United Nations ECE/CES/GE.41/2018/11

Economic and Social Council

Distr.: General
19 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 9 of the provisional agenda
Relation between censuses and other statistics, such as demographic, labour and regional statistics

Updating estimates at the municipal level using population censuses and sample surveys

Note by the Colombian National Statistic Department (DANE)[1]

Summary

The National Administrative Department of Statistics of Colombia (DANE) generates estimates of total population at the municipal level, based on information collected in the population and housing censuses. However, in the inter-census period (10 or more years), the measures and trends of the main demographic indicators might change in smaller areas.

For the national and regional levels, indicators obtained from sample surveys have been used in recent decades. However, at the municipal or local level, there is a need to implement robust methodologies to obtain these estimates. In addition, it must be taken into account that social transformations do not usually happen homogeneously, but they are diverse across large geographic areas.

This document shows the results of the application of the Fay Herriot methodology for updating population estimates in the municipalities of Colombia. Statistical models allow the incorporation of additional information in the estimation process. The inclusion of contextual variables serves to enrich the analysis. In addition, the information provided by the 2018 National Housing and Population Census will make it possible to assess the coherence of the geographical differences identified in the models.

[1] The present document was submitted late due to resources constraints.

GE.18-11941(E)
1811941

I. Introduction

1. In Colombia, in accordance with the Political Constitution in force since 1991, the municipalities are the fundamental entities of the Political-Administrative Division of the country. Currently, the territory is organized into 1101 municipalities, 20 non-municipalized areas and the Island of San Andrés, among which contracted brands are presented not only in terms of their degree of socioeconomic development but also in terms of the quality and quantity of information available.

2. Local planning faces important challenges due to the lack of information at a high level of disaggregation, in addition to the fact that the phenomena are heterogeneous between and within geographical areas. In the intercensal periods, it is where these deficiencies become more evident because the characteristics of the phenomena vary over time and there are not many high-quality sources of information that account for said changes.

3. On the other hand, increasing the sample sizes of the surveys prepared by the National Administrative Department of Statistics is an economic solution in terms of economic resources. The municipalities are generally unplanned domains within the sample surveys developed by the institution. Therefore, DANE seeks to validate the application of mixed logit models to estimate the probability of weighted sums in smaller areas (Hobza and Morales, 2016) based on the results of the National Population and Housing Census 2018 (in execution) in several thematic aspects.

4. The estimation in small areas is one of the alternatives that have been developed to provide a solution to this need to have figures with a high level of disaggregation for decision makers and researchers who can use them for planning purposes and as input within other processes of modelling the socioeconomic reality of the country.

II. Sources of information and methodology

5. The applications under development by DANE's Census and Demography Working Group are to evaluate the possibilities and limits of the application of these methodologies to improve the trends of the indicators of interest in unplanned domains. Within this work it should be noted that in Colombia there are three main groups of sources of demographic information: population censuses, vital statistics registers and sample surveys.

6. These exercises have been possible within the framework of technical cooperation with other statistical institutes in the region, in order to consolidate a system of indicators that incorporates information from the three sources mentioned above, in such a way that censuses and registers can be used as auxiliary information to obtain more robust indicators at municipal level for intercensal periods.

7. In one hand, it is intended to give a solution by explicitly considering the temporal correlation in the phenomena of interest. This, because it has been identified that the territorial differences tend to change in their magnitudes, but basically classifications or rankings of the municipalities with greater and lesser progress in the social, economic, demographic and environmental indicators can be obtained, which do not change largely in the short term. That is to say that in a period of 10 to 15 years, the territorial contrasts are not modified in an extreme way, but rather changes are gradual.

8. On the other hand, it is proposed that in the future models that also incorporate the spatial correlation structure are used. In other words, they take into account that the nearest municipalities tend to resemble each other geographically, while the more distant tend to behave differently in their indicators.

III. Models of Fay Herriot

9. Following the approach of Esteban, Morales, and Pérez (2016) this model can be defined in two stages. In the first, the sample model relates the sampling error of a direct indicator. That is to say:

$$yd = ud + ed$$

10. Where d = 1, ..., D; ud is the characteristic of interest, and yd is the direct estimator of ud, and the errors are independent and have a normal distribution with zero means and known variances.

11. In the second stage, the link model assumes that the characteristics of the area are linearly dependent on p auxiliary area variables.

$$\mu d = xd\,\beta + ud,\ d = 1, ..., D,$$

12. Where xd is a row vector that contains aggregate values of p auxiliary variables per area. This model links the variables of interest for all areas by means of a regression parameter, and can be re-expressed as:

$$yd = xd\,\beta + ud + ed,\ d = 1, ..., D.$$

IV. Progress status of the project

13. Given that historical information is available mainly from population censuses, it is considered that this can be used to improve estimates of periods after it. Three applications that are currently being developed refer to the following topics:

(a) Estimation of the number of migrants based on the information provided by households with migratory experience in demographic and health surveys, and employment surveys;

(b) Estimation of the number of Internet users based on information from sample surveys, administrative registers and data from households with Internet access from population censuses;

(c) Estimation of the number of disabled people, based on information from population censuses, administrative registers and sample surveys.

V. Preliminary conclusions

14. The techniques based on these models had not been used in the generation of official statistics, therefore the preliminary results for the study phenomena are currently being evaluated, using relatively small populations in comparison to the data observed in reality, this approach requires survey data at unit-level to adjust the models, as well as census data at cross-domain level to build the best empirical predictor (PSB).

15. Therefore, it is considered that the development of prototypes of analysis and adaptation of algorithms for the estimation of the indicators and their coefficients of variation is a complex process that requires a longer maturation time before going from this stage of development to a stage of production and publication. This approach requires survey data at the unit level to adjust the models, and cross-domain level census data to construct the PBS.

16. In the tests carried out, the estimators based on the Fay Harriot models show significant improvements in the coefficients of variation estimated in the medium and small

size municipalities, however in the larger municipalities such as the capital of the country – Bogota – and the second largest city – Medellin – on some occasions the estimate shows a greater variability, so it is stated that it is necessary to continue testing with different auxiliary variables and methodological alternatives for the consideration of the temporal and spatial structure of the phenomena of interest.

17. Finally, it should be noted that an important advantage of the area models has been identified, which is that only the totals of the auxiliary variables are needed for the study domains, information that is often available in censuses or administrative registers.

Bibliographic references

Esteban, M. D., Morales, D., & Pérez, A. (2016) Area-level Spatio-temporal Small Area Estimation Models. Analysis of Poverty Data by Small Area Estimation, 205–226.

Hobza, T., & Morales, D. (2016). Empirical best prediction under unit-level logit mixed models. Journal of official statistics, 32(3), 661–692.

Rao, J.N.K. and Molina, I. (2015). Small area estimation, Second Edition. John Wiley.