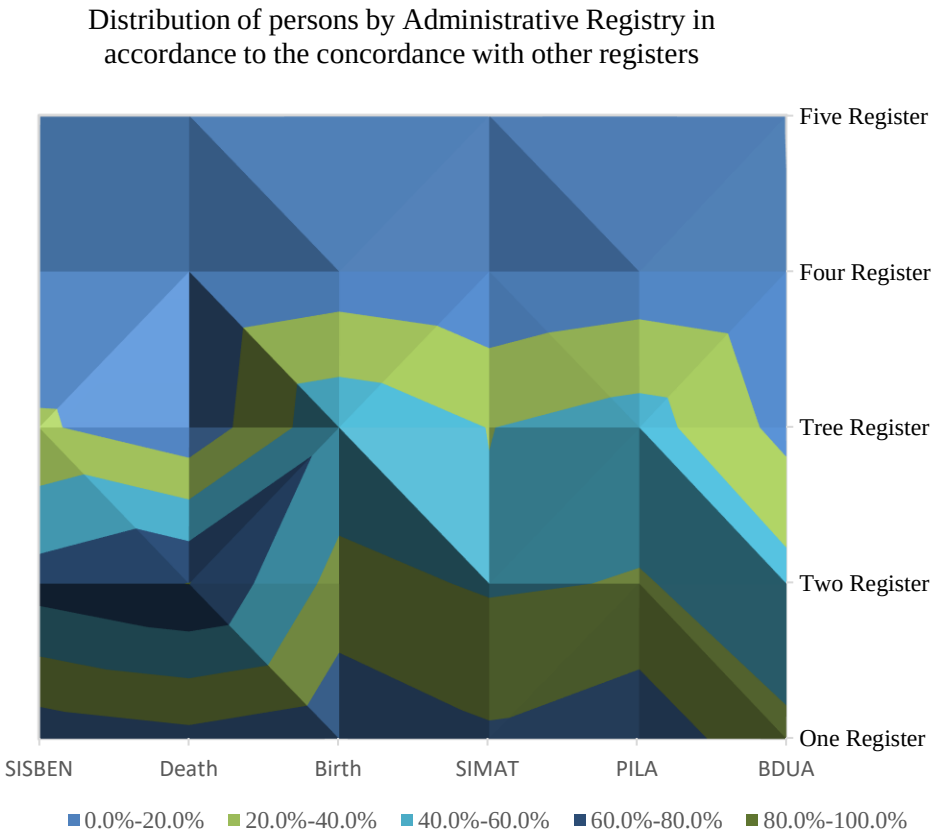
17. Once the quality analysis of the variables in each of the Administrative Registers is made, we continue with the construction of the REBP. At this point, DANE has built the Statistical Base Population Register in some territories in order to evaluate the methodology and the proper use of them for statistical purpose.

18. Table 3 shows the distribution of individuals in one or more Administrative Registers. This was made in the construction of the REBP in the Department of *San Andrés, Providencia y Santa Catalina – Colombia*. It is relevant the high number of registers in more than two Administrative Registers, allowing to diagnose and improve the quality of the information in a REBP for its use.

Table 3
Percentual distribution of people according the number of registers that appears in the REBP in San Andrés, Providencia y Santa Catalina

Concurrence of people into the registers				
One Register	*Two Registers*	*Three Registers*	*Four Registers*	*Five Registers*
41,24%	45,65%	12,47%	0,64%	0,00%

Graph IV
Percentage distribution of people that are in a record according to the agreement in other registers



19. Graph IV shows the concordance of people in one or more registers taking into considerations the register where it appears. Darker parts of the graph show a higher intensity of people in a certain number of records where it appears. We can see that the majority of people who are in SISBEN are in only one other registers while the people who appear in registries like SIMAT, Birth or PILA are also in three or more. In these registries, we can make a better analysis of quality of the data.

IV. Conclusions

20. The quality of the variables in the Administrative Registers is closely related to the metadata that is the basis of the administrative acts. Despite the great efforts made by entities to have high-quality data, the reality presents diverse scenarios that are not contemplated, therefore, it is recommended to apply basic validation meshes that identify missing information and consistent inconsistencies with the metadata before using the registries for statistical purposes.

21. The agreement of people in the various administrative registers provides information to analyse the quality of the Registry. However, we can find equal people in different

Administrative Registers with different information. It is advisable to have the registers that assign the administrative information to the people, for example, the National Registry, who assign the identification information to people, are the ones that give the exact data being the national authority in this matter.

Bibliography

Wallgren, A., Wallgren, B. (c2012), *Estadísticas basadas en registros, Aprovechamiento estadístico de datos administrativos. Instituto Nacional de estadística y geografía* (México) INEGI.