Data mining techniques on the administrative data system to enhance the accuracy of the population census counts

Antonella Bernardini[1], Nicoletta Cibella[1], Giampaolo De Matteis[1], Gerardo Gallo[1], Antonio Laureti Palma[1], Fabrizio Solari[1]

[1]Italian National Statistical Institute (Istat, Italy), lauretip@istat.it

Abstract

Since 2020, the Italian National Institute of Statistics (ISTAT) has been producing fully register-based population count integrating the population register with 'Signs of Life' (SoL) derived from administrative data. On the basis of the recent development of data mining applications, as well as the increasing availability of large amounts of data, the aim of this study is to experiment new methodologies for estimating and assessing population counts. First of all, SoL are used to implement first a supervised classification strategy to distinguish between usually resident and not usually resident population in Italy and after an unsupervised model to assess the most plausible usual place of residence.

In the first step, no specific location was assumed, although location variables, as place of work or tax domicile, were used as SoL identifiers. Supervised training and testing data sets were built using 2021 census sample survey data. A machine learning model is used for classification based on the Support Vector Machine (SVM). Overfitting and underfitting estimators were checked using the cross-validation and the validation curve, while the size of the training set was controlled by the learning curve. Quality assessment of Machine Learning (ML) results was performed, as well as an evaluation of the importance of the 2021 census sample survey data as the training set. The assessment shows the discriminant role played by a population register in a register-based population count estimation, highlighting the Italian situation from countries where population registers are not available.

In the second step, in order to assess the real usual place of residence, we use utility consumption, electricity and gas, as data sources to identify the monthly consumption patterns associated with each point of delivery. Through a cluster analysis of the consumption patterns in association with the information on households included in statistical registers, it is possible to assess the usual place of residence, i.e. where a household, or part of it, actually lives, reducing the possible misplacement errors. Furthermore, under certain conditions of unicity between services provided and households served, through each home's energy consumption model, it is possible to estimate the number of people who live there. This estimate helps to improve the quality of census statistics, especially on Italian household characteristics, and to improve the overall evaluation of populations affected by misplacement errors.

Keywords: administrative data, machine learning, population census counts

1. Introduction

In official statistics production, administrative data are crucial for coping with budget constraints and less willingness on the part of respondents to participate in surveys [1]. Administrative data have undoubted advantages, such as: being inexpensive; having a census-like approach to collection, i.e. data are collected for all the units that are part of the administrative reference population; reducing the statistical burden.

Since 2020, Italy has produced population and housing census count estimates by relying exclusively on the administrative sources organized as 'Signs of Life' (SoL) in an *ad hoc* Integrated Data Base of Usual Residents (IDBUR). Through IDBUR it is possible to define "signs of life", that refer to activities carried out by anonymized individuals, usable for statistical purposes. Being self-employed or working for a company, being a civil servant, attending a school or university are examples of 'direct signs of life' [2, 3].

Other than IDBUR, the availability of data from the first cycle of the Permanent Census survey was a great opportunity to compare information. The response to the survey serves as a proxy for each individual's usual residence in a certain territory and therefore the survey is a perfect candidate to be used as a training set for any supervised classification.

The ML approach for counting the resident and the non-resident population in census statistics is a relatively new field of application and only a few case studies are available in the scientific literature [4]. In present work we applied ML tools on a SoL dataset, in particular we focused on the Support Vector Machine (SVM), which is a single-layer neural network and one of the most flexible and robust prediction methods [5].

The SoL approach depends greatly on the location of the signals reported in the administrative sources. In particular, place of residence discrepancies can cause misplacement errors, producing over and under-counting simultaneously within two different municipalities. To reduce misplacement errors we used utility consumption, electricity and gas, as data sources to identify the monthly consumption patterns associated with each point of delivery in Italy.

2. The accuracy of the population census counts

The ML model used for population census evaluation is the SVM, which is trained using features derived from IDBUR and units derived from the census survey. We have classified two possible states, usually resident or non-resident. Naturally there is a significant imbalance between the two classes both in terms of volume, there are many more usual residents than non-usual residents, and in terms of data quality, non-usual residents are often linked to the absence of information. This asymmetry of the two classes required the careful use of a training set, obtained from the sample surveys (Area Survey and List Survey) designed for census purposes. In particular, from the survey it was possible to obtain non-resident units when the conditions required by the statistical questionnaire were not met.

The main SoL considered in order to detect the usual resident or non-usual resident population in Italy are reported in Table 1. The signs measure the presence of individuals on the national territory. The 12 monthly signs are grouped in a variable which gives greater weight to the

continuous presences in the last months of the reference year. Other variables are of the type (0/1) presence/absence on the territory in the reference year.

Table 1: Variable considered for the classification problem (resident/ non-resident).

Variable	Description
WS	*- work and study signals*
MRC	*- communication from municipal register*
CONTR	*- active contracts (about car, house rental, real estate)*
ABROAD	*- signal of presence abroad (Income Database, Consular population register)*
PS	*- permits to stay from 2012 to 2021*
TAXID	*- Tax ID code*
BCOUNTRY	*- country of birth (Tax register)*
CLAGE	*- age (Tax register)*
CLSEX	*- sex(Tax register)*

In order to evaluate the relevance of the features considered, we carried out a principal component analysis. The expectation variation versus the principle components shows that the cumulative variance explained is equal to approximately 0.75 only after the fifth component, indicating a low correlation between the features considered and therefore the need for a multivariate approach to the ML classification algorithm.

2.1 Machine learning processes for an imbalanced training set

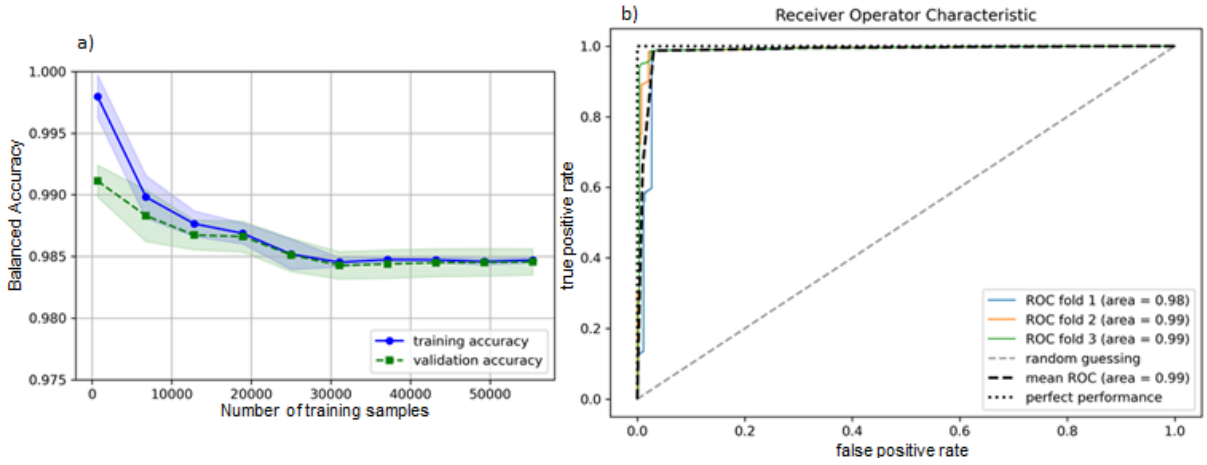
An information set that identifies the usual resident or non-usual resident population faces the challenge of performing classification in a class imbalance problem: the resident population is much larger than the non-resident population, at least in terms of presence of signs of life. This means that the classifier tends to be unbalanced towards the majority class. Typically, this problem is solved using sampling approaches, ranging from under-sampling to over-sampling, or using classification algorithms that are cost sensitive to class imbalance. Random over-sampling and random under-sampling have become standard approaches to improve classification performance although they both have serious drawbacks, such as tendency to overfitting in over-sampling approach. Alternatively, cost sensitive classifiers are also capable of generating informative models to overcome class imbalance problem. Many traditional methods are easily extensible with this goal in mind. Support vector machines, for example, can be cost sensitive and this is one of the reasons we chose it as a classifier. What is required in a cost-sensitive imbalance problem is to use the proper evaluation metric. In fact, accuracy and error rate overemphasize the performance of the majority class at the detriment to the considerations of the performance of the minority class. In order to overcome this issue, it should be used a Balanced Accuracy (BA), the semi-sum of TPR (True Positive Rate) and TNR (True Negative Rate), and the ROC (Receiver Operating Characteristic) curve.

2.2 Model choice and validation

We evaluated three kernel functions: the linear SVM model, the SVM-RBF (Radial Basis Function) kernel model, and the SVM-Polynomial model. The SVM – RBF was chosen as the best ML model. The choice of the SVM – RBF model and the values of the best fitting hyperparameters was based on a nested grid search selection. We obtained a score of about 99% using the following hyperparameters: *C(non-resident)=50* and *C(resident)=1* and *γ=1*, where *C* is inverse of the regularization parameter of each class.

To evaluate the robustness of the chosen model as a function of the size of the training set, we checked the learning curve (figure 1a) and the ROC curve (figure 1b), the size of the training sets as a function of their balanced prediction accuracy of training and validation sets. The figure shows how the two curves tend to overlap completely starting from a training set of 50,000 samples, indicating the minimum size of the training set that must be considered in this ML model.

Figure 1:a) training set as a function of their balanced prediction accuracy of training and validation sets; b) Receiver Operator Characteristic plot.



2.3 Analysis of the SVM results achieved

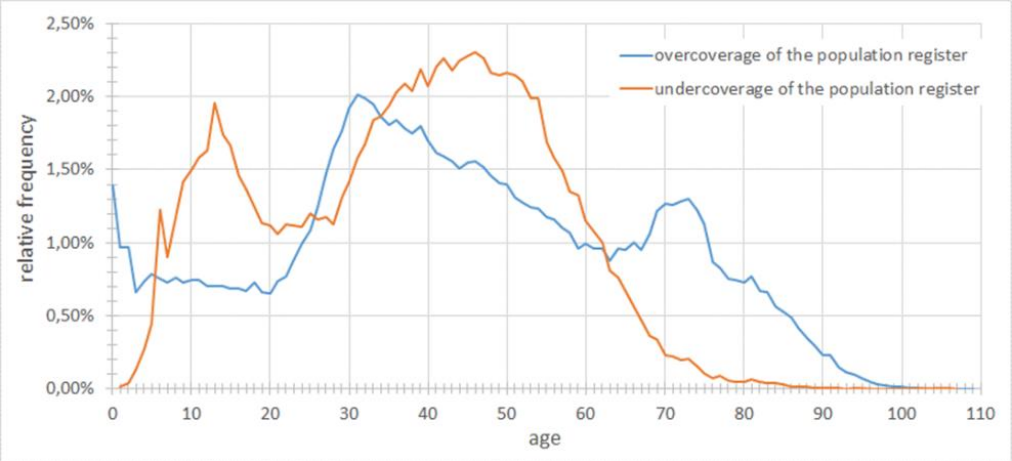
The ML classification results provide the usual resident and non-usual resident population in Italy and allow their comparison with the information in official population registers, based on municipalities. We identified four sub groups representing: non-resident, overcovered population, undercovered population and resident. The resident and non-resident are the two groups of individuals where the SVM output and the official population register are coherent, the overcovered population indicates non-residents for the SVM output who are classified as residents in the official population registers and the undercovered population indicates residents for the SVM output but non-resident in the municipal population registers. The results are shown in table 2, from which the quality of local population registers can be assessed, highlighting the potential role of SoI data and the ML approach.

Table 2: ML vs Population Register

Population Labels	ML	PR	Percentage
Overcovered	non resident	resident	0.33%
Undercovered	resident	non resident	0.10%
Not Resident (Matched)	non resident	non resident	1.76%
Resident (Matched)	resident	resident	97.81%

The relative frequency of the population over and undercoverage concerning the age of individuals is shown in figure 2. From the figure we can assess an overcoverage peak at around the age of 30, which is the age at which people tend to leave the country, mostly for work reasons. In this case, it seems that municipalities do not update the Municipal Register in a timely way. Undercoverage increases, however, at working age, when people may come to Italy for work, but are not always immediately identified as residents.

Figure 2: age versus relative frequency of the under and overcoverages



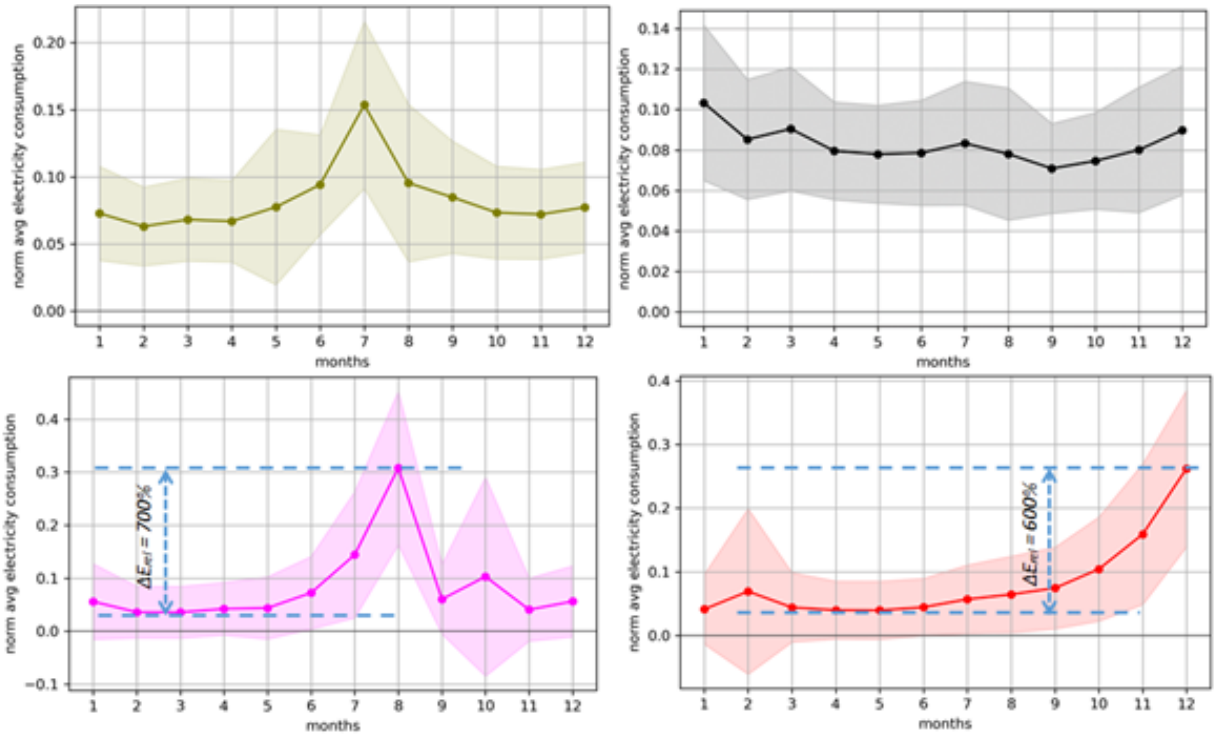
The phenomenon of overcoverage and undercoverage may be strongly correlated with the characteristics of the individual. Indeed, the population function by age could be further broken down by citizenship, showing different patterns for Italians and foreigners in terms of over/undercoverage. Furthermore, the analysis showed the important role of the variable derived from municipal register communications compared to all other SoL. In fact, Italy is particularly favoured compared to countries where official registers do not exist.

3. Evaluating the most probable place of usual residence based on utilities

Place of residence discrepancies can lead to misplacement errors. Through the combination of household information, included in the IDBUR, and utility consumption patterns, associated with each household, the usual place of residence can be assessed. The correct identification of the dwelling used helps to assess location errors and reduce potential statistical biases related to estimation processes of population characteristics at the sub-national level.

We used utility consumption, electricity and gas, provided by the Regulatory Authority for Energy, Networks and Environment (ARERA). All identification data were rendered anonymous, both with regard to the contract holder and the Point Of Delivery (POD). Therefore, it was possible to link an energy contract to an anonymous person, belonging to an anonymous household, only at municipal level. It was possible to distinguish between two household groups, which can be traced back to a single contract or to multiple contracts. Through the consumption profile of each POD it was also possible to estimate the number of users. This estimate helped evaluate the actual location of households, improving the overall quality of the census statistical counts.

Figure 3: consumption profile clusters: a-b) low variation consumptions; c-d) consumption with high annual variation



3.1 Identifying household consumption patterns

We first carried out a cluster analysis to identify the main consumption patterns in the data. In essence, we found four main consumption patterns groupable into two kinds of users:

- consumption for homes, i.e. consumption patterns with limited variations during the observed year (max variations of the order of 100%)
- consumption for not usual or residential homes; i.e. consumption patterns with high variations during the observed year (variations of the order of 1000%)

All PODs associated with a pattern show similar normalized dynamics during the observed consumption year depending on the season. Figure 3 shows the four clusters: figures 3 a and

b appear to represent two typical domestic consumption patterns, while figures 3 c and d appear to represent two consumption patterns with large annual variations, typical of holiday homes or any other seasonal activities.

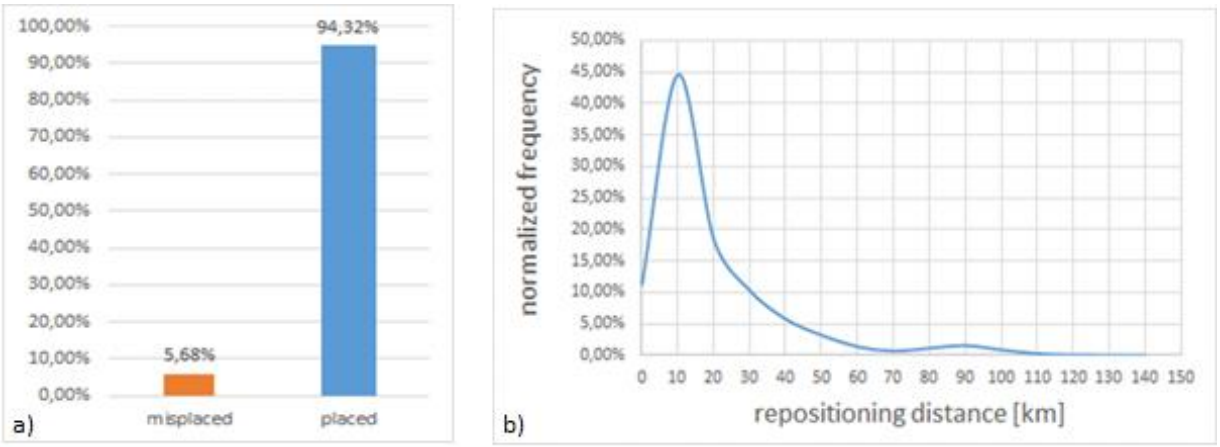
3.2 Defining a forecasting model

The data we worked on are 22,001,036 electricity contracts (13,776,499 civil homes; 8,224,537 legal units) and 14,516,785 gas contracts. In order to best analyse data, two groups were built:

A. families with a single contract: 13,136,120 electricity contracts
B. families with multiple contracts: 8,864,916 electricity contracts, for 3,383,369 families

We used the four consumption patterns as centroid in a k-means cluster analysis applied to all consumption data. The k-means metric was based on the Euclidean distance between different monthly consumptions, also including the second order derivative of the monthly trend. After the k-means clustering all consumption data was associated with a cluster. We used group A as a training set for building a regression function used to predict group B. The regression function allowed us to estimate the most probable consumption contract and the average number of family members in each cluster.

Figure 4: a) misplacement and well placed percentages; b) normalized frequency of misplacement distance



The average value of the prediction error (the number of predicted household members minus the number of household members in IDBUR) compared to the number of components in IDBUR is negative for a number of components lower than three and positive for above three. The number of members equal to three had the minimum error and reflects the average size of Italian families and therefore the average size of their flats.

Compared to the information from IDBUR, it is possible to estimate an error of approximately 5% in the positioning of the habitual residence. Figure 4b shows the misplaced normalized frequency versus misplacement distance of the family, showing a frequency peak at 10 km.

4. Conclusion and future work

This work has highlighted the richness of administrative data for census statistics. In particular, the use of machine learning techniques that learn directly from data appears to be very relevant because they are well suited to a complex scenario such as the census one.

It is very important to carefully evaluate the over and under coverage profiles from the ML model and the official population register in order to apply the proper correction to the official population register. One of the most critical points, the rare incidence of non resident population compared to the resident one, is dealt with by using imbalanced cost sensitive classifiers, supported by a dedicated audit survey [3] to improve the overall quality of the census count.

In addition, the stable acquisition of the consumption source in the production process can provide useful insights in order to increase the quality of territorial statistical estimations, which concern both territorial location and number of household members.

In conclusion, the production of census statistics, based on administrative registers and characterized by a significant volume of data, can benefit from the use of ML techniques through an improvement in the overall quality of the process, both in terms of performance and outputs.

References

[1] UNECE, (2021).*Guidelines for Assessing the Quality of Administrative Sources for Use in Censuses*, United Nations, New York, https://unece.org/sites/default/files/2021-10/ECECESSTAT20214_WEB.pdf

[2] Bernardini, A., Chieppa, A., Cibella, N., Solari, F., (2021). Administrative data for population counts estimations in Italian Population Census. In Perna C., Salvati N. and Schirripa F (Eds.) Book of short papers - SIS 2021, Pearson, pp. 274-278.

[3] Solari, F., Bernardini, A., Cibella, N. (2023). Statistical framework for fully register based population counts. Metron, Vol. 81, pp. 109-129.

[4] Zuppardo, M., Calian, V., Harðarson, Ó. , (2022) Machine learning estimation of the resident

Population, Statistical Journal of the IAOS, pp. 1–14

[5] Casari, A., Zheng, A., (2018). Feature Engineering for Machine Learning. Boston: O'Reilly Media Inc