DISCOVERING INDIVIDUAL PROFILES FROM ADMINISTRATIVE SIGNS OF LIFE USEFUL FOR THE ESTIMATION OF CENSUS RESULTS

Antonella Bernardini, Angela Chieppa, Tiziana Tamburrano

Abstract. The Italian Permanent Population Census (PPC) produces traditional Census results making use of administrative data integrated into statistical registers and survey data. Specific workflows validate administrative records and integrate data related to the same person, producing a standardized data structure that represents the so-called "signs of life" (SoL), referring to a specific reference date or period. SoL classifications and patterns are key for the Permanent Census strategy, especially for the estimation of the usual resident population: each individual in administrative sources is classified as resident according specific conditions related to Sol profiles. Moreover, quality assessment of Census population counts relies on SoL to design an audit survey. SoL can also significantly contribute to estimating thematic aggregates, adding new dimensions to what is collected with the census questionnaire. In this context, continuous evaluation and improvement of SoL classifications are essential. The availability of data from the initial waves of PPC provides a valuable opportunity for experimentation to uncover individual patterns by studying the statistical association between survey responses and the SoL of the same person. In this work, we present the initial results from pattern recognition to evaluate SoL profiles. The data used are derived from the integration of survey data collected in 2021 with administrative SoL for the corresponding year. Multiple Correspondence Analysis and Clustering are employed for an exploratory analysis. Subsequently, a supervised classification tree is used, with the response to the survey as the target variable, and SoL classification is considered among the independent variables. Some patterns and relevant features emerge and point out specific groups of interest as well as issues than call for further analysis and improved SoL classification.

1. Signs of Life for Population Census purposes: the case of the estimation of usual residents

The results of the Italian Permanent Population Census (PPC) are the output of some estimation processes based on survey data, registers, and specific variables derived from administrative sources (Bernardini et al., 2021). These latter sources are designed for administrative purposes, so they need proper processing to meet the quality requirements of official statistics production. Administrative databases also need to undergo processing to derive new features that could serve as variables specifically relevant to the estimation model, in which administrative data are intended to contribute.

The process for producing the census population counts consists of the estimation, for each individual candidate to be resident in Italy, of the dichotomic variable "usual resident (yes/no)" and the categorical one "place of usual residence" (territorial classification of the Italian administrative units). A specific thematic database (or register), called "Integrated Data Base of Usual Residents" (AIDA, from now on) has been implemented in the Italian National Statistical Institute (ISTAT) to exploit, at the individual level, the administrative sources and to derive new valuable information for population count estimation (Bernardini et al., 2019).

The main output of AIDA are the "Signs of Life" (SoL) that could be defined as *structured information derived from administrative sources after proper statistical processing (microlevel linkage, quality evaluation, classification according to expert knowledge, or pattern detection) and designed to support the estimation of usual residents in Italy and their place of residence*.

SoL classifications are key for the Permanent Census strategy. From 2020 on, they constitute the unique source for the estimation of the usual resident population and have also been used as covariates in other Census result estimation models. SoL relevance implies a strong effort to continuously improve and evaluate their quality.

1.1 Initial classification of SoL for Population Census purposes

Each presence of individuals in administrative sources provides data useful to build SoL, according to the definition in the previous paragraph. The sources of the signals are multiple and growing.

The main one is the National Register of the Resident Population (ANPR), that is the national database in which the municipal registry offices (municipal

“Anagrafi”) gradually converged during the latest years. ANPR is managed and fed on a local basis by each Italian Municipality, while the Ministry of Interior officers supervise it at the national level. ANPR is the administrative source with the highest quality to get data about people usually living in Italy, because it is specifically designed to store data about resident people, although for administrative purposes, and covers every local territorial unit. Nevertheless, this source still contains coverage errors, due both to individual habits of late or false personal registration and also to living conditions that are less stable than in the past. Data from ANPR are the primary source of the ISTAT Base Population Register (PBR) and determine the first computation of population amounts for specific dates and territorial units before the Census correction is delivered. Only ANPR data that comply with a set of quality checks defined at ISTAT are loaded into the PBR.

Additional sources are currently under study and will soon be introduced, including those related to electricity supply as well as mobile phone big data.

From the integration of all these sources available at ISTAT, direct or indirect signs of life are derived:

- *direct signs of life* – Activities performed by individuals from which a durable period of time (e.g., a year) and a place can be clearly identified (e.g., being a public or private employee, having a regular rental contract);
- *indirect signs of life* – An individual status or a non-professional condition (e.g., children or other relatives as dependents in tax returns).

The integration and loading of the different administrative sources and registers into AIDA are implemented through standardized workflows that run periodically and that produce specific entries in the database, each of these entries representing a structured version of SoL (ISTAT, 2022). Each SoL occurrence corresponds to a specific person and year (e.g., person A, year 2021) and contains various attributes that condense and summarize the integration of the administrative sources. The most important of these attributes are: 1) a sequence that represents the combination of the specific sources presenting information related to that person for the considered year; 2) a sequence that represents the monthly presence of the individual in the administrative sources over the 12 months of the year considered; and 3) the territorial units to which the administrative individual presence is related.

These attributes are the basis for computing SoL classifications that could be useful for Census estimation. Some initial analyses of the association between administrative data and place of usual residence (Chieppa et al., 2018) have determined the classifications currently implemented in AIDA. For direct SoL, the identification of duration patterns in administrative data has been translated into a specific classification that distinguishes between stable/continuous signals, seasonal signals, discontinuous signals, random signals, and absent or non-useful signals (Bernardini et al., 2019). Additionally, the initial analysis, coupled with expert knowledge, facilitated the formulation of specific computing rules to derive only one 'prevalent place' for each SoL. Moreover, a hierarchy among all indirect signals, established through expert evaluation of different administrative sources, has been utilized to classify each person based on a 'main indirect signal' when multiple signals are present. While these SoL classifications currently implemented in AIDA can label all individuals, there is an ongoing effort to identify new SoL classifications that can enhance the accuracy of predicting usual residence, particularly as additional data sources are integrated.

Table 1 shows the distribution of people with signal in at least one 2021 administrative sources, breakdown according to the presence in ANPR, and the an aggregation of the initial classification of SoL currently available in AIDA.

Table 1 – *Individual profiles based on initial SoL classification and presence in ANPR.*

Presence in ANPR	SoL initial aggregate classification	Counts	%
Not registered in ANPR	Steady signs of work/study	420.287	34,01%
	Signs of university enrollment	26.508	2,14%
	Weak signs of work/study	256.624	20,76%
	Signs others than work/study (permit to stay, rental contracts etc.)	263.722	21,34%
	Episodic presence	268.802	21,75%
Total individual entries with SoL, not registered in ANPR		1.235.943	100,00%
Registered in ANPR	Steady signs of work/study	31.620.418	52,94%
	Retirement/income source signs	16.926.345	28,34%
	Fiscally dependent family member	4.479.095	7,50%
	Weak signs of work/study	1.932.287	3,24%
	Indirect signs of life - several sources	1.218.389	2,04%
	Rental contract	1.042.950	1,75%
	Signs of university studies	991.543	1,66%
	Signs of work/study episodic	422.545	0,71%
	No signs of life	1.096.850	1,84%
Total people registered in ANPR		59.730.422	100,00%

AIDA 2021

The upper rows of the table are related to SoL of people not registered in ANPR but with at least one presence in another administrative source: they’re about 1,2 million of people for 2021, in 36,15% of cases have a strong/steady signal that could be considered as candidates to under-coverage of official population register.

The majority (97,46%) of more than 59 million individual records registered in ANPR have also administrative SoL that are coherent with registered place of usual residence in Italy. Nevertheless, 4 million of these individuals have administrative signs out of registered province.

Moreover, there are 2,5% of people in ANPR without any other administrative sign.

These results are the starting point for a deeper analysis on administrative data to improve SoL classifications, as described in following paragraphs.

2. Learning new classifications for SoL from the integration with census survey data

The availability of data from the first waves of PPC is a great opportunity to discover patterns in administrative individual data associated with usual residence. The response to the survey serves as a proxy for each individual's usual residence in a certain territory. Therefore, the survey outcome can be used to evaluate the administrative patterns relevant to predicting the resident population count. These same patterns can be used to determine improved versions of SoL classifications that can constitute covariates in prediction models using administrative data to estimate the population resident in Italy. In this section, we describe the first results of a multidimensional analysis of integrated data aimed at improving the existing initial SoL classification on duration and type of source, as well as basic individual and territorial characteristics.

Multiple Correspondence Analysis (MCA) and Clustering were employed for an exploratory, unsupervised analysis to unveil the complete association structure in the data, identify relevant dimensions, and detect 'natural' clusters. Subsequently, a decision tree classifier was utilized as a tool to provide a more detailed description of specific associations and patterns related to usual residence. The response to the Census survey served as the outcome, guiding the algorithm in its search for final groups.

2.1 Set up the experimental database for learning

Data from different sources have been integrated into a database useful for learning goals through linkage at the individual level of all the different sources and ensuring the coherence of the time or period referenced by the data.

Survey data have been loaded, taking into account the data collection and validation rules. The variables from surveys that could be very useful in analyzing administrative signs are: survey outcome (respondent or not found); place of residence in the previous year; duration of presence; use of the same or other accommodation for systematic travel; some additional information for people found; type of survey (areal or list); survey mode.

From the AIDA database, variables useful for deeper analysis are: the presence or absence of the administrative signal during the last available year; continuity pattern; type or source of signal; and coherence among localizations of signs from different sources.

From PBR, which manages all ANPR validated data but also other individual and territorial attributes, variables considered are: presence or absence on January 1st, on the date of the survey, and on December 31st. Moreover, some territorial classifications of the registered place of residence are also integrated, such as degree of urbanization, regions, and other geographical attributes.

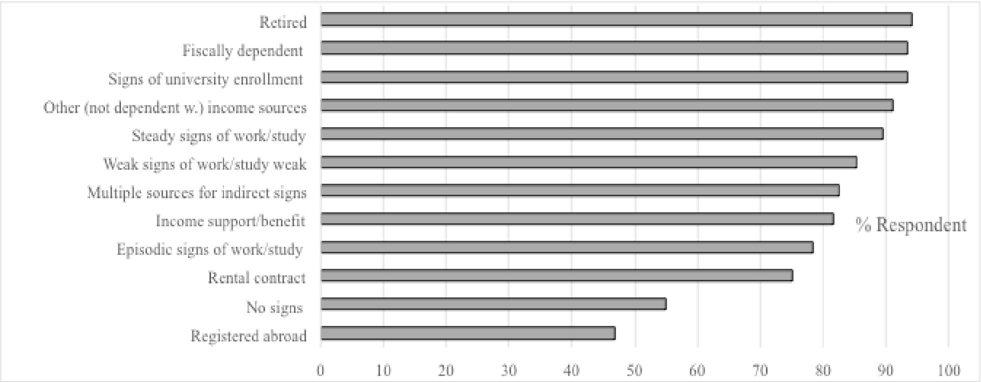
Finally, individual demographic variables are derived from available archives: gender, age, citizenship, and place of birth.

The results described in the following paragraphs are related to the analysis of a specific dataset extracted from the integrated database, where only data related to the year 2021 are considered. This dataset amounts to about 4 million individual records.

3. Multidimensional exploratory analysis on SoL and census survey outcomes

In Figure 1, it is possible to read the response rate at the List Census Survey executed in 2021, with the type of SoL corresponding to sampled people as a breakdown. The survey is executed on a sample of households and individuals extracted from PBR. Only those who confirm that the place of usual residence is the same resulting from the sample list have to respond to the questionnaire, that is to say that respondents are actually residents.

Figure 1 – *Respondents to census surveys and SoL – 2021 L CENSUS SURVEY.*



Respondents to census survey=resident people

For considerations above, the distribution in Figure 1 could be read as a first exploration of the effectiveness of SoL to predict usual residence.

The most evident result is the importance of the pension signal (94,6% of respondents for this group) and of the "indirect" fiscal signals (93,5% of respondents), even stronger than stable work or study signals. For weak signals, i.e., non-continuous, the response rate falls below 90%. People without signals have a response rate of 55%; this group of individuals is clearly critical for prediction based on SoL.

A multidimensional analysis is needed to evaluate the joint effects of SoL classes with other individual or territorial attributes.

MCA is an unsupervised technique for visualizing patterns in large and multidimensional categorical data (Greenacre and Blasius, 2006) by means of identifying principal dimensions that explain and synthesize the variability in the study dataset. Another powerful result when using MCA is the possibility of plotting in the same space defined by the resulting dimensions both categories and cases. When there is a statistical association among them, they are plotted near each other.

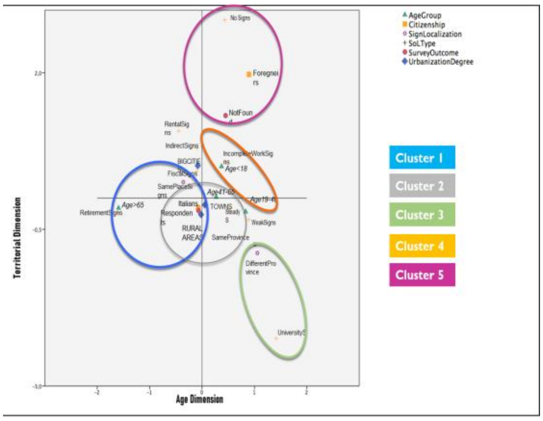
In Figure 2, we present the outcomes of a MCA applied to the dataset integrating SoL and survey data. These results aim to assess the relationships among all variables considered in this analysis, including gender, age, citizenship, urbanization degree, regions of residence, response to the survey, and SoL classes. Importantly, age and the degree of urbanization emerge as the primary dimensions explaining the majority of variability in the dataset, along with citizenship. In the graph, age is represented by the horizontal axis, with older ages on the left and younger ones on the right. The territorial dimension, with citizenship, moves along the vertical axis of the scheme: below we have rural areas and medium-sized municipalities, while above are more urbanized (cities) areas, with a strong associated presence of foreigners. This association structure is very typical when analyzing Census population phenomena on Italian territory.

A K-means cluster technique (Abidioun et al., 2023) is used to better detect significant patterns. Clusters are plotted in coloured ellipses in the same space resulting from MCA principal dimensions in Figure 2.

Five clusters result from the K-means algorithm:

- Cluster 1, with 54% of cases: Italians, respondents, older people, steady and fiscal and retirement signs; SoL with same localization of PBR.
- Cluster 2, with 27% of cases: Italians, living in medium towns or suburbs, adult ages, stable signs; SoL in different place than PBR but same province. This cluster seems to detect commuters.
- Cluster 3, with 3% of cases: Italians, respondents to survey, living in medium towns, university signs or weak SoL, with different localization than PBR. This cluster rapresents the students living seasonaly where university is located.
- Cluster 4, with 13% of cases: young people, absent or incomplete or fiscal SoL, half respondents half not; critical pattern, difficult to predict place of usual residence; could be Neet (not working or studying) or PBR overcoverage.
- Cluster 5, with 3% of cases: foreigners not founded with survey, registered in cities/urban areas, only rental contract sign or absent/incomplete SoL; critical pattern, difficult to predict if people with this SoL are still usually living in Italy by using only available SoL classes and individual/territorial attributes. Need for more information.

Figure 2 - *Clusters on SoL and survey outcomes.*

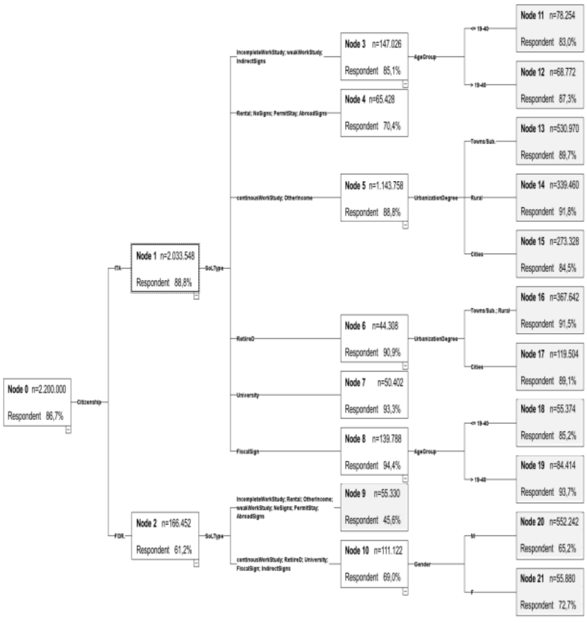


Both MCA and K-means techniques are unsupervised methods; that is, there is no variable to “guide” the pattern detection (Wu, 2012). Clusters derived from MCA and K-means undercover the associations among all considered variables. A supervised classification algorithm is needed to derive patterns that are especially relevant for the prediction of usual residence, considering response to a Census survey as a proxy for being a usual resident. In figure 3, there is a visual representation of the results of a classification tree algorithm: the same variables from MCA and k-means analysis have been used.

The classification tree is a very useful tool to share the results of a pattern recognition and classification model with thematic experts, since the resulting classification could be read through rules that are more easily understood than the parameter output of other predictive models.

The dependent variable chosen is the survey outcome, and the tree classifier adopted is CHAID (Ritschard, 2013), which makes use of the chi-squared association statistic to define, at each level, how to split cases into subgroups, starting from the entire study dataset. On each node, the algorithm splits cases according to rules on different categories of independent variables so that derived subsets are the most associated with the response variable (survey outcome).

Figure 3 - *Patterns according to response to the survey: classification tree result.*



Dataset: census survey sampled individuals CHAID classification tree

The model resulting when using CHAID on the study dataset (see figure 3) does not succeed in identifying rules to accurately predict usual residents in Italy, because almost all subsets detected have a higher percentage of respondents and prediction would be "resident" in almost all cases. Nevertheless, the tree is very useful for a description of individual and territorial profiles associated with the different probabilities of being resident in Italy. Tree results show that first splitting rule coincide with citizenship, that is to say that the patterns of Italians and foreigners respondents are different.

- For Italians, each of the existing SoL classes has a strong association with the probability of being resident: for instance, Italians with a fiscal SoL have a 94,4% of being confirmed, compared to only 70,4% in the case of Italians without administrative signs or only with a rental contract. In some cases (direct signs), the degree of urbanization is needed to better differentiate residents from those who are not found, while in other cases (indirect signs, such as fiscal ones or incomplete or weak signals), the age class is needed.
- Foreigners, on the other hand, have lower probabilities to be found with surveys; this could be related to both undercoverage of the survey or change of usual residence. Moreover, all different available classes of SoL combine in only two relevant groups, forming a critical pattern for prediction: all foreigners without signs, incomplete ones, or only with rental contracts have an almost equal probability of response and not being found; therefore, any prediction one makes (on the probability of response or residence) would make an error with a probability of about 50%. Moreover, the small size of this critical subset (only about 55 thousand individuals out of more than 2 million in the dataset) constitutes an added problem when adjusting a predictive model on these data.

4. Exploring data of critical no-signs group

In previous paragraphs, the group of individuals without signals, both foreigners and Italians, has resulted in a critical or difficult-to-predict pattern by using the variables of the study dataset.

To dissect this group further, additional targeted analyses were undertaken in an attempt to delineate subgroups and identify derived variables that might aid in classifying such cases.

The spatial distribution suggests the presence of distinct subgroups and potential varied living conditions. Noteworthy concentrations of this pattern are evident in several scenarios: municipalities along the country borders (potentially

Italians commuting to neighboring countries); holiday municipalities (lacking signals: "convenience" residents with second homes); southern municipalities where a mix of undeclared workers, genuine unemployed individuals (lacking signals but still residents), and emigrants not yet registered in the census coexist (lacking signals = pending cancellation due to non-residency).

Survey data significantly contribute to supplementing information for this group of people. There are 52.337 individuals without administrative signals who responded to the surveys. 90% of them declared to have resided in the same place or house one year prior, implying that the absence of SoL might not directly lead to removal from the residents register. Furthermore, 18.33% of these individuals indicated a working condition in the Census questionnaire, underscoring potential instances of illicit employment or underreporting in certain labor archives. Survey data from Census questionnaires regarding the marital status of these individuals reveal that 31% are married, suggesting the need to further explore household relationships or indirect signals in AIDA.

5 Some lessons learnt: changes in Population Census design and current experimentations

The use of administrative data for official statistics leads to major changes in the estimation of final official results (UNECE, 2020). Identification of different administrative-specific patterns, associated with each estimation output, is crucial for building the best integration and estimation strategy.

The first results from a multidimensional analysis to evaluate the association of available SoL classifications with usual residence, approximated by survey responses, point out very useful insights. For the majority of people eligible to be residents (at least one administrative sign), existing SoL classes together with individual demographic attributes and the degree of urbanization of the expected place of residence could be used to build accurate predictive models. On the other hand, groups such as foreigners in urbanized areas, people with no signals, and youth from the South with incomplete signs emerge as critical patterns in the sense that they call for an improved SoL classification since the existing classes proved not to be informative enough to predict their usual residence.

To get an accurate estimation also for these critical patterns and to improve existing SoL classes, the current study areas are: 1) improving the pattern recognition analyses by including machine learning techniques and by using all the data from the first cycle (2018–2022) of census waves (Casari and Zheng, 2018); 2) loading new administrative sources with high quality and coverage in AIDA (big data). Moreover, audit surveys are being planned to measure the quality of population estimates based on administrative data and to gain further insight and data for critical patterns (Solari et al., 2023).

References

ABIDIOUN M. I., EZUGWU A.E., ABUALIGAH L., ABUHAIJA B., HEMING J. 2023. K-Means Clustering Algorithms: A Comprehensive Review, Variants Analysis, and Advances in the Era ff Big Data, *Information Sciences*,Vol. 622, pp. 178-210.

SOLARI F., BERNARDINI A., CIBELLA N. 2023. Statistical Framework for Fully Register Based Population Counts, *Metron*, Vol. 81, pp. 109-129.

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, pp. 274-278.

BERNARDINI A., CIBELLA N., FASULO A., FALORSI S., GALLO G. 2019. Empirical evidence for population counting: the combined use of administrative sources and survey data. In *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 Media, Inc.

CHIEPPA A., GALLO G., TOMEO V., BORRELLI F., DI DOMENICO S. 2018. Knowledge Discovery for Inferring the Usually Resident Population from Administrative Registers, *Mathematical Population Studies.*, Vol. 26, pp. 1-15.

ISTAT. 2022. Nota tecnica sulla produzione dei dati del Censimento Permanente. Roma: Istituto Nazionale di Statistica.

GREENACRE M., BLASIUS J. 2006. *Multiple correspondence analysis and related methods*. New York: Chapman and Hall.

RITSCHARD G. 2013. CHAID and Earlier Supervised Tree Methods. In MCARDLE, J.J. and G. RITSCHARD (Eds.) *Contemporary Issues in Exploratory Data Mining in the Behavioral Sciences*, New York: Routledge, pp. 48-74

UNECE. 2020. New frontiers for censuses beyond 2020. Geneva, United Nations.

WU J. 2012. Cluster Analysis and K-means Clustering: An Introduction. In WU J. (Ed.) *Advances in K-means Clustering,* Berlin, Heidelberg: Springer, pp. 1-16.

Antonella BERNARDINI, Istat, anbernar@istat.it
Angela CHIEPPA, Istat, chieppa@istat.it
Tiziana TAMBURRANO, Istat, tamburra@istat.it