ESTIMATION MODELS USED IN 2021 PERMANENT POPULATION CENSUS: CURRENT ACTIVITY STATUS, OCCUPATION, INDUSTRY AND STATUS IN EMPLOYMENT[1]

Diego Chianella, Carolina Ciccaglioni, Dario Ercolani

Abstract. Every year, Istat is involved in Census estimates production, which is obtained by integrating sample data with information from administrative sources. As regards the production of labour estimates, Istat produces Current activity status estimates for Italian dissemination. In 2023, in addition to these estimates, to comply with EU regulations, referred to the year 2021, estimates for Occupation, Industry and Status in employment were also produced. The estimation process for Employed/Not Employed is already well documented. For this reason, the procedure for estimating the Unemployed and Outside the labour force (students, retired people, housewives, other) and the variables related to Occupation, Industry and Status in employment will be here described. The estimates of these variables were produced at municipal level. The estimation models were implemented in R software through the ‘multinom’ function included in the ‘nnet’ package which allows to fit multinomial log-linear models using neural networks. The article aims to describe the different estimation models and the procedures used for choosing and defining the auxiliary variables included in the models. The administrative sources used were mainly social welfare and income sources and they allowed to identify ‘work signs’ of each individual.

1. Context and objectives

The transition from the traditional Census to the Permanent Census has required a total overhaul, both conceptual and methodological, of the estimation process of the census hypercubes. In detail, this evolution was made possible by the arrangement and use of administrative archives. This method provides up-to-date information on the population by integrating survey data with administrative data. A description of this process was provided in D’Alò et al. (2017) and Brogi et al. (2018). As a consequence, it was necessary to implement new estimation strategies to produce reliable estimates at given territorial levels (municipal and provincial).

One of the aims of the Italian Permanent Population and Housing Census is providing estimates of occupational, non-occupational status and so on. Following

[1] Although the document is the result of a joint effort of the authors, D. Chianella realized sections 3, 5, 7, C. Ciccaglioni sections 1, 4, 6 and D. Ercolani section 2. The views expressed in this article are those of the author and do not necessarily represent the views of Istat.

the Commission Regulation (EU) 2017/712 of 20 April 2017 about the production of census hypercubes, we computed estimates of *current activity status*, *occupation*, *industry* and *status in employment*. These variables are part of several census hypercubes. Below a description of the census variables related to the work topics and their breakdowns in accordance with the Commission Implementing Regulation (EU) 2017/543 of 22 March 2017:

- ‘Current activity status’ is the current relationship of a person to economic activity, based on a reference period of one week; his classification is shown in Table 1.
- ‘Occupation’ variable refers to the type of work done in a job (Table 2).
- ‘Industry’ **(**branch of economic activity) topic refers to the kind of production or activity of the establishment or similar unit in which the job of an employed person is locate (Table 3).
- ‘Status in employment’ classification is shown in Table 4.

Table 1 – *Current activity status classification.*

		Current activity status categories
1	Labour force	
	1.1	Employed
	1.2	Unemployed
2	Outside of the labour force	
	2.1	Persons below the national minimum age for economic activity
	2.2	Pension or capital income recipients
	2.3	Students
	2.4	Others

Table 2 – *Occupation variable classification.*

	Occupation categories
1	Managers
2	Professionals
3	Technicians and associate professionals
4	Clerical support workers
5	Service and sales workers
6	Skilled agricultural, forestry, and fishery workers
7	Craft and related trades workers
8	Plant and machine operators, and assemblers
9	Elementary occupations
10	Armed forces occupations

Table 3 – *Industry classification.*

	Industry categories
1	Agriculture, forestry and fishing
2	Mining and quarrying
3	Manufacturing
4	Electricity, gas, steam and air conditioning supply
5	Water supply; sewerage, waste management and remediation activities
6	Construction
7	Wholesale and retail trade, repair of motor vehicles and motorcycles
8	Transportation and storage
9	Accommodation and food service activities
10	Information and communication
11	Financial and insurance activities
12	Real estate activities
13	Professional, scientific and technical activities
14	Administrative and support service activities
15	Public administration, defence; compulsory social security
16	Education
17	Human health and social work activities
18	Arts, entertainment and recreation
19	Other service activities
20	Activities of households as employers; undifferentiated goods- and services producing activities of households for own use
21	Activities of extraterritorial organisations and bodies

Table 4 – *Status in employment classification.*

	Status in employment categories
1	Employees
2	Employers
3	Own-account workers
4	Other employed persons

An ‘employee’ is a person who works in a ‘paid employment’ job, that is a job where the explicit or implicit contract of employment gives the incumbent a basic remuneration, which is independent of the revenue of the unit for which he/ she works (this unit may be a corporation, a non-profit institution, government unit or a household).

An ‘employer’ is a person who, working on his or her own account or with a small number of partners, holds a ‘self- employment’ job and, in this capacity, on a continuous basis (including the reference week) has engaged one or more persons to work for him/her as ‘employees’.

An ‘own-account worker’ is a person who, working on his/her own account or with one or a few partners, holds a ‘self- employment job’ and has not engaged, on a continuous basis (including the reference week), any ‘employees’.

‘Other employed persons’ includes persons who are ‘contributing family workers’ and ‘members of producers’ cooperatives'.

2. Focus on available data sources and variables created

The Census is the only survey that allows the dissemination of data on employed people, people seeking employment and the inactive up to municipal detail. With the new system, census data on the labour and non-labour force no longer derive from exhaustive field surveys, but from the integration of information from administrative sources with that collected on a sample of families. To support the estimate process of census topics we selected the information available from the archives considered to be the most correlated with the professional and non-professional status. Specifically, with reference to all the units included in the Population Register (RBI), an information structure has been created extending its contents by integrating data regarding both employment and income, as well as social areas (pensions, subsidies and monetary grants). The information on employment contained in administrative sources is of fundamental importance for the purposes of measuring employment. The administrative content that will be exposed consists of an extension of the RBI through an individual archive called Prevalent Occupation Base (BOP), derived from the study of the Labour Register (RL), and through the integration of social and income information (several variables of the realized dataset are on Table 5). The RL is configured as a system of micro data at the basis of a series of internal processes of the Institute - that is, usable by various internal users - able to increase coherence between the processes themselves, reduce redundancies and unify the statistical treatment of the available archives on employed persons, their job, contributions and salaries. The information present in the RL is used as input for two other registers: 1) employment data based on Business Enterprise Registry (Asia); 2) annual register on wages, hours and individual Labour costs, aimed at producing and analysing the remuneration variables and the work input from the worker side starting from the employment relationship linking the worker to the economic unit. The different sources used were ordered in a hierarchical way on the basis of the study conducted on the quality of the sources, in terms of consistency and completeness of the information and are obviously different for each subgroup of workers. For the "private employees" component, supplies were used: UNIEMENS INPS source that represents the main source both in terms of coverage (approximately 90% of job positions come from this source) and in terms of information content; employees in agricultural sector INPS and CIG. The priority criterion between sources allows to create intermediate versions of the register even in the absence of one or more sources; these versions in fact allow to produce timely and accurate output to support other processes. As previously reported, starting from the identification codes in the RBI, which consistently align with the survey data, a data structure was created consisting of administrative information regarding the following fields.

Occupational field. With reference to the activities on the Labour Register, a database has been set up to allow the main occupation of the employed to be defined (BOP). This activity can be summarized in the deterministic identification, the investigation among all the work positions, observed through RL, of the individual and his main work activity, his work characteristics, connected to the ILO definitions, from which the specifications of the Labour Force survey (FOL) and Census Surveys are derived.

The reference population of the BOP consists in all the subjects identified in at least one of the administrative sources on employment available in Istat identified within RL. The aim is to identify, for each subject in at least one of the administrative sources on occupation acquired by the Istat and available for the reference year, his main work activity in a specific period of the year, limited to the month.

The basic definitional aspects applied to the data sources analysed are connected to the administrative information, from which the aim is to extract characteristics similar to the work definitions dictated by the International Labour Organization (ILO). Therefore, the classifications of work characteristics used in the FOL survey and in the Population Census were identified through administrative data.

The choice of the main work activity depends on criteria linked to the information power of the individual sources and on the comparison between different sources. Main sources: RL – Employed employment; INPS – Domestic work; INPS – Voucher; INPS – Parasubordinate Collaborators; INPS - Parasubordinate Freelancers; INPGI - Collaborators; INPGI - Freelancers; INPS - Artisans and traders; INPS - Autonomous Agriculture; INPS Ex Enpals – Self-employed workers; ASIA Enterprises – Individual VAT numbers with positive turnover. Auxiliary sources: Chamber of Commerce - Business Persons Archive; Chamber of Commerce - Shareholder Archive of joint-stock companies; ASIA Active and non-active companies.

Income aspects. The availability of fiscal sources acquired by the Ministry of Finance and the Revenue Agency has allowed the reconstruction of income from work, pensions and capital.

For the residents observed in RBI and their family characteristics, the individual indicator called equivalent income was calculated according to criteria defined by the OECD.

Social spheres. From the Population Register, information on educational qualifications and attendance at educational courses was extracted and partly reclassified, as well as the types of pension (source of origin: INPS Social Security Source) such as old-age, indemnity and welfare pensions, disability and survivors. The possibility of having the source on non-pension monetary treatments has made it possible to associate with each individual in RBI the possible sums paid by INPS and the related types of benefits regarding social unemployment benefits, family allowance for workers, household allowance family member for families with economic difficulties, maternity or sickness allowances, student subsidies.

Among the reconstructed information, the available variables, which can be used as covariates, are: personal data (sex and age); residence and citizenship; annual incomes; presence of pension; enrolment in study courses; type of administrative information about work.

The following table summarises the variables used and the way they were build.

Table 5 – *Several variables used and reconstructed in this work on employment estimation.*

Source		Individuals type	Variable
Labour Register (LR)	Employees – Social Security of Private and Public sector (INPS-UNIMENS)	Employees	- Work distance - Distance from last work signal in terms of months - Work signals in terms of yearly weeks - Continuity of administrative work signals presence - Qualify classification of Employees - Main job classification of Not-Employees
	Domestic sector Social Security	Employees	
	Journalism Social Security	Employees	
	Agricultural Autonomus Social Security	Not-Employees	
	Artisans and traders Social Security	Not-Employees	
	Freelance Employer Coordinated Social Security	Not-Employees	
	Freelance Self Employed Social Security	Not-Employees	
ASIA Business Register (VAT Self-employed)		Not-Employees	
ASIA Business Register (Enterprises)		Work charateristics	Work activity sector (NACE)
Individuals Register (RBI)		Students	Course of study frequency (0,1)
		Registry personal data	Gender / Age / Citizenship
Social Security Source (INPS)		Work retired persons	Retirement allowance (0,1)
Not Social Security Benefits source (INPS)		Unemployed persons	Unemployment benefit indicator
Income source (Agenzia Entrate – Ministero Economia e Finanze)		Income data	Earnings from employment (OECD)

3. Estimation Methodology

As mentioned in section 1, traditional decennial census was replaced by continuous data collection integrating survey and administrative data.

The sampling design used to draw the census sample involves a two-stage stratified approach: municipalities and households/individuals. Larger municipalities are surveyed every year, while smaller municipalities rotate annually, covering all municipalities over a five-year cycle. This ensures comprehensive data collection, even though not all municipalities are surveyed each year.

To produce estimates on professional and work conditions at the municipality level for 2021 (covering all the Italian municipalities), we used logistic multinomial models (Agresti, 2013) implemented in R with the "multinom" function from the nnet package, except for the estimation of employed individuals, which is carried out through latent class models (Boeschoten et al., 2021). It is important to note that individuals estimated as employed through latent class models were excluded from the units on which the multinomial models were applied. The models were run separately for each region. These models are designed to fit multinomial log-linear relationships using neural networks, allowing for accurate estimation of categorical responses.

The estimation process uses a variety of administrative data sources, including ISTAT thematic registers on labour, the statistical register on enterprises for self-employed workers and ISTAT base individual register for demographic data. Additional data sources provide specific information such as education records and pension benefits.

The response variable, representing the specific category j of interest (e.g., employment status), is observed in the census sample data, while the predictors include a rich set of individual and area-level variables. Individual-level predictors, derived from administrative data might include age, sex, citizenship, and employment history, while area-level predictors might encompass regional unemployment rates derived from the labour-force survey and geographical classifications.

For each individual i (excluding those under 15 years old and individuals estimated as employed through latent class models) in RBI, probabilities of belonging to different categories (j) are calculated based on the predictor values X, where X_i represents the vector of covariates for individual *i*:

$$log(P(Y_i = j\ |X_i)/P(Y_i = k\ |X_i)) = X_i\beta_j; \quad for\, j = 1, \dots, k-1,$$

where β_j represents the set of regression coefficients associated with the predictors for category j and k is the reference category in the model. From this relationship, we can derive the probability of belonging to each category j for individual *i*:

$$\hat{P}_{ij} = \hat{P}(Y_i = j \mid X_i) = exp(X_i \hat{\beta}_j) / \left(1 + \sum_{l=1}^{k-1} exp(X_i \hat{\beta}_l)\right).$$

Summing these probabilities within a specific domain, such as a municipality (M), provides the estimated number of individuals in each category *j*:

$$\hat{Y}_j^M = \sum_{i \in M} \hat{P}_{ij} = \sum_{i \in M} \hat{P}(Y_i = j \mid X_i).$$

The β parameters are estimated using maximum likelihood estimation (MLE). The nnet package internally sets up the likelihood equations and uses optimization techniques to find the parameter estimates that maximize the likelihood of the observing given data. This involves solving the following optimization problem:

$$\hat{\beta} = \arg max_\beta \sum_{i=1}^{n} \sum_{j=1}^{k} I(Y_i = j) \log P(Y_i = j \mid X_i, \beta),$$

where $I(Y_i = j)$ equals to 1 if the response variable for the individual *i* is in the category *j*, and 0 otherwise.

Variable selection for the models was carried out using Classification and Regression Trees (CART) models (Breiman et al., 1984), with model comparisons based on the Akaike Information Criterion (AIC) (Akaike, 1973) and the Bayesian Information Criterion (BIC). The BIC consistently favored simpler, more parsimonious models compared to the AIC, balancing model fit and complexity.

In addition, to select the best model, confusion matrices were generated for different models and their accuracy was evaluated. These confusion matrices were calculated on the test dataset, which was not used during the model training phase, following a standard cross-validation procedure. This approach ensures a more realistic assessment of the model performance and generalization capability.

Some covariates included in the model were grouped into classes to reduce model complexity, prevent overfitting, and increase computational efficiency. With a high number of categories, it was observed that some profiles had few or no observations, leading to potentially unstable or inaccurate estimates. By grouping categories, the number of observations per class increases, improving the stability of the estimates.

This methodology allows ISTAT to provide detailed and accurate estimates that inform policy-making, economic planning, and social services delivery at both national and municipal levels.

To produce a measure of accuracy associated with the municipal-level estimates, experiments are recently being conducted (Chianella et al. 2024) to account for both sampling error and model error. This experimentation is based on a previously introduced generic measure of global uncertainty (GMSE) (Alleva et al. 2021).

4. Current activity Status

Referring to the categories described in Table 1, the ‘labour force’ category (1) comprises all persons who fulfil the requirements for inclusion among the employed or the unemployed. ‘Employed’ category (1.1) was estimated through latent class model (as mentioned in section 3) and it is out of scope for this work.

‘Persons below the national minimum age (15 years old) for economic activity category’ (2.1) were derived from RBI and therefore is not part of the estimation process.

The remaining categories in Table 1 (1.2, 2.2, 2.3, and 2.4) were estimated using a multinomial logistic model (as describe in section 3).

The categories of the target variable, recorded in the census sample and used as the response variable in the model, has a broader classification: ‘unemployed’ (1.2) individuals are divided into ‘person seeking for first employment’ and ‘person seeking for new employment’; ‘others’ category (2.4) comprises ‘housewife’ and ‘other condition’.

The variables found to be significant in the model are both individual and municipal level variables:

- Individual level variables: gender, age group, citizenship, attendance of a course of study by RBI, distance from the last work signal (in classes), continuity pattern of administrative work signals (in classes), retirement indicator from labour statistical register, unemployment benefit indicator and labour income (in classes).
- Area level variables: provincial unemployment rate, target variables municipal level estimates from previous census and inner areas (urban, suburban, and rural).

5. Occupation

Referring to Table 2, the classification is effective for both employees and non-employees. The categories of the response variable ‘Occupation’ recorded in the census sample and used in the multinomial model align with those required by European regulations.

The predictor variables (X_i) used in the models include both individual level and area level variables:

- Individual level variables: gender, age group, citizenship, education level, main job classification of non-employees and qualify classification of Employees.
- Area level variables: provincial unemployment rate from the Labour Force Survey, estimates from the previous census and inner areas (urban, suburban, rural).

The covariates "Main job classification of non-employees" and "Classification of employees" are derived from the BOP database described in the previous paragraphs. The main data sources for defining these variables are detailed in Table 5.

The most significant covariate was found to be the Qualify classification of Employees. The categories for this variable are: Manual Worker, Office Worker, Middle Manager, Apprentice, Senior Manager and Other Employee. Unfortunately, this was the most detailed reconstruction possible from the administrative data. It was not possible to create a covariate matching the response variable more closely, which would have allowed for a higher model accuracy.

6. Industry

For the Industry variable, some individual values were directly derived from the register, while the residual part was estimated using a model based on administrative and survey data. It was not possible to link the administrative data on the type of enterprise for about 30% of the estimated employed people. In particular, for this portion of individuals, the model was fitted on the subset of the census sample without administrative signals on industry. In fact, the auxiliary variables were related to individual demographic characteristics (gender, age group, and citizenship), educational level and territorial features (urbanization degree, coastal areas).

For each individual the industry value was generated by means of random draws based on the estimated probabilities.

Model selection was carried out on the basis of AIC and p-values of regression coefficient analysis comparison.

7. Status in Employment

The variable Status in Employment, represents the type of employment status held by individuals and is categorized into specific types as required by EU regulations (Table 4).

The categories of the variable of interest, recorded in the census sample and used as the response variable in the multinomial model, follow a different classification system: Employee (1), Continuous collaboration worker (2), Occasional worker (3), Entrepreneur (4), Freelancer (5), Self-employed worker (6), Member of a cooperative (7), Contributing family worker (8).

The predictors (X_i) used in the model include both individual- and area level variables:

- Individual level variables: gender, age group, citizenship, education level, presence of income from self-employment, presence of income from employment and main job classification of non-employees.
- Area level variables: the same area-level variables used for the Occupation estimation are applied here.

The most correlated variable with the Status in Employment is “Main job classification of non-employees”. The categories for this variable are: Contributing family worker (1), Collaborator (2), Entrepreneur (3), Freelancer with employees from census sample (4), Freelancer without employees from census sample (5), Own-account worker with employees from census sample (6), Own-account worker without employees from census sample (7), Member of a cooperative (8).

It is evident that the categories required by European regulations, those recorded in the census sample, and those of the covariates differ from each other. Consequently, a mapping operation was carried out to align the categories of this covariate with those required in the census sample and those mandated by the regulation (Table 6). For example, individuals who responded in the census sample as "Entrepreneur”, along with “Freelancer” and "Own-account worker” and who also declared having employees in the census sample, were assigned to category 2 (employers). Following these mapping procedures, the multinomial model was run using the reclassified variable from the census sample with its four categories instead of eight. This alignment ensures consistency and compliance with the regulatory requirements while improving the accuracy and reliability of the estimates.

Table 6 – *Mapping Operation between European Regulation Classification of the "Status in Employment", Values Recorded in the census sample, and the Main Covariate Used in the Estimation Model.*

No	Status in Employment (Eurostat)	Status in Employment (census sample)	Main job classification of non-employees
1	Employees	1	Null + Cond[2]=Employee
2	Employers	4 + (5+6)*Employees	3+4+6
3	Own-account workers	2+3+ (5+6)*Not Employees	2+5+7+9
4	Other employed person	7+8	1+8

[2] “Cond” is another variable contained in BOP: it indicates whether the worker is an employee or not.

References

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

AKAIKE H. 1973. Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd International Symposium on Information Theory*, Budapest: Petrov, B.N., Csáki, F. (eds) Akadémia Kiadó, pp. 267–281.

ALLEVA G., PETRARCA F., FALORSI P. D., RIGHI P. 2021. Measuring the Accuracy of Aggregates Computed from a Statistical Register. *Journal of Official Statistics*, Vol. 37, No. 2, pp. 481-503.

BOESCHOTEN L., FILIPPONI D., VARRIALE R. 2021. Combining multiple imputation and hidden markov modelling to obtain consistent estimates of employment status. *Journal of Survey Statistics and Methodology*, Vol. 9, No.3, pp. 549–573.

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

BROGI F., CICCAGLIONI C., FALORSI S., FASULO A., QUONDAMSTEFANO V., SOLARI F. 2018. Preliminary experimental: results on the Italian Population and Housing Census estimation methods. In *Proceedings of Twentieth Meeting of UNECE Group of Experts on Population and Housing Censuses*, Geneva.

CHIANELLA D., DELIU N., DI ZIO M., FALORSI P.D., FALORSI S., ROCCI F., SIMEONI G. 2024. Process and output quality evaluation measures for Istat Integrated System of Statistical Registers. In *Proceedings of the Seventh International Conference on Establishment Statistics (ICES VII),* Glasgow.

D’ALO’ M., FALORSI S., FASULO A., SOLARI F. 2017. Estimation strategies combining different sources of data. In *Proceedings of 61st World Statistics Congress ISI 2017*, Marrakech.

Diego CHIANELLA, Istat, chianella@istat.it
Carolina CICCAGLIONI, Istat, ciccaglioni@istat.it
Dario ERCOLANI, Istat, ercolani@istat.it