MEASUREMENT OF DAILY COMMUTING IN THE ITALIAN PERMANENT POPULATION CENSUS[1]

Carolina Ciccaglioni, Loredana Di Consiglio, Tiziana Pichiorri, Fabrizio Solari

Abstract. Recently, the Italian National Statistics Institute decided to move from traditional enumeration census to a combined census, whose backbone is given by statistical registers and administrative archives. The main goal of the new census process is to provide yearly information. In this work, the estimation process for both work-related and study-related commuting is presented. Predictive modelling approach is used to produce commuting estimates. Explicitly, a multinomial logistic model is assumed. Estimates are produced at municipality level using 2019 census survey data and the information included in statistical registers and administrative archives. Results are presented, displaying the relevance of the administrative data in the estimation process.

1. Introduction

Over the past few years, the Italian National Statistics Institute (Istat) has to face more and more complex information requirements referred to both the improve the supply and quality of statistical information and update to a data production system adherent to the other EU countries.

The goal of producing information on an annual basis and the possibility of using administrative data sources made necessary a reconsider the design of the Italian Population and Housing Census. Therefore, Istat has replaced the traditional census with a combined census - named Italian Permanent Population and Housing Census - placing the integrated system of statistical registers and administrative sources at the core of statistical production. Census sample surveys are planned to support registers, providing information where this is missing, incomplete or of unsatisfactory quality in registers and administrative sources.

This work describes the estimation process of the census tables (multidimensional tables describing certain phenomena) related to resident

[1]Sections 1 and 5 were written by L. Di Consiglio and F. Solari, sections 2 and 3.2 were written by T. Pichiorri, sections 3.1 and 4 were written by C. Ciccaglioni.

population commuting for work or study reasons. Commuting is usually associated to only work-related commuting flows. In this study commuting is intended in a broader sense, including also study-related commuting flows. In addition to this, attention is paid to commuting within and outside the municipality of residence.

Estimates were produced by means of predictive modelling, using a multinomial logistic model to combine survey data with register and administrative data.

Commuting is a relevant phenomenon in Europe which increased consistently in recent years, due to the changes in production systems that led to a higher labour mobility. Furthermore, improvements in transport and communication infrastructures has simplified goods and services' movements leading to commuter routes expansion.

This paper describes census tables estimation process referred to commuting population in 2019 and presents the first results.

Section 2 is dedicated to the target variables and target parameters definition. Section 3 is devoted to the estimation process description. The results are shown in Section 4 and final remarks are provided in Section 5.

2. Notation and definitions

As previously described, the aim is to produce estimates for work and study commuting flows. A commuter is defined as an individual making the same journey between home and place of work or place of study back and forth, at least three days per week. It is considered as a commuter also an individual moving inside his/her own municipality of residence. The target population is the resident population moving daily by work or study reasons, by place of destination (inside/outside the residence municipality - Italy or other countries), by gender (male/female), for all the Italian municipalities, that is at LAU2 level.

The estimation process takes as input the binary variable occupational status, according to which each individual is classified as either employed or non-employed. The occupational status is the result of a different estimation process. Specifically, it is predicted through means of a latent class model in an independent Istat estimation process (for details about latent class models, see Biemer, 2011). The definition adopted in the Italian census assumes that an employed individual can be only a work commuter and a non-employed individual can only be a study commuter. Furthermore, the variable gender is available for all the individuals in the population.

The following notation is used throughout the manuscript. Let q $(q = \text{work}, \text{study})$ and g $(g = \text{male}, \text{female})$ denote the generic commuting reason and

the generic gender, respectively. Besides, let j $(j = 1, \dots, M)$ be the generic municipality. The resident population in the municipality j can be classified according to their occupational status, i.e. employed and non-employed. Let ${}_gN_j^{\text{work}}$ and ${}_gN_j^{\text{study}}$ denote the employed and the non-employed population size in municipality j, respectively, having gender g. Furthermore, let i denote the generic individual in the municipality j, where or depending on whether is employed or not. Then, for each individual i in the municipality j, the target variable can be defined in the following way

$$Y_{ij}^q = \begin{cases} 0, & \text{if } i \text{ is not a commuter for the commuting reason } q \\ 1, & \text{if } i \text{ is a commuter inside } j \text{ for the commuting reason } q \\ 2, & \text{if } i \text{ is a commuter outside } j \text{ for the commuting reason } q \end{cases}.$$

The target parameters are given by the commuting rates for study and work at municipality level, that is, for $j = 1, \dots, M$, $q = \text{work}, \text{study}$, $g = \text{male}, \text{female}$,

$$\bar{Y}_{.j}^q(k) = \frac{\sum_{i=1}^{{}_gN_j^q} 1_{ij}^q(k)}{{}_gN_j^q}, \dots k = 0, 1, 2,$$

where

$$1_{ij}^q(k) = \begin{cases} 1, & \text{if } Y_{ij}^q = k \\ 0, & \text{otherwise} \end{cases}.$$

3. Estimation method

3.1. Model definition

Since the target variable Y^q is a categorical one, the natural choice is to assume a multinomial distribution for Y^q and a multinomial logistic model. Suppose a sample of a municipalities is selected from the overall set of the Italian municipalities. Furthermore, a sample of individuals is drawn within each sampled municipality. Let m be the size of the sample of municipalities and let n_j^q (q= work, study) denote the size of individuals in the sampled municipality j related to the employed and non-employed sub-populations. The multinomial logistic model

adopted for all the individuals $(i = 1, \dots, n_j^q)$, for all the municipalities $(j = 1, \dots, m)$ and for $q = \text{work}, \text{study}$ can be expressed as follows

$$\log \frac{\text{Prob}\left(Y_{ij}^q = k\right)}{\text{Prob}\left(Y_{ij}^q = 0\right)} = x_{ij}^q \beta_k, \dots k = 1, 2, \tag{1}$$

where x_{ij}^q is the value assumed for the individual i in the municipality j by the set of auxiliary variables chosen for the target variable Y^q.

Equation (1) implies

$$\text{Prob}\left(Y_{ij}^q = k | x_{ij}^q, \beta_k^q\right) = \frac{1}{1 + \sum_{k=1}^{2} \text{e}^{x_{ij}^q \beta_k^q}}, \dots k = 0$$

and

$$\text{Prob}\left(Y_{ij}^q = k | x_{ij}^q, \beta_k^q\right) = \frac{\text{e}^{x_{ij}^q \beta_k^q}}{1 + \sum_{k=1}^{2} \text{e}^{x_{ij}^q \beta_k^q}}, \dots k = 1, 2.$$

For $j = 1, \dots, M$, $q = \text{work}, \text{study}$ and $g = \text{male}, \text{female}$, the target parameters are estimated as

$$_g\hat{\bar{Y}}_{.j}^q(k) = \frac{\sum_{i=1}^{{}_gN_j^q} \hat{1}_{ij}^q(k)}{{}_gN_j^q}, \dots k = 0, 1, 2,$$

where $\hat{1}_{ij}^q(k) = \widehat{\text{Prob}}\left(Y_{ij}^q = k | x_{ij}^q, \hat{\beta}_k^q\right)$ is the predicted probability to assume the value k of the variable Y^q for the individual i in the municipality j under the model (1).

3.2. Model selection

The auxiliary variables used in the estimation process can be split into individual level and municipality level variables. The former set of variables includes individual socio-demographic and administrative variables. The list of the socio-demographic variables used in the models is given by gender, age, citizenship (Italian, foreigner) and educational level (illiterate, literate but no formal educational attainment, primary education, lower secondary education, upper secondary education, bachelor's or equivalent level, master's or equivalent level, doctoral or equivalent level). An important role is played by administrative variables denoting work and study activities. These variables provide information on location and reference month to which each activity refers to. The latter set includes information available at municipality level. In particular, these variables are geographical coordinates (longitude, latitude), population density and some municipality level indexes, that are administrative municipality classification (capital, regional chief town, provincial chief town), inner area classification (urban pole, inter-municipal pole, urban belt, intermediate area, peripheral area, ultra-peripheral area), degree of urbanization, mountainousness index, altitude index, climate index, seismic risk index. Furthermore, inter-municipal distances, commuting rates at 2011 census, municipality presence in smaller islands and, only for the study commuting variables, the presence of school establishments within the municipal territory (primary, lower secondary and upper secondary school).

The sampling design is supposed to be informative, since both sampling weights and target variables depend on the municipality population size. The above described municipality level variables are considered in the model because they are strongly correlated with the municipality population size and they are supposed to integrate the information deriving from the sampling weights (for instance, see Little, 1983).

Variable selection was carried out using a backward stepwise procedure based on the Akaike Information Criterion (AIC) (Akaike, 1973). Moreover, Classification and Regression Trees (CART) (Breiman et al., 1984) was used to validate the results provided by the AIC and to define proper re-classifications for some auxiliary variables. In detail, the segmentation algorithm suggested a reclassification of the variables age, distances between residence municipality and work/study municipality, assuming a non-linear relationship with the target variable. This analysis was carried out at national and regional levels, for both work and study commuting flows, within and outside the residence municipality.

Separate models are defined for employed and non-employed individuals, for work and study commuting. Non-employed individuals are highly heterogeneous in terms of commuting propensity and availability of predictive auxiliary variables

from administrative sources. Therefore, CART recommended to split the non-employed individuals into two sub-populations: individuals in compulsory education age, with ages ranging from 6 to 16 years, and the remainder, aged from 0 to 5 years and aged 17 years and older. For each different sub-populations, distinct model estimation was performed. Furthermore, for all commuting populations, distinct model selection was carried out for each Italian region in order to maximize the predictiveness of the adopted set of auxiliary variables.

As far as work commuting, the non-parametric analysis suggests that the most influential territorial variables are 2011 census commuting rates, inter-municipal distances. Moreover, significant variables are the degree of urbanization, the inner areas classification and municipality presence in smaller islands (for the regions Latium, Campania, Apulia).

With regard to the study commuting variables, the inter-municipal distances played a relevant role in the model fitting process. Scholastic institutes presence or absence turned out to be a strongly significant variable. Specifically, presence of primary and lower secondary schools for the 6-16 aged non-employed model and upper secondary schools for the other non-employed model resulted to be an important input for the target variable prediction.

Location of administrative signs of work or study (the same as the residence municipality, in the same province of the residence municipality, in the same region of the residence municipality, outside the region of the residence municipality) turned out to be the most predictive individual auxiliary variable. Other relevant individual auxiliary variables are citizenship, educational level and age-class.

4. Results

2019 is the reference year for the commuting census process presented in this work. The reference year for the commuting census process presented in this work is 2019, which consists in household survey and an area survey on enumeration areas and addresses. A two-stage sampling design was used for both survey components. Municipalities are the primary sampling units, while the secondary ones are given by households for first component and by enumeration areas or addresses for the component area. The surveys involved 2.848 municipalities and 2.836.208 individuals. Figures 1,2,3 display the commuting rates for all the Italian municipalities. Municipality level commuting rates are shown only for ‘commuting inside the residence municipality’ category and for ‘commuting outside the residence municipality’. No commuting individuals' rates are not displayed since they are not a census target.

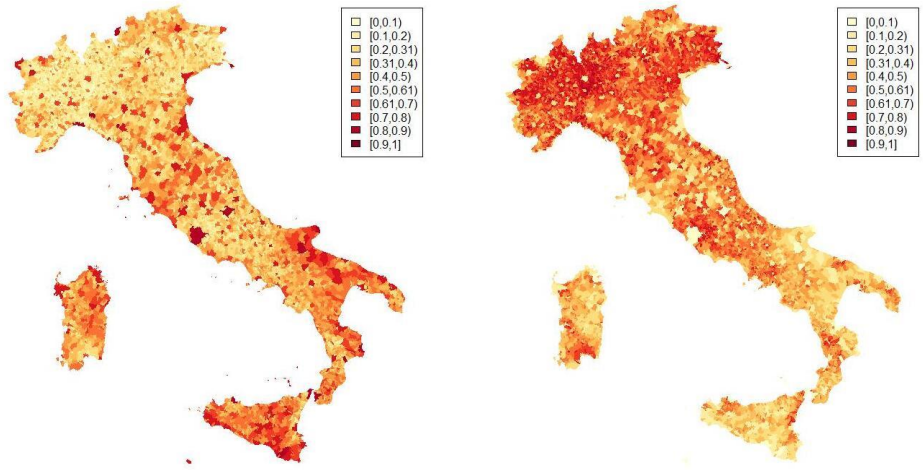
Figure 1a shows work commuting flows inside the same municipality mainly involve larger cities. Furthermore, this phenomenon is also relevant for urban and inter-municipal poles. High commuting rates are also observed for peripheral and ultra-peripheral areas with high degree of urbanization. Symmetric behaviour is found for the work commuting flows outside the residence municipality displayed in Figure 1b. As expected, work commuting flows are smaller for large municipalities and in general for municipalities with high level of urbanization.

With regard to study-related commuting, Figures 2a and 2b shows the commuting rates for 6-16 aged sub-population, inside and outside the residence municipality, respectively. It can be observed that the 6-16 study commuting displays analogous mobility trend to work commuting.

Commuting rates for not 6-16 aged sub-population are displayed in Figures 3a and 3b. Only largest municipalities and some peripheral and ultra-peripheral areas display large inside municipality commuting rates.

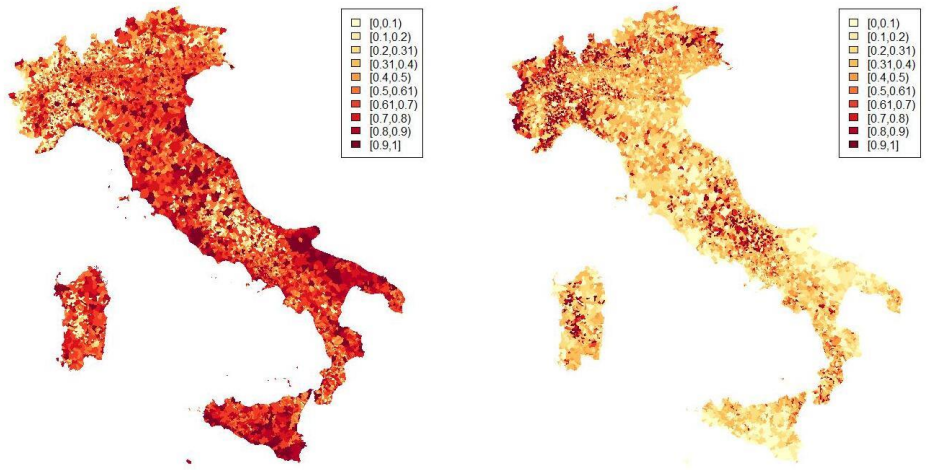
As a general comment, it can be noticed that commuting rates outside the residence municipality are higher in the Northern regions than in central and southern regions.

Figure 1 – *Work commuting rates: (a) inside the municipality, (b) outside the municipality.*



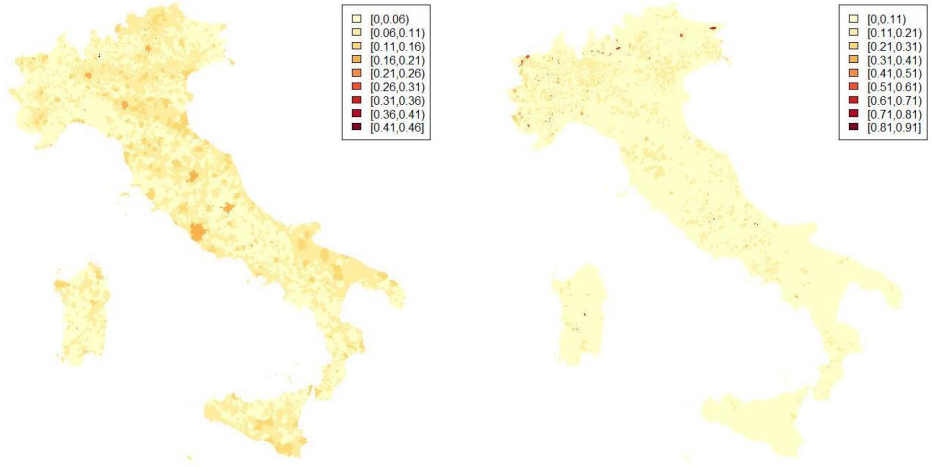
(a) (b)

Figure 2 – *6-16 aged study commuting rates: (a) inside the municipality, (b) outside the municipality.*



(a) (b)

Figure 3 – *No 6-16 aged study commuting rates: (a) inside the municipality, (b) outside the municipality.*



(a) (b)

5. Conclusions

This work describes the estimation process leading to the production of the census tables about work and study commuting. The estimation methodology is consistent with the modernization process embraced by Istat in recent years. Specifically, the modernization process provides for a general integration framework between register and administrative data on the one hand and survey data on the other hand (Falorsi, 2016). Here, coherence is attained by means of a model prediction approach, with survey data integrated with register and administrative data inside a multinomial model context. Alternative solutions can be provided by mixed effects multinomial logistic models (see, for instance, Agresti, 2013, Chapter 13) or adopting machine learning techniques, supervised or unsupervised learning. An example of application of machine learning technique is given in Ciccaglioni et al. (2022). Specifically, the authors carried out an experimental study implementing neural networks to provide estimates for work-related commuting.

Here, estimates and analyses refer to 2019. It must be underlined that commuting is a constantly evolving phenomenon, especially after the Covid-19 pandemic. Significant changes in commuting are expected after the introduction of smart work and study modes. In order to take into account the new situation, from 2021 specific answers have been added to the census survey questionnaire. Then, the measurement of commuting represents a real challenge for the next few years.

References

AGRESTI A. 2013. *Categorical data analysis (3rd edition).* Hoboken NJ: John Wiley & Sons.

AKAIAKE H. 1973. Information theory and an extension of the maximum likelihood principle. In: *Proceedings of the 2nd International Symposium on Information Theory*, Budapest: Akadémiai Kiadó, pp. 267-281.

BIEMER P.P. 2011. *Latent Class Analysis of Survey Error*. Hoboken NJ: John Wiley & Sons.

BREIMAN L., FRIEDMAN J., OLSHEN R., STONE C. 1984. *Classification and Regression Trees.* Monterey CA: Wadsworth and Brooks.

CICCAGLIONI C., DI CONSIGLIO L., PICHIORRI T., SOLARI F. (2022). Machine learning approach for estimation of commuting in the Italian Population Census, 7th ITAlian COnference on Survey Methodology ITACOSM2022, 8-10 June 2022, Perugia (Italy).

FALORSI P.D. 2016. Centralisation of data collection: a pillar of Istat's modernisation. In *Data Collection Workshop*, The Hague.

LITTLE, R.J.A. 1983. Comment on An evaluation of model dependent and probability sampling inferences in sample surveys by M.H. Hansen, W.G. Madow and B.J. Tepping. *Journal of the American Statistical Association*, Vol. 78, pp. 797-799.

McCULLOCH W.S., PITTS W. 1943. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, Vol. 5, No. 4, pp. 115-133.

Carolina CICCAGLIONI, Istat, ciccaglioni@istat.it
Loredana DI CONSIGLIO, Istat, diconsig@istat.it
Tiziana PICHIORRI, Istat, pichiorri@istat.it
Fabrizio SOLARI, Istat, solari@istat.it