Catalogue no. 11-522-x
ISSN: 1709-8211

Proceedings of Statistics Canada Symposium 2021
Adopting Data Science in Official Statistics to Meet Society's Emerging Needs

Administrative data for the estimation of population counts: statistical learning from the first waves of Italian Permanent Population Census

by Angela Chieppa, Nicolandta Cibella, Antonella Bernardini, Giampaolo de Matteis and Silvia Farano

Release date: October 22, 2021

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2021
Adopting Data Science in Official Statistics to Meet Society's Emerging Needs

Administrative data for the estimation of population counts: statistical learning from the first waves of Italian Permanent Population Census

Angela Chieppa, Nicoletta Cibella, Antonella Bernardini, Giampaolo De Matteis, Silvia Farano[1]

Abstract

The Permanent Census of Population and Housing is the new census strategy adopted in Italy in 2018: it is based on statistical registers combined with data collected through surveys specifically designed to improve registers quality and assure Census outputs. The register at the core of the Permanent Census is the Population Base Register (PBR), whose main administrative sources are the Local Population Registers.

The population counts are determined correcting the PBR data with coefficients based on the coverage errors estimated with surveys data, but the need for additional administrative sources clearly emerged while processing the data collected with the first round of Permanent Census. The suspension of surveys due to global-pandemic emergency, together with a serious reduction in census budget for next years, makes more urgent a change in estimation process so to use administrative data as the main source.

A thematic register has been set up to exploit all the additional administrative sources: knowledge discovery from this database is essential to extract relevant patterns and to build new dimensions called signs of life, useful for population estimation.

The availability of the collected data of the two first waves of Census offers a unique and valuable set for statistical learning: association between surveys results and 'signs of life' could be used to build classification model to predict coverage errors in PBR.

This paper present the results of the process to produce 'signs of life' that proved to be significant in population estimation.

Key Words: Administrative data; Population Census; Statistical Registers; Knowledge discovery from databases.

1. Introduction

1.1 Italian Permanent Population Census and Population Counts

The *Permanent Census of Population and Housing* is the census strategy adopted in 2018 in Italy, based on statistical registers continuously updated with administrative and surveys data. Two sampling surveys are carried out yearly to improve the quality of registers and complete them with the information they do not provide, so to assure Census outputs with same dissemination level of the past traditional editions.

Census Population counts, i.e. total amounts of usually resident population for each Municipality, are the primary Census output: in the new framework of an official statistical production based on registers, population counts estimation is based on the *Population Base Register* (*PBR*) corrected by means of survey results and, possibly, integration with other administrative data. PBR main administrative sources are the Local Municipal Population Registers.

The PBR may be affected by errors of over-coverage, i.e. inclusion in the register of individuals who are no longer on the territory, and by errors of under-coverage, i.e. non-inclusion in the register of individuals who are on the territory.

[1] Italian National Institute of Statistics (ISTAT), Population Census Dept., (chieppa@istat.it , cibella@istat.it).

For the start-up cycle of Permanent Census, the methodological design to compute the correction factors for PBR consists in two steps:

- Direct estimation for Municipalities in the yearly survey sample: the capture-recapture model is adopted to estimate the coverage errors of PBR; the 'first capture' is the presence in the register, while the 'second capture' (the necessary independent measurement of the same target population) are data collected with the Census surveys.

- Indirect estimation for non-sampled Municipalities, making use of Small Area Models.

1.2 Administrative '*signs of life*'

A thematic database (register), called *AIDA-Integrated Archive of Usual Resident Population*, has been set up to exploit all administrative sources useful for population estimation.

Specific workflows have been developed to manage data coming from administrative sources, such as: Labour and Education, Tax Returns, Earnings, Retirement and Non-Pension Benefits, Permits to Stay, apartments' Rental and Sales. Administrative signals of each individual, i.e. the presence in one or more of the sources, could be considered as *signs of life (SoL)* of a person in Italy, possible proxy of usual residence. These signs of life are classified according to duration patterns, type and reliability of the specific source, relation with other individual signals (e.g. household relations).

AIDA signs of life and surveys data are both valuable sources for the estimation of under and over-coverage of PBR data. For the first Permanent Census waves, SoL have been used to manage survey under-coverage: all sample units from PBR not respondent to the survey but associated with strong SoL in AIDA were recovered and considered 'usually resident'.

1.3 Pushing factors to improve Permanent Census design and current investigation areas

The first cycle of Permanent Census has achieved its main objectives: the two waves of Census annual surveys registered very high response rates and the Permanent Census permits disseminating the Census results respecting cost constraints and timeliness.

Nevertheless, some critical issues emerged in relation to the estimation design, that call for a general rethinking:

- Sampling issues:

    - even minimal variations in response rates and in survey outcomes produce very substantial impacts on the population counts, both in absolute and relative terms;

    - given the sample sizes, the correctors are significant at the municipal*citizenship level, but more detailed domains are requested;

    - need to manage the Municipal bias.

- Surveys costs, respondents burden, survey quality:

    - need to leverage administrative records to reduce costs;

    - non-ignorable survey nonresponse;

    - the complexity and the strong dependencies from the survey network.

- Need for a shared, accountable estimation process for population counts:

    - Census results on population has a strong administrative and political impact, determining funding levels, availability of services and so on: estimation process has to be shared with public stakeholders (understandable estimation process).

    - decimal correcting factors for groups or records are not easy to share when Municipal estimated counts are disputed by local authorities; lists of individuals records to revise Municipal data are often needed (usable results of estimation process).

Moreover, the COVID Pandemic was a strong pushing factor in the use of administrative data in absence of surveys, because Census surveys were suspended in 2020.

There are two main areas of investigation underway: a) new Census surveys' design for the post-2021 cycle, to ensure an optimal trade-off between costs and quality of estimates; b) designing and testing alternative methodological solutions for the estimation of the population counts by making the most efficient use of administrative data integrated in the AIDA database. This last area includes the experimentation described in the following paragraphs.

2. A data-science approach to design a new solution

Availability of both the first two waves of Permanent Census surveys and administrative sources, covering the same reference period, makes possible to identify patterns in data relevant for the estimation of the population counts and to develop a data driven solution for the prediction of usual residence.

Given the complexity of the phenomenon and the fact that administrative data are not specifically designed for statistical purposes, it is extremely important to extract knowledge from these databases so to improve their quality and choose/build proper dimensions for the estimation of Census results. An interdisciplinary team is needed to develop this knowledge discovery process: methodological issues are as critical as database management or process automation skills and as expert knowledge on thematic issues (such as usual residence, administrative sources quality and surveys data quality).

Interdisciplinary team, statistical learning from data to derive a predictive model for usual residence, database management to improve workflows of administrative data are key components of a so-called *data-science approach*. An experimentation of this kind of approach, using all available data, is currently underway at ISTAT to improve the production of Census results for next years by exploiting administrative data.

In terms of a statistical learning problem, the final goal of this experimentation is to find the best 'learner' to predict usual residence in Italy at individual level (target/response variable) by classifying administrative signs of life and presence in PBR (covariates/auxiliary/independent variables).

Also, the intermediate results of this learning process are crucial to improve the Permanent Census design: 1) to improve the workflows that make use of administrative data (standardized workflow to manage different sources and managing variables of each source, whose number could vary from one year to another; 2) features selection for usual residence estimation (identification of a subset of relevant variables); 3) tuning of AIDA SoL database (standardization of variables computed from administrative sources that could vary from one year to another); 4) detection of critical patterns, associated with high uncertain prediction.

The experimentation steps are the following:

- *Set up of the experimental databases*: pre-processing and loading data from surveys, base register, administrative sources; executing linkage at individual level among different sources.

- *Data Exploration*: study of associations between administrative signals and PBR information, making use of survey data as outcome measurement of usual residence and so identifying patterns of under- and over-coverage in the register.

- *Feature engineering*:

  - formalization of expert knowledge

  - introduction/building new variables, based on data exploration results and on expert rules

- *Predictive modelling*:

  - training/test of classification models (latent models, decision tree and others);

  - deterministic rules defined by experts.

- *Statistical and thematic evaluation of results*:

  - correction of PBR based on individual classification for over and under coverage;

  - analysis of the results; comparison among different models results and with 2018-2019 estimations.

Between different steps, there is not only a sequential direction, but also continuous feeding between each other.

Shared and interpreted results of every single step, including data preparation, are the intermediate goals that are so important in developing a feasible new solution (redesign of data collection and sampling scheme for surveys; new workflow to manage administrative sources and so on).

3. First steps of experimentation

3.1 Exploratory Data Analysis on integrated experimental database

Data from different sources have been integrated in a database useful for learning objectives, with linkage at individual level and assuring coherence of time/period reference among sources.

- Survey data have been loaded taking into account the data-collection experts' validation rules; in addition to microdata collected, quality indicators deriving from field-operations are loaded.

- For Administrative sources (SoL) loading workflows on more than 50 different archives were standardized to produce comparable 'signals', differentiated for duration and quality of source. Time reference is managed with proper time-window on duration of signal, centered on 31st December of each year; different period are tested, from a longitudinal window of more years to at least 12 months. Continuity patterns on duration of administrative signs reflect in so-called 'strength of signal'.

- From register (PBR), core demographic variables and place of usual residence are loaded; time reference is 31st December of each years and Census survey day, to assure same reference/period of surveys data.

This experimental integrated database is continuously modified with knowledge process achievements, so there are different releases of same database, each release with some new or transformed dimension.

At the starting point of the learning process, the joint distribution of 'presence in PBR' and 'presence in administrative sources' shows different subpopulations that need further investigations. Out of more than 60 million individual records in the set of people with presence in PBR in years 2018-2019 and/or with administrative signals in the same period, there are:

- 2,5% of people in PBR without any administrative signs: they have to be investigated to verify which part are really over-coverage of register;

- about 1,8 million of people that are not registered in PBR, but with administrative signs in same period: 28,7% of them are strong signs, that is to say steady/continuous presence derived from qualified sources: are they all really under-coverage of register?

- about 97% of people in PBR with signs in administrative sources, but

    - 5% with 'episodic' signs (duration of sign or source is not significant according usual residence definition) could be related with over-coverage.

    - 4 million of these individuals have signs out of PBR province: which part are error of place of usual residence in PBR (local component of coverage error)?

In these investigations, response to survey could help to differentiate subpopulation and detect individual records eligible of coverage errors in PBR.

The analysis dimensions are individual ones such as age, sex, administrative signs /profile (when signs exist), information on place of residence (when the person in recorded in register) and localization of administrative signal, but also municipal and regional attributes; finally, for the subset of those who entered the Census Surveys samples, responses to surveys and other quality indicators from the fieldwork are included. Multiple Correspondence Analysis and Classification Trees result to be very good tools to detect patterns in data and also for sharing results with thematic experts.

Some patterns clearly emerged, together with relevant 'features'/variables:

- *Municipal Population size is clearly a discriminant dimension*: small towns are associated with very high response rate for surveys and good quality in Register, while in big cities there are critical patterns for which is more difficult to evaluate Register quality.

- *Elderly/retired people* as well as *familiar relationship derived from fiscal signals* are profiles of people that result to be predictive of good quality in the register

- Further data are needed for better understand and reduce uncertainty on *critical patterns, such as people with no signs, one-person household, foreigners in big cities*, because the quality of prediction, using available variables on administrative signs and current surveys, results to be too low.

- In *medium size towns,* there is a significant group of individuals whose administrative signs are localized outside the PBR province: in this case, there is also a need for further analysis, to correctly dimension effective misclassification in PBR place of residence.

- For *Foreigners*, a relevant dimension to predict usual residence is the Area/Country of citizenship.

- *Longitudinal SoL*, on more than one year, seize better the *foreign component of PBR under-coverage*.

Analysis results and comparison between survey and administrative data were also useful in the improvement of administrative loading workflow and microlevel linkage (to solve some false signal).

3.2 Using expert knowledge to compute new variables

Analysis results show that there is a need for more information that could be partly derived from existing variables or that could also involve some changes in the loading processes from original sources. The experts' knowledge, about social characteristics and administrative sources, is key in this phase that is known as *Feature engineering step*. Feature engineering has a high impact on the performance of final predictive models.

In current experimentation, the formalization of 'expert rules' to detect some specific subpopulation, making use of existing variables, has allowed to convert expert knowledge into new variables/data. For example, expert rules have been used to detect, among all individuals with work/study signs localized outside the province of usual residence recorded in the PBR, which of them are commuters, so to split apart the cases of misclassification/coverage error. Once formalized, new relevant variables were added to the experimental database to implement these expert rules such as 'family ties', 'travel routes', 'typology of municipalities based on commuting habits'.

In addition, expert were asked to define rules to detect in which cases (what type of SoL and demographic characteristic in PBR) an individual record could be considered an error of coverage of the base register. This deterministic classification would result useful when training and test statistical model: on one hand, it could be used as a benchmark; on the other hand, this formalization of expert knowledge could be of a great help in the interpretation of the results of statistical predictive models.

Even using deterministic rules based on expert knowledge, there are individual data that are difficult to classify, so critical areas remain. A data-driven statistical model will probably reduce this aggregate, but the quality of prediction of usual residence for some critical patterns would probably imply new specific targets for the surveys design.

4. Next steps and some final remarks

After the setting up of the experimentation and the first steps of exploratory analysis and feature selection/engineering, the learning process is currently in the final decisive steps:

- management of new sources (related to Covid - peculiarities of the year 2020) in the AIDA register;
- application of refined rules and new variables on the database;
- selection of statistical predictive model
- evaluation of selected model and comparison with results of experts deterministic rules.

The Permanent Population Census strategy has determined a radical change in statistical production of Census results.

Nowadays, a further step is needed towards a massive use of administrative data for improving the quality of PBR data and Census population counts.

Best solution implies not only methodological aspects and model assumptions, but also surveys costs and feasibility issues as well as expertise in designing and managing databases workflows: an interdisciplinary approach is needed.

The data of the first two waves of Census surveys constitute a valuable source for learning from data how to integrate administrative signs of life in the estimation process.

First steps of this kind of investigations result in:

- an experimental integrated database useful for training/testing/evaluating alternative models for estimation
- selection of relevant dimensions, including new variables derived from existing data ad expert rules
- detection of critical patterns: the new design of surveys could be used to collect data to solve this uncertainty.

New estimation models are currently being evaluated, based on statistical predictive model derived from learning process; the usefulness of mixed solutions based on classifiers that also take into account criteria entered by thematic experts is emphasized.

References

Bernardini A., Chieppa A., Cibella N., Solari F., (2021), "Administrative data for population counts estimations in Italian Population Census", in Perna C. et al (eds) *Book of short papers - SIS 2021*, Pearson, pp. 274-278

Bernardini A., Cibella N., Gallo G., & Al. (2019), "Empirical evidence for population counting: the combined use of administrative sources and survey data", paper presented at ESS Workshop on the use of administrative data and social statistics, Valencia, Spain.

Casari A, Zheng A. (2018), *Feature Engineering for Machine Learning*, Boston, O'Reilly Inc.

Chieppa A., Gallo G., Tomeo V. & Al. (2018), "Knowledge discovery for inferring the usually resident population from administrative registers", *Mathematical Population Studies. International Journal of Mathematical Demography*, 26:2, pp. 92-106.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009), *The elements of statistical learning: data mining, inference, and prediction*, New York, Springer

Kim J.K., Rao J.N.K. (2012), Combining data from two independent surveys: a model assisted approach, *Biometrika*, 99(1), pp. 85-100.

Pfeffermann, Danny & Eltinge, John & Brown, Lawrence. (2015), "Methodological Issues and Challenges in the Production of Official Statistics: 24th Annual Morris Hansen Lecture", *Journal of Survey Statistics and Methodology*, 3, pp. 425–483

UNECE (2018), *Guidelines on the Use of Registers and Administrative Data for Population and Housing Censuses.*

UNECE (2020), *New frontiers for censuses beyond 2020.*

Zhang, Li-Chun (2019), "A note on dual system population size estimator", *Journal of Official Statistics*,35(1), pp. 279-283.