



# Sae estimation of related labor market indicators for different overlapping areas

Michele D'Alò<sup>1</sup> · Danila Filipponi<sup>1</sup> · Silvia Loriga<sup>1</sup> 

Accepted: 23 March 2024

© The Author(s), under exclusive licence to Società Italiana di Statistica 2024

## Abstract

The aim of this study is to provide a comprehensive description of the statistical methodology used to produce estimates for various labor market variables at both the City and FUA levels, along with an analysis of the results obtained. To achieve this goal, small area estimates were computed using a unit-level multivariate model. This model was specifically designed to enable coherent estimation of the variables of interest collected by the Labour Force Survey, exploiting information derived from administrative data and statistical Registers. The use of such administrative data at the unit-level represents a novel approach to estimation based on Italian Labour Force Survey data. The estimator used in this work is based on a multivariate model implemented through the Mind R package, which was developed by Istat. The method presented in this study represents an extended multivariate version of the conventional linear mixed model at the unit level. To ensure consistency across different domains, a single cross-classification model was employed, encompassing all relevant domains of interest. The outcomes of this analysis reveal significant improvements in efficiency compared to direct estimates. This is particularly noteworthy in the estimation of unemployed individuals (both total and by gender), where direct estimates are prone to relatively high sampling errors.

**Keywords** Small area models · Labour force survey · Administrative data · Coherence · Functional urban areas · Cities

---

Michele D'Alò, Danila Filipponi and Silvia Loriga contributed equally to this work.

---

✉ Silvia Loriga  
siloriga@istat.it

<sup>1</sup> ISTAT: Istituto Nazionale di Statistica, Rome, Italy

## 1 Introduction

Eurostat plays a crucial role in providing a comprehensive understanding of European territories, particularly at the small area level, and in monitoring the targets set by European regional policies. To achieve this, Eurostat actively promotes the collection of statistical information on various aspects of quality of life, with a particular focus on Cities and their commuting zones, known as Functional Urban Areas (FUA). This data collection initiative, formerly recognized as Urban Audit, is a collaborative endeavor involving National Statistical Institutes, the Directorate General for Regional and Urban Policy, and Eurostat. It is worth noting that the provision of data is voluntary, as there is currently no EU legislation requiring the collection of these statistics. As a result, the availability of data differs from one topic to another and from year to year. The statistics related to Cities and FUA can be accessed through Eurostat's City database (EUROSTAT 2017).

This paper aims to provide a detailed description of the statistical methodology applied to produce estimates for a specific set of labor market variables at both the City and FUA levels. Furthermore, it aims to analyze the results obtained from these estimates. The estimation of labor market variables is executed through a unit-level multivariate model carefully designed to ensure internal coherence among all the estimated parameters of interest. The estimates are derived from data obtained from the Labor Force Survey (LFS), while relevant covariates are sourced from the Labor Register and Population Register. It's worth emphasizing that the use of this extensive set of administrative variables at the unit-level represents a novelty in small area estimation based on Italian LFS data: the SAE estimates regularly produced for Labour Market Areas indeed use a model with spatially correlated area effects and temporally auto-correlated effects, leveraging quarterly LFS time series, and utilize administrative auxiliary variables limited to the demographic composition of the population.

The provision of such statistics is outlined in the Grant Agreement "Sub-national statistics Italy" between Istat and Eurostat. The computation of these estimates serves as the specific objective of Work Package 3 in the Pilot Study "Small Area Estimation for city and functional urban area statistics."

The paper is structured as follows. In Sect. 2, an overview of the reference information context is provided. Specifically, the definition of the target variables and the desired level of territorial disaggregation is discussed in Sect. 2.1. Section 2.2 provides a comprehensive overview of the Italian Labour Force Survey (LFS) and includes an analysis of the direct estimates. Due to notable sampling errors observed in the direct estimates, particularly for the estimation of unemployed persons, the use of small area estimation methods becomes necessary. The auxiliary information used to specify the unit-level mixed model for computing the small area estimates is described in Sect. 2.3. This auxiliary data is derived from the Labour Register, Population Register, and other relevant administrative sources. Section 3 describes the employed method, which can be considered an extended multivariate version of the standard linear mixed model at the unit level. The analysis of the results is presented in Sect. 4, including model diagnostics

and a comparison between small area and direct estimates, to highlight the efficiency gains achieved with the former approach. Finally, Sect. 5 presents an analysis of the main results obtained from the small area estimates produced at the City and FUA levels for the years 2018, 2019 and 2020, the periods covered by the available administrative data. The key findings of the study are outlined in Sect. 6.

## 2 Data description

### 2.1 Target indicators and domains

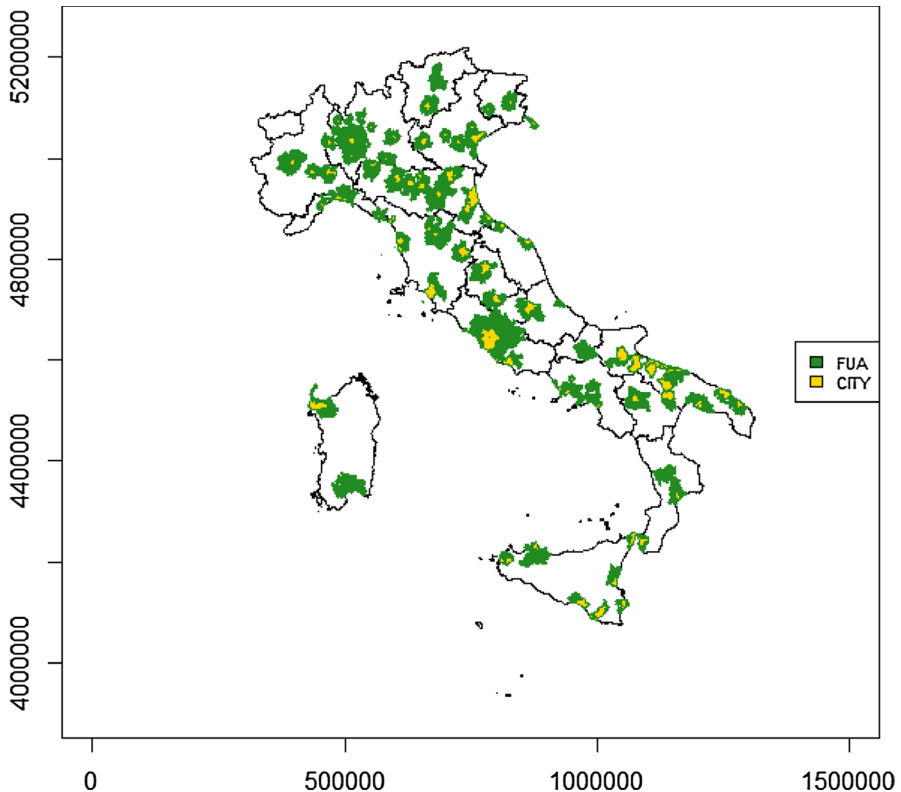
The goal of this paper is the estimation of the main labour market indicators for specific geographical domains. The target indicators are the two labour status, Employed Persons (EMP) and Unemployed Persons (UNE), along with Economically Active Population (EAP).

According to the definitions settled in European regulations, which are consistent with ILO definitions, Employed Persons are people having worked at least one hour, for pay or profit, in the reference week, or having a job or business but being absent during that week, for instance due to illness, holidays, etc.. Unemployed Persons are not employed people, being actively searching for a job during a four weeks period, constituted by the reference week and the three previous weeks and available to start working within two weeks. The last indicator, Economically Active Population, is given by the sum of employed and unemployed persons. A reference week is assigned to each household in the sample, representing the reference period for the information on the labour market participation; the allocation of the reference weeks is done assuring that the sample is uniformly spread over all the weeks of the year.

The classification of labor market statuses encompasses different domains and cross-classifications based on geographical areas, sex, and age groups. For Employed Persons, the domains are defined by geographical areas (Cities and Functional Urban Areas, FUAs). Within these areas, further cross-classifications are applied based on sex and a specific age group (20–64 years). Similarly, Unemployed Persons are categorized based on geographical domains, and within them, they are cross-classified by sex and a specific age group (15 years and above). Finally, the Economically Active Population is classified by geographical domains, and within these areas, by sex and two age groups (15 years and above, and 20–64 years).

The geographical domains are defined as:

- Cities that are local administrative units (municipalities) where at least 50% of the population lives in one or more urban centres, which are defined as clusters of contiguous grid cells of  $1 \text{ km}^2$  with a population density of at least 1.500 inhabitants per  $\text{km}^2$  and collectively a minimum population of 50.000 inhabitants.
- Greater Cities, that are formed by the group of local administrative units sharing the same high density cluster (in Italy this applies only in Milan and Naples)



**Fig. 1** Italian geographical areas: FUAs and Cities

- Functional Urban Areas, FUAs, that are formed by groups of local administrative units constituted by the City plus the municipalities included in its commuting zone. FUAs are based on the OECD-EC definition and they represent territories that are highly integrated from an economic point of view. FUAs (especially larger ones) often intersect the administrative boundaries of provinces and, in some cases, even of regions.

The spatial distribution of the small area of interest and how the Cities are nested within the FUAs in Italy is shown in Fig. 1.

The domains of interest show intersections with administrative units. In Italy there are 87 Cities and 83 FUAs; the FUA are present in 19 regions (Valle d'Aosta being the only exception without FUAs) and in 83 provinces. In the case of Milan and Naples, the Greater Cities are considered instead of the single Cities inside them. Both Greater Cities fall within a single region. Some considerations on the relationship between FUAs and Cities are:

- All FUAs have at least one City;

- The FUA of Milano has 5 Cities within its boundaries;
- The FUA of Bari has 3 Cities within its boundaries;
- The FUA of Roma, Napoli, Palermo have 2 cities within their boundaries;
- All other 80 FUAs have only one City within their boundaries.

For an insight into the definition of these geographical areas, refer to (EUROSTAT 2018).

## 2.2 The Italian labour force survey

The estimation of target indicators relies on data from the Labour Force Survey (LFS), a social survey which follows a two-stage sampling design. The primary statistical units are municipalities, while the final statistical units are households. In the first stage, Italian municipalities are stratified at the provincial level based on their population size. The largest municipalities form self-representative strata and are always included in the sample. The remaining municipalities are grouped into non self-representative strata, comprising municipalities with similar population sizes, and one municipality is selected from each stratum. In the second stage, a simple random sampling is used to select households within the municipalities chosen in the first stage. Once a household is selected, it participates in the survey four times, following a 2–2–2 rotation scheme. This means that the household is interviewed for two quarters, then temporarily excluded from the sample for two quarters, and finally interviewed again for two quarters. The planned domains for the survey are provinces and regions. Direct annual estimates are produced for provinces, while quarterly estimates are released for regions. Additionally, direct estimates are also computed for the 13 largest municipalities, which have a population exceeding 250,000 individuals.

The direct estimates are derived through a calibration estimator (Deville and Sarndal 1992). The primary goal is to ensure that the estimates of certain structural variables, derived from the sample, are consistent with the known totals of the corresponding variables in the reference population, obtained from external sources. The structural variables that are used in the LFS calibration refer to the distribution of the population in regions, provinces and 13 largest municipalities, by gender, age groups and citizenship. The use of such information allows to improve efficiency and coherence of the estimates. Calibration weights are computed using ReGenesee a software developed by Istat, as well as sampling errors, that are estimated exploiting the convergence for big samples of the calibration estimator to the corresponding GREG estimator. For further details on the Italian LFS methodology refer to ISTAT (2006).

Cities, Greater Cities, and FUAs are unplanned domains for the Labor Force Survey, therefore the survey's sample coverage varies across these areas of interest. The direct estimates for the twelve target indicators have been produced applying the same calibration estimator used for producing the survey's planned estimates. The accuracy of these estimates is assessed using their coefficient of variation. The coefficient of variation exhibits notable variation across different

**Table 1** Percentiles of the direct estimates CVs for the 12 target indicators, in Cities and Fuas domains, 2020 LFS data

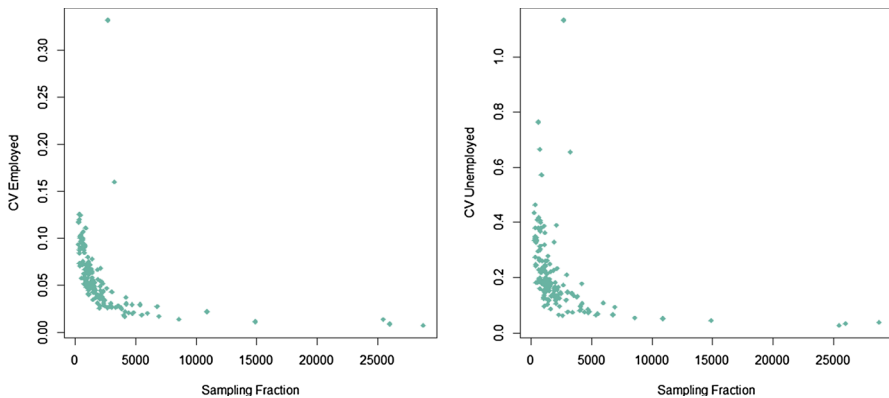
	min	pc5	pc10	median	pc90	pc95	max
eap_t	0.8	1.7	2.3	4.9	9.0	10.3	29.9
eap_m	1.1	2.4	3.2	6.8	11.8	13.3	42.8
eap_f	1.2	2.6	3.5	7.7	15.9	17.4	45.0
eap2064_t	0.8	1.7	2.4	5.0	9.1	10.5	31.4
eap2064_m	1.1	2.5	3.2	7.0	12.1	13.7	47.0
eap2064_f	1.2	2.6	3.6	7.8	16.2	18.0	45.1
une_t	2.7	6.3	7.6	17.9	38.0	41.9	113.4
une_m	3.6	8.3	11.4	25.3	54.3	63.5	190.3
une_f	4.3	9.6	11.5	26.6	57.6	80.1	177.6
emp2064_t	0.8	1.8	2.5	5.3	10.0	11.0	33.2
emp2064_m	1.1	2.5	3.4	7.3	12.8	14.5	49.2
emp2064_f	1.2	2.8	3.8	8.4	17.2	19.4	48.2

areas and indicators of interest. In Table 1 the percentiles of the coefficients of variation of the direct estimates of the twelve indicators of interest, in Cities and Fuas domains, are presented.

Focusing the analysis on the two principal indicators, employed persons aged 20 to 64 (EMP 20 – 64) and unemployed persons (UNE), it is worth noting that regarding employment, the coefficient of variation for the direct estimates in the year 2020 is acceptable for all areas. The highest coefficient of variation, reaching 33%, is associated with the FUA of Novara. Excluding this outlier, the highest coefficient of variation is 16% for the FUA of La Spezia. In contrast, the estimates of unemployment, being lower in magnitude compared to employment, exhibit higher coefficients of variation. Even after excluding the anomalous data for the FUA of Ferrara (which stands at approximately 113%), the coefficients of variation surpass 33% in 30 areas. They range between 16% and 33% in 73 areas, while 65 areas display coefficients of variation below 16%. It should be noted that coefficients of variation for unemployment estimates were lower in previous years due to higher estimate levels compared to the pandemic-affected year of 2020.

Figure 2 illustrates the relationship between the coefficients of variation of direct estimates and the sampling fraction in the areas, respectively for employed individuals aged 20–64 and unemployed individuals in Cities and FUAs. It is noteworthy that the highest coefficients of variation are observed in areas with the lowest sampling fraction.

The considerations regarding the characteristics of the domains of interest, the variability in sample coverage across areas, and the distribution of coefficients of variation calculated for the direct estimates of the twelve indicators all converge towards the necessity of employing small area estimation methods. This is crucial to effectively achieve the estimation objectives of this study, given the nuanced challenges associated with the surveyed areas and the need for more robust and accurate estimations in the face of varying sample coverage and data characteristics.



**Fig. 2** Distributions of the direct estimates CVs in Cities and Fuas domains, for employed 20–64 (left) and unemployed persons (right), versus the sampling fraction in the area. LFS 2020

### 2.3 Auxiliary information

In the past decade, National Statistical Institutes in Europe have introduced the use of Administrative Data (AD) in their statistical production processes. The Italian National Statistical Institute is currently undergoing a significant transformation of its statistical production processes through the implementation of the Italian Integrated System of Statistical Registers (ISSR) (ISTAT 2016). This system integrates administrative and statistical data within a unified framework to ensure consistent statistical processes and accurate outputs. The ISSR comprises four main Base Statistical Registers that collect relevant statistical units for official statistics. These registers include the Population Register, which contains information on individuals, households, and cohabitations; the Economic Units Register, which contains data on enterprises, farms, and institutions; the Places Register, which contains addresses, enumeration areas, and geographical coordinates; and the Labor Register, which contains job position information. The variables associated with these units, derived from administrative sources, are considered core variables as they can be identified at the unit level and remain stable over time. One major advantage of using the ISSR for statistical production is the increased availability of linkable information at the unit level.

In estimating labor market indicators, variables derived from the Labor Register play a central role. The Labor Register is a statistical register that integrates different social security and fiscal data, encompassing information on all Italian job contracts and social security details. Administrative sources related to employment differ in terms of quality and informativeness. Some sources provide detailed information on employment contract dates, while others only provide an overall signal for the entire year. Additionally, certain statistical units, such as irregular jobs or those with salaries below a certain threshold, may not be covered by administrative information. After preprocessing and data harmonization, the information is organized within an information system that establishes links between employers and employees,

**Table 2** Distribution of employment status by LFS and AD, 2020 data

	AD	
	OUT (Not employed)	IN (Employed)
LFS		
Not employed	62.0	2.3
Employed	3.0	32.7

with the primary unit of analysis being the employee job position (see (Baldi et al. (2018))). From this data structure, information on individual workers, including their employment status and characteristics based on International Labour Organization (ILO) definitions, can be derived. To ensure comparability with the ILO definition of employment and integration with the Labour Force Survey (LFS), the Register information is harmonized at the weekly level, providing weekly employment status information for each individual. To assess the discrepancies between LFS data and the Labor Register, the available data are linked at the individual level. Table 2 summarizes the distribution of employment status based on both LFS and AD, where AD employment refers to measurements taken during the same week as the LFS interview. The two measures show a low level of disagreement, approximately 6% of the interviews, with values outside the main diagonal in the table representing classification errors.

It is crucial to highlight that discrepancies in the information obtained from the LFS and AD can arise due to variations in definitions and data acquisition processes. Examples of these discrepancies include temporal misalignment of sources, especially for occasional jobs, a lack of administrative information in case of irregular work, and inconsistencies in the definition of employment across different available sources, among other factors.

Table 3 describes the variables available in the ISSR used in the study. The *equivalent income* consists of the family income divided by the equivalent dimension of the family (considering the presence of scale economies affecting the consumption needs of the family).<sup>1</sup>

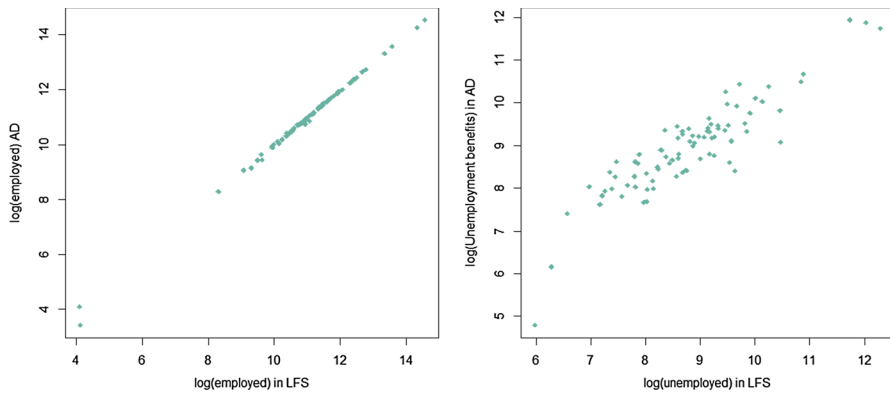
An exploratory data analysis was conducted to assess the potential predictive power of auxiliary variables. Figure 3 displays scatterplots illustrating the relationship between the logarithm of LFS employed and the logarithm of AD employed (left plot), as well as the logarithm of LFS unemployed and the logarithm of the number of individuals receiving unemployment benefits (right plot). The estimates in the scatterplots are based on the 2020 LFS sample data for Functional Urban Areas and Cities. From the scatterplots, it is evident that the first plot demonstrates a strong correlation between LFS data and Register data, indicating a reliable association between the employment measurements in the Labour Register and the LFS. On the other hand, the second plot exhibits a greater level of variability. These findings suggest that the employment measurements derived from the Labour Register serve

<sup>1</sup> The equivalent dimension of the family is obtained applying a specific equivalence scale (defined by the OECD): it is computed assigning value 1 to the first adult component, 0.5 to other components aged over 13 and 0.3 to the components that are aged 13 and less.



**Table 3** Auxiliary variable present in the ISSR

Register	Variable name	Values
Labor Register	Average number of working weeks	[0,1]
	Redundancy fund	0 = no, 1 = yes
	Unemployment benefits	0 = no, 1 = yes
	Households financial assistance	0 = no, 1 = yes
	Sickness and maternity allowances	0 = no, 1 = yes
Population register	Sex	M, F
	Age in class	1 = 0–15 2 = 16–19 3 = 20–24 4 = 25–29 5 = 30–34 6 = 35–39 7 = 40–44 8 = 45–49 9 = 50–54 10 = 55–59 11 = 60–64 12 = 65–69 13 = 70–74 14 = 75+
Fiscal data	Italian citizenship	0 = no, 1 = yes
	Education level	1 = Primary education 2 = secondary education 3 = university degree
	Enrolment in school or university	0 = no, 1 = yes
	Retirement pension	0 = no, 1 = yes
	Disability pensions	0 = no, 1 = yes
	Work income in class	1 = 0 2 = (0,10.000] 3 = (10000,25.000] 4 = (25000,50.000] 5 = 50.000+
	Pension income in class	1 = 0 2 = (0,10.000] 3 = (10000,17.000] 4 = (17.000,25.000] 5 = 25.000+
	Capital income in class	1 = 0 2 = (0,1.000] 3 = (1.000,6.000] 4 = 6.000+
	Equivalent income in class	1 = (0,6.000] 2 = (6.000,10.000] 3 = (10.000,17.000] 4 = (17.000,27.000] 5 = 27.000+ 6 = 0



**Fig. 3** Log of LFS employed and LFS unemployed versus log of AD employed and log of people with unemployed benefits in FUAs and Cities. Year 2020

as a valuable covariate for modeling employment conditions. However, the covariate for unemployment may have limited predictive power.

### 3 Small areas estimator

Small Area Estimation (SAE) methods are crucial for making inferences about finite populations, particularly when sample sizes within specific domains are insufficient to provide precise estimates using direct domain estimators. SAE is commonly employed to estimate the number of individuals in different labor market statuses at the small area level. In the literature on small area estimation, various approaches have been proposed for binary or multi-category responses, such as logit or multinomial mixed models, that shares the same random effects for all categories within each domain (Saei and Chambers 2003; Molina et al. 2007; López-Vizcaíno et al. 2013, 2015; Chambers et al. 2016; Marino et al. 2019 and Dawber et al. (2022)). In particular, López-Vizcaíno et al. (2013) and (López-Vizcaíno et al. 2015) proposed a multinomial logistic mixed model for SAE of multi-category responses that allows for category-specific random effects. However, these models are primarily designed to produce area-level estimates and rely on categorical explanatory variables aggregated within each area by category. Furthermore, as the number of categories increases, computing the empirical best predictor can become computationally challenging due to integrals lacking closed forms, and estimating the mean squared error (MSE) can be computationally prohibitive even with a limited number of areas.

The objective of this paper is to develop SAE modeling approaches at the individual level for two key reasons. The ISSR provides valuable unit-level variables that can be linked with survey data and possess significant predictive power. For instance, the labor register contains valuable data such as employment status, which exhibits strong correlations with the target variables of interest. By employing appropriate statistical modeling techniques, this auxiliary information

can be effectively incorporated into the estimation process, enhancing the accuracy and reliability of the estimates. Furthermore, the goal is to produce coherent estimates across different domains. With twelve related indicators of interest, a method that can better accommodate multiple constraints and interconnections among the estimated parameters is necessary. Modeling and predicting the target variable at the individual level facilitate harmonization of estimates across various domains and ensure consistency in the produced estimates. It is worth noting that the utilization of administrative information at the unit level represents a novel approach to estimating indicators of interest using labor force survey data.

The Battese, Harter, and Fuller model (Battese et al. 1988) is a highly influential paper that focuses on unit-level Small Area Estimation (SAE) models. These models take advantage of additional information that is available at the unit level. By incorporating both sampling information and auxiliary data known for each population unit, this model utilizes a linear mixed model to make predictions about parameters of interest in small areas. In this particular application, we employ a multivariate version of the standard linear mixed model at the individual level. This approach allows us to generate predictions at the unit level, take into account the multivariate nature of the data and address computational challenges effectively. It is important to underline that the use of normal models for binary variables is generally discouraged in classical statistics. However, when it comes to small area statistics, there is no clear-cut evidence demonstrating the definitive superiority of logistic models over normal models in terms of performance (D'Alò et al. 2012; Ranalli et al. 2018). Moreover, area-level models generally offer fewer efficiency gains compared to unit-level models, as the linkage between register and survey data allows for more predictive unit-level models. The estimator used was implemented through the Mind R package - Multivariate model based INference for Domains -, developed by Istat (see D'Alò et al. (2021)). This method can be viewed as an expanded multivariate version of the standard linear mixed model at the individual level. By incorporating various extensions, it allows for the specification of a multivariate linear mixed model, following the approach described in Datta et al. (1999). Additionally, this approach enables the inclusion of multiple random effects in the model, offering in this way the flexibility to fit possible marginal effects. These marginal random effects, in addition to or instead of the usual random area effects, can be particularly advantageous when dealing with a substantial number of small or out-of-sample areas. By incorporating these marginal random effects, the synthetic part of the EBLUP can better manage the bias of the predictor and produce less smoothed estimates. The marginal random effect can be derived from the variables used to define the strata in the sampling design or from other variables employed in defining the planned domains or by other relevant groups obtained by cross-classifying the the population units.

The model can be expressed as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{u} + \mathbf{e} \quad (1)$$

where

- $\mathbf{y}$  and  $\mathbf{e}$  are respectively the vector of the sample values of the target variables and of the residuals, of  $(n \times C)$  elements, where  $n$  is the number of units observed in the sample, while  $C$  represents the number of categories assumed by the target variable;
- $\mathbf{X} = X \otimes \mathbf{I}_C$ , where  $X$  is the design matrix of the sample values of the auxiliary variables considered for the fixed effects whose dimension is  $(n \times G)$ , with  $G$  being the number of variables (or the categories in case of categorical values) considered in the model and  $\mathbf{I}_C$  is an identity matrix of  $C$  order. The order of matrix  $\mathbf{X}$  of the multivariate model is  $[(n \times C) \times (G \times C)]$ ;
- $\boldsymbol{\beta}$  is the vector of the regression parameters whose length is  $(G \times C)$ ;
- $\mathbf{Z} = Z \otimes \mathbf{I}_C$ , where  $Z$  is the design matrix of the sample values of the random effects of dimension  $(n \times Q)$ , where  $Q$  is the total number of modalities of the random effects considered in the model. Obviously the  $\mathbf{Z}$  matrix of the multivariate model is of  $[(n \times C) \times (Q \times C)]$  order;
- $\mathbf{u}$  is the vector of random effects whose length is  $(Q \times C)$ .

Assuming that the vector of variance components  $\boldsymbol{\omega}$  is known, the *Best Linear Unbiased Predictor* or *BLUP* of the population target parameters  $\boldsymbol{\eta}$  is obtained using the sample data  $\mathbf{y}$ :

$$\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\eta}}(\boldsymbol{\omega}, \mathbf{y})$$

In real situations the variance components are usually unknown, therefore plugging in an estimator  $\hat{\boldsymbol{\omega}} = \hat{\boldsymbol{\omega}}(\mathbf{y})$  of  $\boldsymbol{\omega}$  the corresponding *Empirical Best Linear Unbiased Predictor* (*EBLUP*) is obtained (Rao and Molina 2015) as

$$\hat{\boldsymbol{\eta}}(\hat{\boldsymbol{\omega}}) = \mathbf{X}\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}) + \mathbf{Z}\hat{\mathbf{u}}(\hat{\boldsymbol{\omega}}) \quad (2)$$

where  $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}})$  and  $\hat{\mathbf{u}} = \hat{\mathbf{u}}(\hat{\boldsymbol{\omega}})$  are respectively the estimators of  $\boldsymbol{\beta}$  and  $\mathbf{u}$ , obtained plugging in the estimates of  $\hat{\boldsymbol{\omega}}$  into the correspondent *BLUP* estimator,  $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\boldsymbol{\omega})$  and  $\hat{\mathbf{u}} = \hat{\mathbf{u}}(\boldsymbol{\omega})$ , of  $\mathbf{u}$  defined under the hypothesis that  $\boldsymbol{\omega}$  were known. The estimate of variance components of the mixed effects model is computed with the REML derived by the algorithm proposed by Chambers and Clark (2012), using the approach developed by Fellner (1987) and Harville (1977).

In order to produce the target indicators at the required domains, the Mind R package has allowed to:

- Specify the dependent variable as a vector of three dichotomous variables representing the labor market categories: employed, unemployed, and inactive. It is important to note that these three groups encompass the entire population and are mutually exclusive.
- Ensure coherence among various domains throughout the definition of a single unit level incorporating category-specific random effects for the marginal groups, which correspond just to the geographical domains of interest.
- Obtain EBLUP and MSE estimate within each territorial domain further categorized by sex and age groups. Specifically, indicators for the employed refer to the 20-64 age group, indicators for the unemployed cover individuals aged 15 and

above, and indicators for the economically active population encompass both age groups. The estimate of target indicators are obtained specifying in a proper way the matrices  $X$  and  $Z$  in the predictor (2)

Let  $y_{d,j,k}$  represent the multivariate target variables observed in a specific area  $d$ , within an age group  $j$ , and for a particular sex category  $k$ . The relationship between these variables can be expressed as:

$$y_{d,j,k} = X_{d,j,k}\beta + \tau_d + e_{d,j,k}$$

The equation provided represents a particular specification of the general model (1). In this special case, EBLUP can be computed for each area within a more detailed age group of interest, denoted by  $j$ , along with two sexes. The age group classification is defined as follows:  $j = 1$  corresponds to the age class [15, 20),  $j = 2$  corresponds to the age class [20, 64), and  $j = 3$  corresponds to the age class 64+. Regarding the sex category,  $k = 1$  represents males, while  $k = 2$  represents females. The random effects are incorporated at the marginal level, specifically at FUAs and Cities level, respectively.

#### 4 Analysis of the results

In this paragraph, some analysis of the small area estimates produced at the City and FUA levels are described. These estimates are evaluated on the basis of two aspects: goodness of fit and comparison with the direct estimates. The results discussed here specifically refer to the year 2018, as it was the initial year in which the estimator has been applied. However, similar outcomes were observed for the subsequent years 2019 and 2020. The goodness of fit analysis assesses how well the estimation model aligns with the observed data, by examining various statistical measures, such as the coefficient of determination (R-squared). The comparison between model-based small area and direct estimates allows to evaluate the consistency and reliability of the model-based approach. By quantifying the differences between the two sets of estimates, we gain insights into any systematic biases or discrepancies that may exist.

The selection of auxiliary information and specification of the fixed part of the linear mixed model were based on analysis of the relationship between the variables of interest and the available group of covariates. Specifically, the aim is on identifying the auxiliary information associated with the response variables, which allowed for a more precise model specification. The variable selection process was carried out separately for the employed and unemployed variables, taking into account both Cities and FUAs. To accomplish this, a stepwise selection of relevant covariates has been applied to the full model specified for employed and unemployed. Most of the covariates demonstrated significant associations with the response variables, indicating their relevance in the models. Minimal differences were observed among the different models, suggesting that the selected covariates had consistent effects across various models. Finally the

**Table 4** Model fitting result for employed

Variable	Estimate	Std.Error	t	p-values	
(Intercept)	2.79E-02	2.71E-03	10.3	2.00E-16	***
sex = F	-2.30E-02	8.98E-04	-25.598	2.00E-16	***
SCHOOL = 1	-7.20E-03	2.02E-03	-3.56	0.000371	***
UN.BEN = 1	-7.81E-02	2.47E-03	-31.682	2.00E-16	***
EMP.ADMIN	6.83E-01	1.68E-03	407.878	2.00E-16	***
PENSION = 1	-5.77E-02	2.25E-03	-25.662	2.00E-16	***
CL.WORK.INCOME = 2	4.95E-02	1.34E-03	37.074	2.00E-16	***
CL.WORK.INCOME = 3	1.51E-01	1.87E-03	80.884	2.00E-16	***
CL.WORK.INCOME = 4	1.57E-01	2.17E-03	72.279	2.00E-16	***
CL.WORK.INCOME = 5	1.69E-01	3.16E-03	53.349	2.00E-16	***
AGE = 2	4.30E-03	2.62E-03	1.64	0.101103	
AGE = 3	2.15E-02	2.55E-03	8.439	2.00E-16	***
AGE = 4	8.49E-02	2.88E-03	29.447	2.00E-16	***
AGE = 5	1.27E-01	2.89E-03	44.003	2.00E-16	***
AGE = 6	1.31E-01	2.80E-03	46.771	2.00E-16	***
AGE = 7	1.34E-01	2.71E-03	49.58	2.00E-16	***
AGE = 8	1.41E-01	2.58E-03	54.666	2.00E-16	***
AGE = 9	1.43E-01	2.52E-03	56.785	2.00E-16	***
AGE = 10	1.31E-01	2.50E-03	52.479	2.00E-16	***
AGE = 11	8.25E-02	2.49E-03	33.104	2.00E-16	***
AGE = 12	2.55E-02	2.61E-03	9.751	2.00E-16	***
AGE = 13	4.15E-02	2.80E-03	14.816	2.00E-16	***
AGE = 14	4.66E-02	2.58E-03	18.088	2.00E-16	***
CL.EQUIV.INCOME = 2	-6.72E-03	1.71E-03	-3.928	8.56E-05	***
CL.EQUIV.INCOME = 3	-1.20E-02	1.52E-03	-7.911	2.55E-15	***
CL.EQUIV.INCOME = 4	-1.54E-02	1.56E-03	-9.888	2.00E-16	***
CL.EQUIV.INCOME = 5	-2.42E-02	1.70E-03	-14.196	2.00E-16	***
CL.EQUIV.INCOME = 6	5.70E-02	1.41E-02	4.051	5.10E-05	***
CL.PENS.INCOME = 2	-1.53E-02	2.83E-03	-5.411	6.26E-08	***
CL.PENS.INCOME = 3	-2.19E-02	2.91E-03	-7.517	5.61E-14	***
CL.PENS.INCOME = 4	-1.58E-02	2.90E-03	-5.45	5.05E-08	***
CL.PENS.INCOME = 5	-2.68E-02	2.90E-03	-9.236	2.00E-16	***
ED.LEVEL = 2	1.65E-02	1.13E-03	14.629	2.00E-16	***
ED.LEVEL = 3	4.55E-02	1.48E-03	30.762	2.00E-16	***

model selected for the unemployed was adopted, considering that the unemployed is the most “critical” variable to predict, due to the deficiencies of covariate.

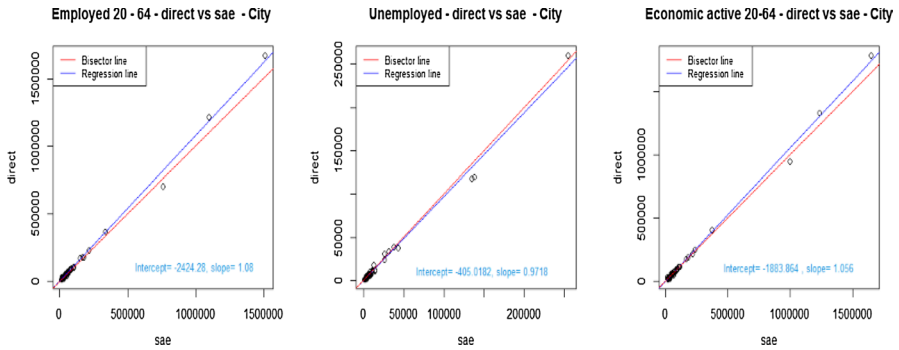
Tables 4 and 5 display the regression parameter estimates and relative standard errors for both the employment and unemployment categories. Additionally, they provide the p-values for testing the null hypothesis  $\beta = 0$ . Table 6 shows the goodness of fit analysis. The coefficients of determination  $R^2$  show a very good

**Table 5** Model fitting result for unemployed

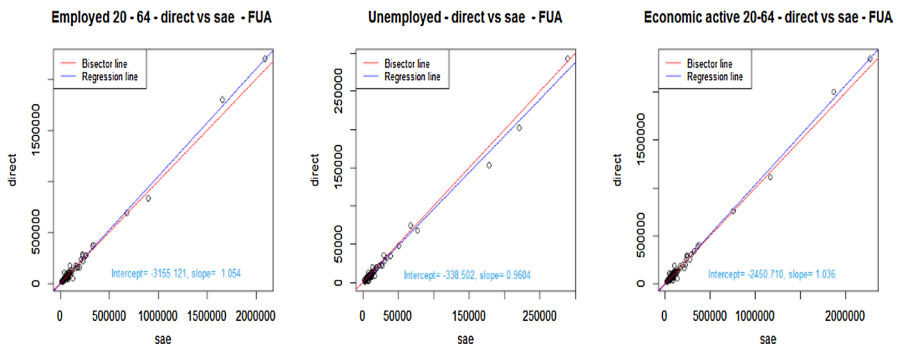
Variable	Estimate	Std.Error	t	p-values	
(Intercept)	4.14E-02	2.06E-03	20.091	2.00E-16	***
sex = F	-1.82E-02	7.34E-04	-24.755	2.00E-16	***
SCHOOL = 1	-1.46E-02	1.65E-03	-8.805	2.00E-16	***
UN.BEN = 1	1.38E-01	2.02E-03	68.404	2.00E-16	***
EMP.ADMIN	-1.00E-01	1.37E-03	-73.321	2.00E-16	***
PENSION = 1	-4.08E-02	1.84E-03	-22.218	2.00E-16	***
CL.WORK.INCOME = 2	1.91E-02	1.09E-03	17.548	2.00E-16	***
CL.WORK.INCOME = 3	-2.40E-03	1.53E-03	-1.573	0.115632	
CL.WORK.INCOME = 4	-3.18E-03	1.77E-03	-1.793	0.073035	
CL.WORK.INCOME = 5	5.28E-03	2.58E-03	2.042	0.041138	*
AGE = 2	1.35E-02	2.14E-03	6.303	2.92E-10	***
AGE = 3	1.04E-01	2.08E-03	49.727	2.00E-16	***
AGE = 4	1.44E-01	2.36E-03	60.974	2.00E-16	***
AGE = 5	1.43E-01	2.36E-03	60.613	2.00E-16	***
AGE = 6	1.27E-01	2.29E-03	55.696	2.00E-16	***
AGE = 7	1.24E-01	2.21E-03	55.94	2.00E-16	***
AGE = 8	1.17E-01	2.11E-03	55.549	2.00E-16	***
AGE = 9	1.11E-01	2.06E-03	53.714	2.00E-16	***
AGE = 10	1.01E-01	2.04E-03	49.696	2.00E-16	***
AGE = 11	7.32E-02	2.04E-03	35.932	2.00E-16	***
AGE = 12	4.25E-02	2.14E-03	19.911	2.00E-16	***
AGE = 13	2.83E-02	2.29E-03	12.394	2.00E-16	***
AGE = 14	2.54E-02	2.11E-03	12.039	2.00E-16	***
CL.EQUIV.INCOME = 2	-2.63E-02	1.40E-03	-18.815	2.00E-16	***
CL.EQUIV.INCOME = 3	-3.12E-02	1.24E-03	-25.084	2.00E-16	***
CL.EQUIV.INCOME = 4	-3.55E-02	1.27E-03	-27.936	2.00E-16	***
CL.EQUIV.INCOME = 5	-3.58E-02	1.39E-03	-25.777	2.00E-16	***
CL.EQUIV.INCOME = 6	-3.98E-02	1.15E-02	-3.462	0.000536	***
CL.PENS.INCOME = 2	1.70E-02	2.31E-03	7.363	1.80E-13	***
CL.PENS.INCOME = 3	1.46E-02	2.38E-03	6.152	7.68E-10	***
CL.PENS.INCOME = 4	1.06E-02	2.37E-03	4.499	6.82E-06	***
CL.PENS.INCOME = 5	9.32E-03	2.37E-03	3.935	8.33E-05	***
ED.LEVEL = 2	-5.89E-05	9.23E-04	-0.064	0.94915	
ED.LEVEL = 3	-5.94E-03	1.21E-03	-4.917	8.81E-07	***

**Table 6** Indicators of the goodness of fit of the selected model

Indicator	AIC	BIC	Sigma	Marg. $R^2$	Cond. $R^2$
Employed	-17142.4	-16697.2	0.233	0.773	0.775
Unemployed	-76704.8	-76259.7	0.205	0.132	0.137



**Fig. 4** The relationship between direct and small area estimates - City



**Fig. 5** The relationship between direct and small area estimates - FUA

fit of the model related to the Employed and a poor fit for the one related to the Unemployed

The quality of the small area estimates that have been produced, in terms of correctness and variability with respect to the direct estimates, is assessed. Model-based small area estimates are obtained by employing statistical models that assume certain relationships between the target variable and the available covariates. Comparing these model-based estimates with direct estimates is useful for identifying potential biases inherent in the model-based approach. If the model-based estimates consistently differ from the direct estimates in a particular direction, it indicates the presence of systematic biases that require attention. On the other hand, if the model-based estimates closely align with the direct estimates, it suggests that the model effectively captures the underlying patterns and relationships present in the data.

Figures 4 and 5 illustrate the comparison between the small area estimates obtained using the MIND estimator and the corresponding direct estimates. The analysis reveals that the small area estimates do not exhibit any clear systematic bias when compared to the direct estimates. This finding holds true across for the various parameters of interest, both for Cities and FUAs.



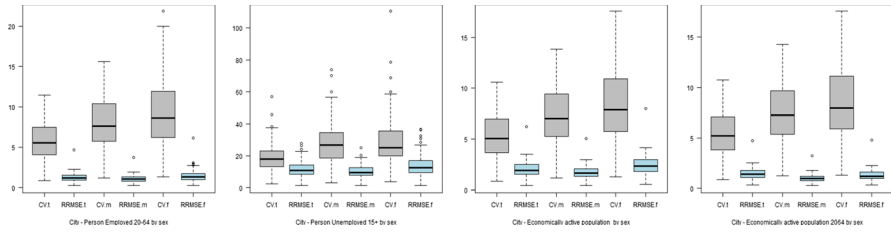


Fig. 6 The distribution of CV% - City

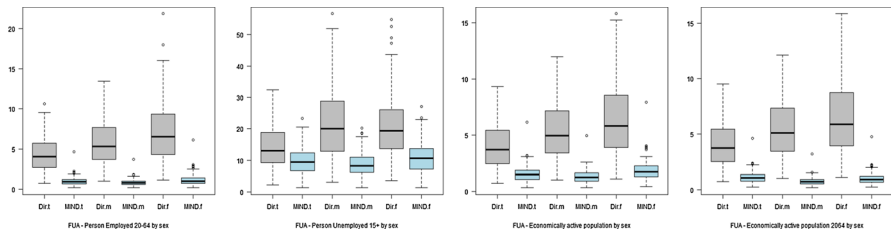


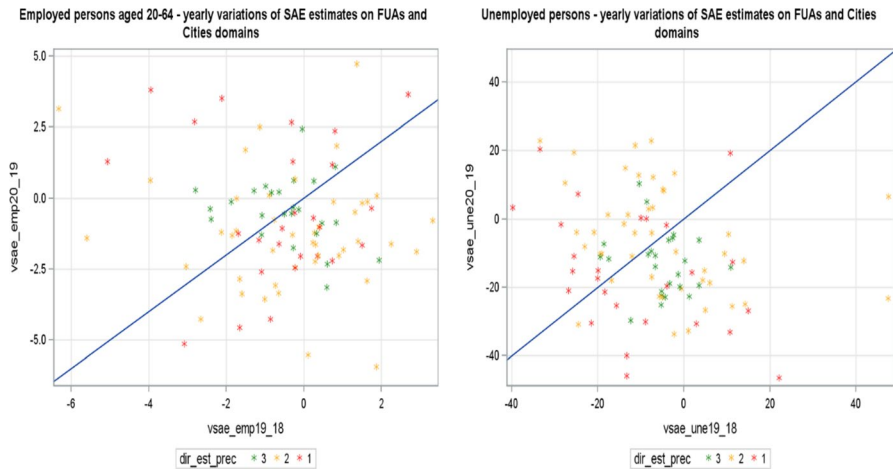
Fig. 7 The distribution of CV% - FUA

Figures 6 and 7 show the distribution of the coefficient of variation for both small area estimates and the corresponding direct estimates at the City and FUA levels. Notably, the coefficients of variation for the direct estimates exhibit high variability, whereas the small area estimation method allows significant gains of efficiency for all indicators within both the City and FUA domains. This highlights the effectiveness of the small area estimation approach in reducing the variability and producing more precise estimates compared to the direct estimates.

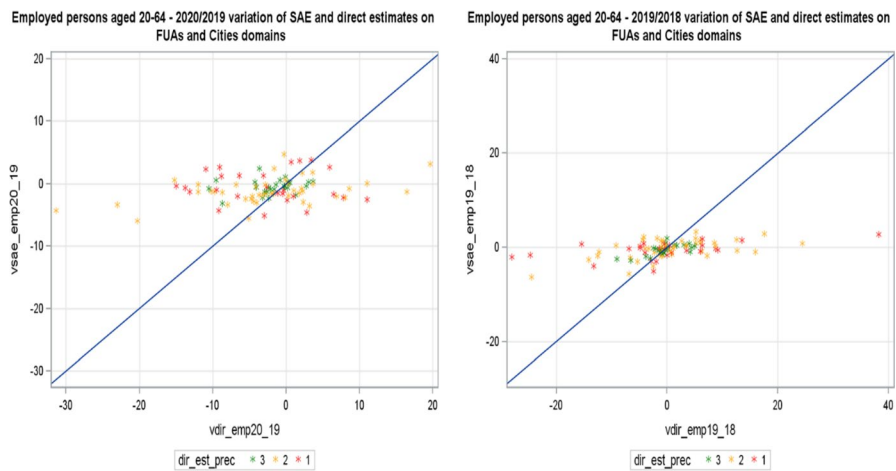
### 5 An analysis of the estimate over the years

Small area estimates have been computed for the years 2018-2020. The focus of this paragraph is analyzing the consistency and patterns of the estimates of yearly changes, to assess the robustness of the SAE methodology and the reliability and stability of the estimates over time.

Figure 8 shows the percentage yearly variations of small area estimates for employed individuals aged 20-64 and unemployed persons, focusing FUAs and the 13 largest cities, the metropolitan municipalities. The scatter plot represents the data, with each dot color-coded to reflect the precision of the corresponding direct estimate. Green dots represent the 25% areas with the lowest coefficient of variation (CV) of the direct estimates (high precision), red dots represent the 25% of areas with the highest CV of the direct estimates (indicating low precision), and yellow dots represent the remaining 50% areas with direct estimates with medium precision. The graphs reveal that the range of yearly variations for the estimates of employed individuals is approximately 8 percentage points for the 2019/2018



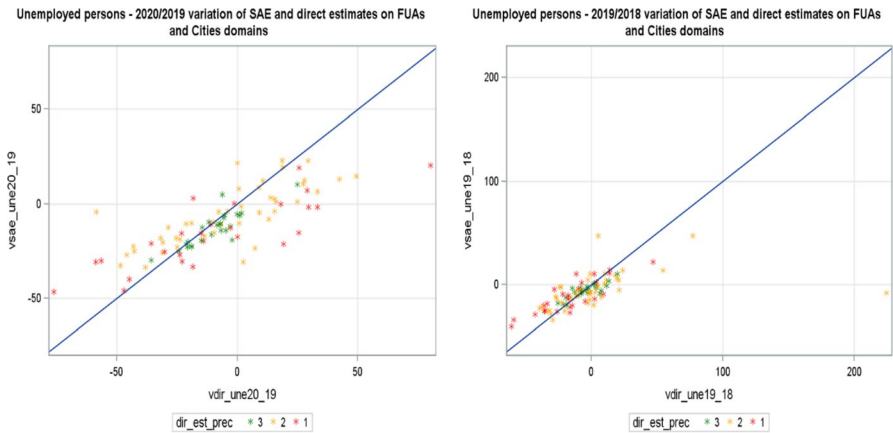
**Fig. 8** Percentage yearly variations 2020 vs 2019 compared with 2019 vs 2018



**Fig. 9** Percentage yearly variations of SAE estimates compared with direct estimates - employed persons aged 20-64 over FUA and 13 greatest Cities (2020 vs 2019 and 2019 vs 2018)

variations and around 10 percentage points for the 2020/2019 variations. The estimates of unemployed individuals instead exhibit a higher range of yearly variations, reaching approximately 60 percentage points. This greater variability is primarily attributed both to the lower level of the parameter and the limited predictability of the auxiliary variables.

Figure 9 shows a comparison between the percentage yearly variations (2020/2019 and 2019/2018) of small area estimates and direct estimates for employed individuals aged 20-64 in FUAs and in the 13 greatest cities. The graphs clearly illustrate that the yearly variations of SAE estimates are generally lower



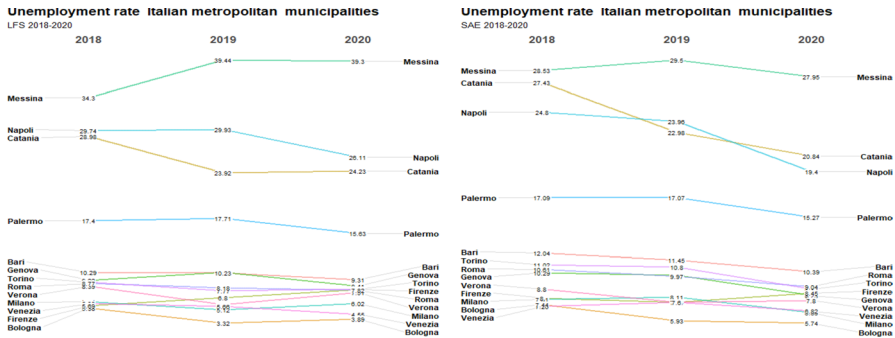
**Fig. 10** Percentage yearly variations of SAE estimates compared with direct estimates - unemployed persons over FUA and 13 greatest Cities (2020 vs 2019 and 2019 vs 2018)

when compared to the corresponding variations of direct estimates. This pattern is particularly evident in areas where the precision of direct estimates is lower, those identified by yellow and red dots.

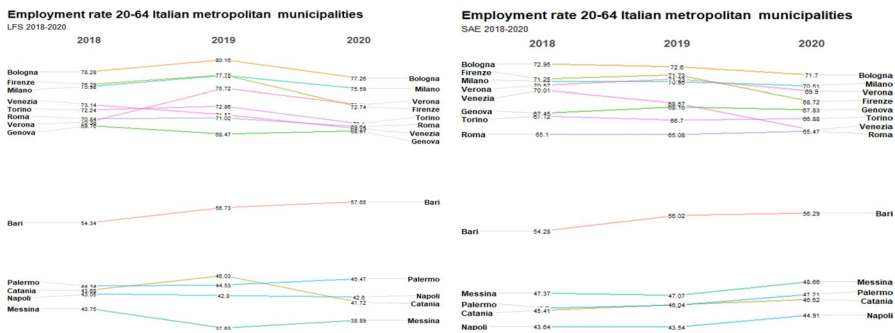
The use of SAE models, which incorporate a wide range of relevant covariates highly correlated with employment status, significantly enhances the precision of the estimates compared to direct estimates, reducing variability and uncertainty. This improvement in precision can also contribute to reduce yearly variations, even if that is usually due to the fact that the SAE estimates are more shrinkage than the direct ones.

Figure 10 illustrates the comparison of yearly variations for unemployed persons of small area estimates with respect to direct ones. The graphs demonstrate that in this case the yearly variations of SAE estimates closely align with those of direct estimates, although some outliers present in the latter are corrected by the former. This is due to the fact that unlike the employment-related covariates used in the SAE models, the covariates associated with unemployment are not as strongly correlated with the unemployment indicator. As a result, the SAE estimates for unemployment tend to align closely with the direct estimates, particularly when the direct estimates exhibit higher precision. This indicates that the SAE methodology, while offering some improvements, is less influential in mitigating the variability and correcting outliers for unemployment estimates compared to employment estimates, showing the importance of having correlated auxiliary information.

Nevertheless, the SAE models still provide valuable insights by refining the estimates and addressing certain outliers in the direct estimates. That can be observed looking at Figs. 11 and 12 that focus on the largest municipalities with the largest sample size. Figure 11 provides a comparison between the trend of unemployment rates calculated using the direct estimator and the corresponding trend of small area estimates (SAE). The graph reveals that the SAE estimates exhibit a smoother trend, with reduced volatility over the three-year period. This increased stability aids in better



**Fig. 11** Temporal variation of direct and SAE estimates of unemployment rate for metropolitan municipalities)



**Fig. 12** Temporal variation of direct and SAE estimates of employment rate for metropolitan municipalities)

understanding and analyzing the changes in unemployment rates, allowing for more informed decision-making and policy interventions. Similar conclusions can be drawn when examining Fig. 12, which presents the same results for employment rates in 20–64 age group. These findings support the value of utilizing the SAE methodology in obtaining reliable and robust estimates for the variables of interest.

In conclusion, the small area estimates not only exhibit a reduction in coefficient of variations (CVs), but they also provide more stable estimates year by year, thereby reducing the volatility of the estimated time series. The SAE methodology proves effective in producing estimates that display greater consistency and smoothness over time, mitigating the fluctuations that are often present in direct estimates computed over unplanned domains.

## 6 Conclusions

This paper presents the methodology developed by Istat for estimating a selection of labor market indicators at the City and FUA level. In Italy, these indicators, based on the LFS (Labor Force Survey) data, are typically released at the regional and provincial level, but not for the specific domains of interest defined by the OECD-EC classification. The estimation covers the years 2018, 2019, and 2020.

The primary objective was to develop a proper small area method for estimating the selected indicators for the domains of interest, aiming to enhance the efficiency compared with direct estimates. Given the large number of target indicators involved, and the relationships among them, a challenging aspect was the production of coherent and consistent estimates. The chosen estimator is based on a multivariate mixed-effects linear model implemented through the Mind R package. To estimate the labor market indicators at the City and FUA level, the multivariate approach developed in MIND was exploited in two ways: to define the dependent variable and to specify the random effects. In this case, the dependent variable was defined as a vector consisting of three dichotomous variables representing the employed, unemployed and inactive individuals. The coherence of indicators across different domains was achieved through a single cross-classification model encompassing, within the graphical areas, all the domains of interest.

The choice of a unit-level model estimator was motivated by the objective of efficiently harnessing the auxiliary information available from the new integrated system of statistical registers. The Population and Labor Register served as the primary sources of information for individuals' demographic characteristics and employment conditions. Additionally, the dataset was supplemented with data on social aspects, welfare benefits, and income types from the Italian Ministry of Finance and the Social Security Agency.

The selection of estimation model for the City and FUA level indicators was carried out separately for the employed and unemployed populations. After testing an initial model with all available covariates, the predictors highly associated with the response variable were selected to define the fixed component in the mixed-effects linear model.

The estimates obtained using the chosen estimator were compared to the direct estimates computed in the initial phase of the study. The results demonstrate significant efficiency gains, particularly for estimating the total number of unemployed persons and their breakdown by sex. This improvement is particularly significant as the direct estimates faced high sampling errors. These estimates are currently available online as part of Eurostat's Cities database, allowing users to download them by selecting Italian Cities and FUAs in the Labor Market section.

Future developments will focus on incorporating time series data into the estimation process, provided that a sufficient number of survey occasions are available. This will entail introducing a temporal random effect into the mixed model framework. Furthermore, comprehensive comparisons are planned between small area estimation models based on logistic/multinomial mixed models and those

utilizing area-level models. These comparative analyses will explore alternative modeling approaches, aiming to identify the most effective strategies for improving the estimation process.

Moreover, a more extensive quality assessment of the estimated results will be undertaken to ensure their reliability and accuracy. Overall, these planned advancements aim to enhance the robustness and applicability of the statistical methodology employed in estimating small area labor market variables. By incorporating different modeling approaches and conducting thorough quality assessments, researchers can refine and validate the estimation process, leading to more reliable and insightful results.

## References

- Baldi C, Ceccarelli C, Gigante S, Pacini S, Rossetti F (2018) The labour register in Italy: the new heart of the system of labour statistics. *Rivista Italiana di Economia, Demografia e Statistica*, LXXI I(2):95–105
- Battese GE, Harter RM, Fuller WA (1988) An error-components model for prediction of county crop areas using survey and satellite data. *J Am Stat Assoc* 83:28–36. <https://doi.org/10.1080/01621459.1988.10478561>
- Chambers RL, Clark RG (2012) *An introduction to model-based survey sampling with applications*. Oxford University Press, Oxford
- Chambers R, Salvati N, Tzavidis N (2016) Semiparametric small area estimation for binary outcomes with application to unemployment estimation for local authorities in the UK. *J R Stat Soc A Stat Soc* 179:453–479
- D'Alò M, Di Consiglio L, Falorsi S, Ranalli MG, Solari F (2012) Use of spatial information in small area models for unemployment rate estimation at sub-provincial areas in Italy. *J Indian soc Agric Stat* 66: 43–53
- D'Alò M, Falorsi S, Fasulo A (2021) mind R package, <https://cran.r-project.org/web/packages/mind/index.html>
- Datta GS, Day B, Basawa I (1999) Empirical best linear unbiased and empirical Bayes prediction in multivariate small area estimation. *J Stat Plann Inference* 75(2):269–279
- Dawber J, Salvati N, Fabrizi E, Tzavidis N (2022) Expectile regression for multi-category outcomes with application to small area estimation of labour force participation. *J R Stat Soc Ser A Stat Soc* 185(2):590–619
- Deville JC, Sarndal CE (1992) Calibration estimators in survey sampling. *J Am Stat Assoc* 87:376–382
- EUROSTAT (2017) Methodological manual on city statistics, <https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/ks-gq-17-006>
- EUROSTAT (2018) Methodological manual on typologies, <https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/ks-gq-18-008>
- Fellner WH (1987) Spare matrices and the estimation of variance components by likelihood. *Commun Stat Simul* 16:439–463
- Harville DA (1977) Maximum likelihood approaches to variance component estimation and to related problems. *J Am Stat Assoc* 72:320–340
- ISTAT (2006) *La rilevazione sulle forze di lavoro: contenuti, metodologie, organizzazione, Metodi e norme*, 32
- ISTAT (2016) Istat's Modernisation Programme, [https://www.istat.it/it/files//2011/04/IstatsModernisationProgramme\\_EN.pdf](https://www.istat.it/it/files//2011/04/IstatsModernisationProgramme_EN.pdf)
- López-Vizcaíno E, Lombardía MJ, Morales D (2013) Multinomial-based small area estimation of labour force indicators. *Stat Model* 13(2):153–178
- López-Vizcaíno E, Lombardía MJ and Morales D (2015) Small area estimation of labour force indicators under a multinomial model with correlated time and area effects. *J R Stat Soc Ser A (Stat Soc)* pp. 535–565

- Marino MF, Ranalli MG, Salvati N, Alfò M (2019) Semiparametric empirical best prediction for small area estimation of unemployment indicators. *Ann Appl Stat* 13(2):1166–1197
- Molina I, Saei A, Lombardía MJ (2007) Small area estimates of labour force participation under a multinomial logit mixed model. *J R Stat Soc Ser A Stat Soc* 170(4):975–1000
- Ranalli MG, Montanari GE, Vicarelli C (2018) Estimation of small area counts with the benchmarking property. *Metron* 76:349–378
- Rao JNK, Molina I (2015) *Small area estimation*. Wiley, London
- Saei A, Chambers R (2003) *Small area estimation under linear and generalized linear mixed models with time and area effects*. S3RI Methodology working paper M03/15

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.