D3.4 Small area estimates pilot study - final report

WP3 - Pilot study “Small Area Estimation for city and functional urban area statistics”

Grant Agreement “Sub-national statistics Italy” n. 882021

30th of March 2022

D’Alò Michele, Filipponi Danila, Loriga Silvia (ISTAT)

Index

1. Introduction

The aim of this report is to describe the statistical methodology used to produce estimates of a selection of labour market variables at City and Fua level and to analyse the results obtained. The variables are included in Eurostat's City database[1] and their release is envisaged in the Grant Agreement "Sub-national statistics Italy" between Istat and Eurostat. The calculation of these estimates is the specific objective of the WP3 - Pilot study "Small Area Estimation for city and functional urban area statistics".

The parameters of interest are the following: number of unemployed persons, total and by sex; number of economically active persons, total and by sex; number of economically active persons aged 20 to 64, total and by sex; number of employed persons aged 20 to 64, total and by sex. All the estimates refer to the years 2018, 2019, 2020.

The estimates of these indicators are calculated through a unit level multivariate model that has been designed in order to allow internal coherence among all the target parameter's estimates. The estimates are based on Labour Force Survey (LFS) data and relevant covariates are taken from the Labour Register and Population Register.

Concerning the report's structure, the definition of the variables and the domains of interest is presented in paragraph 2. In this paragraph, the sampling variance of the direct estimates is analysed: the presence of areas of interest not covered by the sample or having low sampling fraction and the low frequency of observation units, in particular for the estimate of the unemployed persons, imposes the use of small area estimation methods. Finally, the auxiliary variables derived from the Labour Register and the Population Register, together with others administrative information that can be used in the estimation process are described. The method, that can be seen as an extended multivariate version of the more standard linear mixed model at unit level, is explained in paragraph 3. A particular focus is devoted to the estimator used, a function named MIND, available in the R package MIND - Multivariate model based INference for Domains (https://cran.rproject.org/web/packages/mind/index.html). The analysis of the results is exposed in paragraph 4: some model diagnostics are reported in paragraph 4.1; in paragraph 4.2, a comparison between small area estimates and direct estimates helps highlighting the gain in efficiency that can be obtained while applying the methods for small area estimates that were defined. Finally, in paragraph 5 some analysis on the main results of small area estimates produced at City and Fua level for years 2018-2020 are shown, comparing them with the direct estimates and the main conclusion are reported in the paragraph 6.

2. Variables' definition and direct estimates

The variables of interest are represented by the main aggregates of labour supply, measuring the participation of individuals to the labour market at City and Fua level (total and by sex). The indicators concerned are twelve:

[1] https://ec.europa.eu/eurostat/web/cities/data/database

- Economically Active Population (EAP), total, male and female
- Economically Active Population, 20-64 (EAP_20-64), total, male and female
- Persons Unemployed (UNE), total, male and female
- Persons Employed, 20-64 (EMP_20-64), total, male and female

The first two indicators refer to the labour force, given by the sum of employed and unemployed persons. Where the age group is not explicit, it refers to those aged 15 and over. The indicators of interest are strongly related and the relation between indicators is the following, and it holds for sex:

$$EAP\_20\text{-}64 < EMP\_20\text{-}64 + UNE < EAP$$

In addition, each total indicator coincides with the sum of the corresponding male and female indicator.
Given this relationship, in order to preserve the coherence among the estimates of the different indicators it is necessary to apply a multivariate approach.

The areas of interest are typically "small areas": they are represented by Cities that consist of a single municipality and Fuas that are formed by groups of municipalities (a City plus the municipalities included in its commuting zone). Fuas are based on the OECD-EC city definition and they represent territories that are highly integrated from an economic point of view. Consequently, Fuas (especially larger ones) often intersect the administrative boundaries of provinces and, in some cases, even of regions. In two cases only, the domain of interest is given by the Greater City, that is formed by the group of municipalities sharing the same high density cluster (*core city*). In Italy this applies only in Milan and Naples. For an insight on the definition of these territories, please refer to Eurostat's *Methodological manual of territorial* typologies[2].

Cities, Greater Cities and Fuas do not represent planned domains of the Labour force survey. The survey's planned domains are provinces and regions, for which direct annual and quarterly estimates are released. The sample design envisages a stratification of municipalities at provincial level and the calibration used for producing the direct estimates of the survey is carried out with respect to the total population of provinces and regions. Beyond provinces and regions, the calibration weights allow to produce the direct estimates for the 13 greatest municipalities (having a population greater than 250 thousand persons).

As above mentioned, the domains of interest show intersections with administrative units. In Italy there are 83 Fuas in total; they are present in 19 regions (only in Valle d'Aosta there is no Fua) and in 83 provinces. The intersection among FUAs and provinces generates 116 areas:

- there are in total 21 FUAs intersecting different provinces;
- the FUA of Milan intersects 9 provinces;
- the FUA of Rome intersects 6 provinces;
- the FUA of Naples intersects 3 provinces;
- the other 18 FUAs intersect 2 provinces.

[2] https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Territorial_typologies_manual

In total, there are 110 areas resulting from the intersection of FUAs and regional boundaries (90 if we exclude the areas of regions that do not present FUAs). The FUAs that intersect different regions are 7 in total (each intersecting 2 regions):

- IT001L3 Roma – intersects Lazio (12) and Abruzzo (13)
- IT006L3 Genova – intersects Liguria (7) and Piemonte (1)
- IT013L3 Cremona – intersects Lombardia (3) and Emilia-Romagna (8)
- IT033L2 Piacenza – intersects Emilia-Romagna (8) and Lombardia (3)
- IT036L2 La Spezia – intersects Liguria (7) and Toscana (9)
- IT507L2 Ferrara – intersects Emilia-Romagna (8) and Veneto (5)
- IT515L2 Terni – intersects Umbria (10) and Lazio (12).

In the case of these 7 FUAs that intersect 2 regions, there is always a prevalent part in one region and a residual part in another. In the residual region fall max 3 municipalities belonging to a Fua, corresponding to a max 7% of the total population.

The cities are 87 in total (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:

- 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.

Therefore, the domains of interest (Cities and Fuas) are partly overlapping and, since they are defined independently from the provincial and regional administrative boundaries, they intersect them.

Since they are not planned domains, the survey sample coverage in the areas of interest varies. On the basis of the analysis of the overlap between the LFS sample (year 2020) and Fuas and Cities, it is possible to observe that:

- One FUA and its City (Trani) are not covered by the LFS sample (as there is no sample observation in the area);
- While the sample fraction at national level is slightly lower than 1%, some areas of interest present smaller sample fractions; in particular, this occurs for the FUAs of Brescia, Venezia, Treviso, Verona and Varese, showing a sample fraction that does not exceed the half value of the national average;
- Some areas of interest present sample fractions that are higher than the national average. Overall, there are 16 areas with a sample fraction that is one and a half times higher than the national average;
- In some areas of interest, the direct weights produce a population estimate deviating of more than 20% from the values based on administrative data;
- The highest population underestimates are registered in relation to the Fuas of Bergamo (-62%), Treviso (-59%), Ancona (-33%);
- The highest population overestimates are registered in relation to the Fuas of Sassuolo (+139%), Vicenza (+53%), Trapani (+49%), Caserta (+47%), Pordenone (+43%), Alessandria (+33%).

The direct estimates of the twelve indicators mentioned earlier were produced for the areas of interest; their degree of accuracy has been evaluated through their coefficient of variation. These estimates represent the first step towards the production of estimates with small area models. The direct estimates are produced applying the same estimator used for producing the survey's current estimates, namely a calibration estimator for which the constraints are represented essentially by the population distribution by age and sex at different territorial scales (provinces, regions).

The coefficient of variation strongly varies among the areas and the indicators of interest. Since the objective is to deliver estimates for both Fuas and Cities (as well as for the Greater Cities of Milan and Naples), the coefficient of variation in the areas generated by the intersection between Fuas and their Cities/Greater Cities are reported below. To simplify, we consider only two indicators: the employed persons aged 20 to 64 (EMP_20-64) and the unemployed persons (UNE), referring to the total (male + female).

With reference to employment, the coefficients of variation of the direct estimates are all "acceptable": the highest is equal to 33% and is associated to the Fua of Novara; apart of it the highest is equal to 16% in the Fua of La Spezia. The estimates of unemployment, as they are lower than those of employment, present instead higher coefficients of variation: even if the anomalous data of the Fua of Ferrara is deleted (about 113%), the coefficients of variation exceed 33% in 30 areas, while they range between 16% and 33% in 73 areas and in 65 areas they are inferior to 16%. It is worth noting, that coefficients of variation of the direct estimates of unemployment were lower in previous years because the level of the estimates was higher compared with 2020 (in 2020 it was affected by the Covid pandemic). In table 2.1 the statistics for the twelve indicators of interest are presented.

Figure 2.1 Distribution of the direct estimates CVs in the estimation domains (Cities and Fuas – LFS 2020)

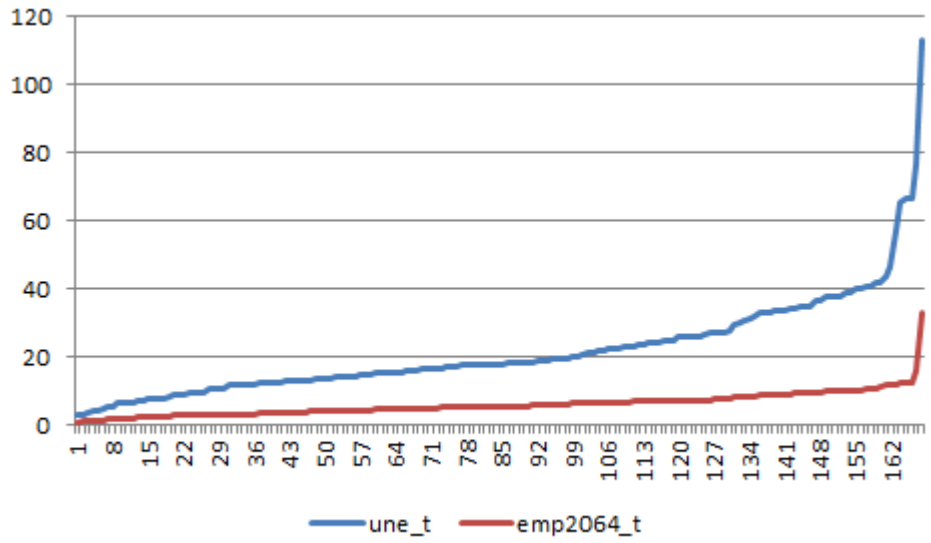


Table 2.1 CV percentiles of the direct estimates in the estimation domains (Cities and Fuas – LFS 2020)

	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

All the considerations exposed so far regarding the features of the domains of interest, the variability of the sample coverage in the areas and the distribution of the coefficient of variation calculated for the direct estimates of the twelve indicators, lead towards the necessity of adopting small area estimation methods for achieving the estimation goal of the present work.

2.1 Administrative information

The information available in the statistical registers and administrative data are essential for the production of statistics at small area level. In particular, in order to estimate labour market indicators, variables derived from the Labour Register and from the Population Register, together with others administrative information, were used.
The Labour Register is a statistical register that includes all Italian jobs and social security information built integrating social security and fiscal data with the aim of both increasing the consistency of statistical processes on the employment of enterprises, labor market indicators and national accounts and aligning information on employment, pension and income.
The Population Register is a statistical register of residing persons in Italy. The register is built integrating administrative population registers, held by the Italian municipalities, with two statistical surveys conducted annually to correct the administrative data for under and over coverage errors. The Population Register provides population counts together with some demographic characteristics of individuals.
Among all the available variables, those used as covariates focus on the following domains:

- demography
- employment
- income
- social aspects.

The demographic variables such as gender, age, place of residence and citizenship were obtained from the Population Register.

The main occupational characteristics, on the other hand, were obtained from the Labour Register. Starting from the job positions, which are the statistical units of the register, the variables related to the occupational and non-occupational status, were obtained trying to be as consistent as possible with the definitions of the labour force survey. The variables derived from the labour register are:

- *weekly employment condition, for the 52 weeks of the reference year;*
- *Information on events of job-protected leave for which the worker receives an allowance, with specific focus on events connected to the redundancy fund;*
- *Information on the termination of the employment relationship.*

The availability of fiscal sources from the Italian Ministry of Finance allowed the computation of:

- *work income*
- *retirement income*
- *capital income*

For the resident individuals observed in the population register and their family, the individual *equivalent income* indicator has been computed. The equivalent income consists of the family income divided by the equivalent dimension of the family - while taking into consideration the presence of scale economies affecting the consumption needs of the family. The equivalent dimension of the family is obtained applying a specific equivalence scale modified by the OECD: at operational level, the equivalent dimension of the family is calculated attributing 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.

Finally, additional information on social aspects was collected from social security data and the Population Register. The variables are:

- *education level and enrollment in school or university courses.*
- *retirement pension*
- *invalidity pension*
- *other types of financial support (unemployment benefits, family allowances for workers, transfers to families with economic problems, sickness and maternity allowances, subsidies for students).*

3. Small areas estimator

The estimator used is based on a multivariate model implemented through the R MIND function, developed by Istat. The method may be seen as an extended multivariate version of the standard linear mixed model at unit level. The main extensions give the possibility to specify a multivariate linear mixed model following the approach described in Datta, Day e Basawa (1999) and can allow a model specification with more than one random effect, so that, possible marginal effects can be fitted in the model. These possible marginal random effects, in addition or instead to the usual random area effects, can be especially useful if a relevant number of areas are very small or even out of sample. This might take the bias of the synthetic estimator under control allowing more local smoothed and less shrinkage synthetic estimates, especially for the out of sample areas. The marginal random effect may be

derived from the variables used to define the strata or from some other variables utilized for defining the planned domains or for cross-classify the population units.
The R function applied, named MIND - Multivariate model based INference for Domains, is available on the correspondent R package available on Cran (see D'Alò et al 2021). The model can be expressed as follows:

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

where

- **y** and **e** are respectively the vector of the sample values of the target variables and of the residuals, of (n×C) elements, where n is the number of units observed in the sample, while C represents the number of values assumed by the target variable;

- $\mathbf{X}='X \otimes I_C$, where $'X$ is the design matrix of the sample values of the auxiliary variables considered for the fixed effects whose dimension is (n×G), with G being the number of variables (and/or the categories in case of categorical values) considered in the model and I_C is an identity matrix of C order. Naturally the order of matrix **X** of the multivariate model is [(n×C)×(G×C)] ;

- β is the vector of the regression parameters whose length is (G×C);

- $\mathbf{Z}='Z \otimes I_C$, where $'Z$ is the design matrix of the sample values of the random effects of dimension (n×Q), where Q is the total number of modalities of the random effects considered in the model. Naturally the **Z** matrix of the multivariate model is of [(n×C)×(Q×C)] order;

- **u** is the vector of random effects whose length is (Q×C)

The estimates that can be obtained with MIND fall within the group of estimators classified as Projection. The general formulation of the Projection estimator of the vector of totals for domain d (d=1,…,D), $Y_{d,}$ is given by the sum of predicted vector values on the basis of the above mentioned model, $\tilde{y}_{d,k}$ for all the units of the target population falling within the d domain U_d:

$$\hat{y}_d^{(PR)} = \sum_{k \in U_d} \tilde{y}_{d,k} \qquad d = 1, \dots, D$$

The composite estimator is an EBLUP estimator where predicted values are used only for the subgroup of units that are not included in the sample; for the s_d sample units falling within the d domain , instead, the values $y_{d,k}$ that are directly observed with the survey are used. The EBLUP estimator of the total is the following:

$$\hat{y}_d^{(EBLUP)} = \sum_{k \in s_d} y_{d,k} + \sum_{k \in U_d - s_d} \tilde{y}_{d,k} \qquad d = 1, \dots, D$$

The estimate of variance components of the mixed effects model is computed with the REML derived by the algorithm proposed by Saei e Chambers (2003), using the approach developed by Fellner (1986, 1987).

In order to produce the labour market indicators at City and Fua level, the potential of the multivariate approach proposed by MIND has been exploited in two ways: to specify the dependent variable y and to define the random effects.

Since the variables of interest are the employed, the unemployed and the economically active population (that is the sum of employed and unemployed persons), the vector $\boldsymbol{y}$ of the dependent variable can be defined as a vector composed by three dichotomous variables representing the categories of the employed, unemployed and inactive. It is also important to recall that these three groups represent an exhaustive and mutually exclusive classification of the population. Obviously, the joint modelling of the three labour market categories allows the coherence with the population of the domain of interest. The territorial domains of interest for each variable are given by City and Fua. Within each territorial domain, the units are further specified according to sex and age group (the indicators for the employed refer to the 20-64 age group, the indicators for the unemployed to those aged 15 and over, while the indicators for the economically active population refer to both age groups).

In order to guarantee the coherence of indicators across different domains, it is possible to define a single cross-classification model that includes all the domains of interest. The model for the vector $y_{dj,k}$ associated to the k (k=1,…,N_{dv}) individual in the domain *(d,v), d=1,…,D e v=1,…,V,* can be expressed as a special case of model (1):

$$y_{dj,k} = X_{djk}\beta + \tau_d + \delta_j + \gamma_{dj} + e_{djk}$$

where $\tau, \delta\ e\ \gamma$ are the random effects. Here, we considered a model where the random effects $\delta\ e\ \gamma$ are degenerate at zero.

4. Analysis of the results

The main goal of the section is to present the analysis of the obtained results. The goodness of fit of the applied model is described in the first paragraph, followed by the analysis of the estimates produced with the selected model and by the comparison of these estimates with the direct estimates. All the results refer to the year 2018, the first year in which the estimation model has been selected, but similar results have been obtained for the years 2019 e 2020. In the last paragraph an analysis of the results over the three years is carried out, comparing SAE and direct estimates.

4.1 Analysis for the model selection

The LFS sample data have been integrated with the selected register information and a model selection has been carried out considering separately the employed and unemployed variables. A mixed linear model with specific area random effects, at City and FUA level respectively, has been considered. Tables 4.1 and 4.2 show that for

the four models the inter-class correlation coefficient (ICC), a measure of the proportion of variance across areas with respect to the total variance, is quite low.

Table 4.1 – Proportion of variance between Cities and individuals - LFS 2018

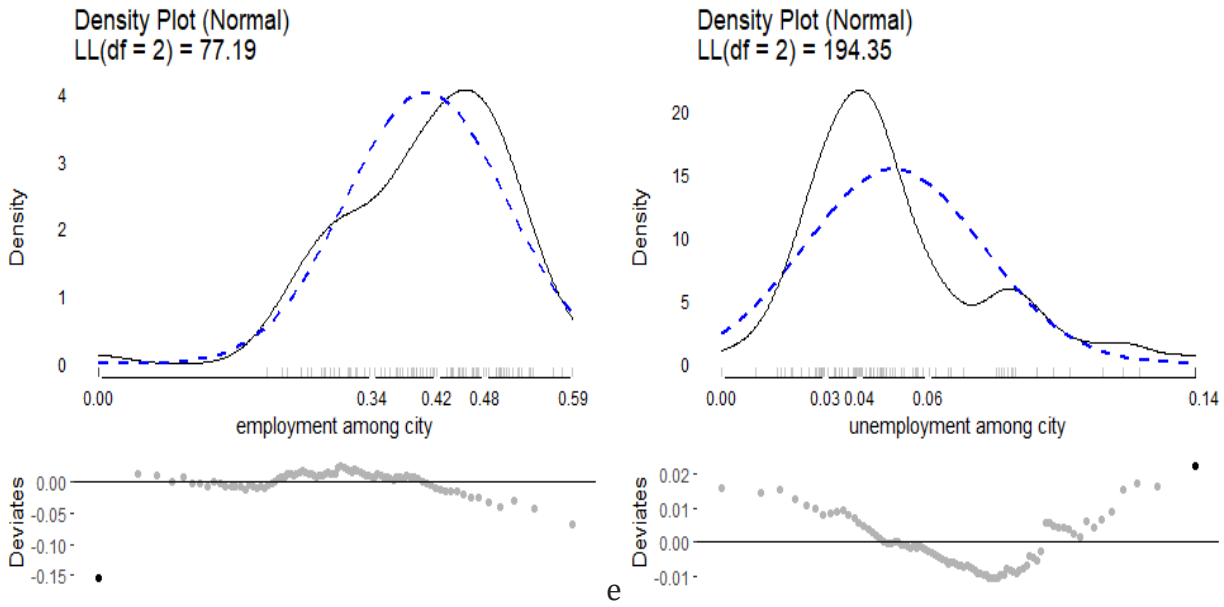
	Persons employed, 20-64, total		Persons unemployed, total	
Variance components	Sigma	ICC	Sigma	ICC
City	0.0078373	0.0323084	0.0006054	0.0117914
Residual	0.2347415	0.9676916	0.0507329	0.9882086

Table 4.2 – Proportion of variance between Fuas and individuals – LFS 2018

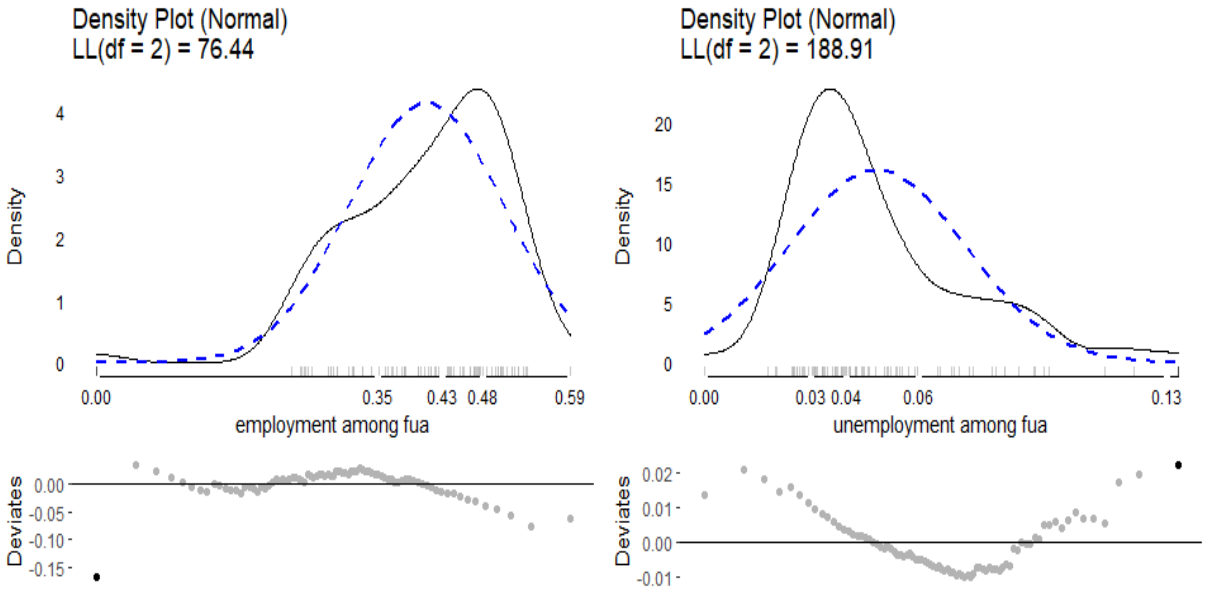
	Persons employed, 20-64, total		Persons unemployed, total	
Variance components	Sigma	ICC	Sigma	ICC
FUA	0. 0071719	0. 0294422	0.0005682	0.0116885
Residual	0. 2364211	0. 9705578	0.0480401	0.9883115

Anyway as showed in Figure 4.1 not negligible differences of employment and unemployment levels among areas, both for City and FUA, are present. In the case of employed persons, the distribution is right skewed, with a long left tail, in both the City and Fua domains. The unemployed distribution instead is left skewed and leptokurtic. The marginal random effect at regional level seems to be not significant.

Figure 4.1 –Distribution of the employed and unemployed persons across domains (City and FUA - LFS 2018)



e



The aim of the next step is to assess the relationship between the variables of interest and the group of predictors available, in order to select the auxiliary information that are highly associated with the response variables and as consequence to specify the fixed part of the linear mixed model on which the estimator used is based.

In tables 4.3 and 4.4, it is possible to have an overview of the regression coefficients and the standard errors of the models selected with a stepwise method for the employed and the unemployed, with respect to the Cities, using 2018 sample data. Of course, a similar model selection has been carried out also for FUAs; and with respect to both areas using data collected in 2019 and 2020. The auxiliary variables in the tables correspond respectively to:

- ✓ the dichotomous variable sex that takes values 1 if the unit is female and 0 otherwise;
- ✓ the indicator variable that takes values 1 for the class of age to which each unit belongs to;
- ✓ the dichotomous variable student that takes values 1 if the unit is student and 0 otherwise;
- ✓ the dichotomous variable unemployment benefit that takes values 1 if the unit perceives it and 0 otherwise;
- ✓ the dichotomous variable layoff benefit that takes values 1 if the unit perceives it and 0 otherwise;
- ✓ the variable indicating the number of working months in the year for each unit;
- ✓ the dichotomous variable retirement that takes values 1 if the unit is retired and 0 otherwise;
- ✓ the indicator variables that take values 1 for the class of working wage to which each unit belongs to;
- ✓ the indicator variables that take values 1 for the class of equivalent income to which each unit belongs to

Table 4.3 Significance of the auxiliary variables in the linear mixed-effects model – Employed persons in Cities

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	0.062	0.007	3244.173	9.108	0.000
Sex 1	-0.027	0.001	144831.354	-20.829	0.000
Age 16-19	0.024	0.006	144805.417	3.867	0.000
Age 20-24	0.091	0.007	144809.987	13.621	0.000
Age 25-29	0.129	0.007	144814.076	19.210	0.000
Age 30-34	0.139	0.007	144816.452	20.922	0.000
Age 35-39	0.139	0.007	144816.548	21.274	0.000
Age 40-44	0.137	0.007	144815.238	21.039	0.000
Age 45-49	0.130	0.006	144813.498	20.065	0.000
Age 50-54	0.119	0.007	144817.255	18.356	0.000
Age 55-59	0.084	0.007	144821.082	12.829	0.000
Age 60-64	0.054	0.007	144827.911	7.983	0.000
Age 65-69	0.056	0.007	144829.680	8.280	0.000
Age 70+	0.053	0.007	144831.050	7.953	0.000
Student	-0.027	0.006	144811.932	-4.316	0.000
Unemployment benefit	-0.061	0.003	144843.452	-17.799	0.000
Layoff benefit	-0.100	0.010	144845.624	-10.206	0.000
Working months in the year	0.657	0.002	144858.415	287.512	0.000
Retirement	-0.112	0.003	144868.958	-40.486	0.000
Class of working wage 2	0.068	0.002	144867.947	35.773	0.000
Class of working wage 3	0.176	0.003	144829.741	67.404	0.000
Class of working wage 4	0.193	0.003	144842.233	63.568	0.000
Class of working wage 5	0.210	0.004	144824.575	49.913	0.000
Class of equivalent income 1	-0.028	0.002	144866.840	-13.113	0.000
Class of equivalent income 2	-0.026	0.002	144741.804	-11.793	0.000
Class of equivalent income 3	-0.027	0.002	144341.055	-11.618	0.000
Class of equivalent income 4	-0.034	0.002	143971.405	-14.395	0.000
Class of equivalent income 5	0.016	0.023	144822.801	0.673	0.501

Table 4.4 Significance of the auxiliary variables in the linear mixed-effects model – Unemployed persons in Cities

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	0.155	0.006	6930.01	26.087	0.000
Sex 1	-0.026	0.001	144850.39	-22.325	0.000
Age 16-19	0.056	0.006	144818.94	9.722	0.000
Age 20-24	0.096	0.006	144824.45	15.875	0.000
Age 25-29	0.087	0.006	144829.85	14.310	0.000
Age 30-34	0.070	0.006	144833.16	11.612	0.000
Age 35-39	0.064	0.006	144833.35	10.781	0.000
Age 40-44	0.058	0.006	144831.48	9.881	0.000
Age 45-49	0.049	0.006	144829.66	8.293	0.000
Age 50-54	0.034	0.006	144834.25	5.723	0.000
Age 55-59	-0.003	0.006	144838.77	-0.590	0.555
Age 60-64	-0.036	0.006	144846.15	-5.974	0.000
Age 65-69	-0.048	0.006	144848.27	-7.859	0.000
Age 70+	-0.050	0.006	144849.74	-8.343	0.000

	Estimate	Std. Error	df	t value	Pr(>|t|)
Student	-0.080	0.006	144827.54	-14.116	0.000
Unemployment benefit	0.103	0.003	144861.25	32.976	0.000
Layoff benefit	0.090	0.009	144861.92	10.141	0.000
Working months in the year	-0.135	0.002	144773.45	-65.198	0.000
Retirement	-0.040	0.003	144848.31	-15.835	0.000
Class of working wage 2	0.027	0.002	144859.01	15.782	0.000
Class of working wage 3	0.007	0.002	144848.05	3.081	0.002
Class of working wage 4	0.008	0.003	144860.25	2.801	0.005
Class of working wage 5	0.015	0.004	144841.94	3.871	0.000
Class of equivalent income 1	-0.041	0.002	144862.92	-20.810	0.000
Class of equivalent income 2	-0.055	0.002	144398.91	-27.314	0.000
Class of equivalent income 3	-0.064	0.002	143361.70	-30.964	0.000
Class of equivalent income 4	-0.064	0.002	142452.45	-29.703	0.000
Class of equivalent income 5	-0.034	0.021	144841.01	-1.611	0.107

The following table 4.5 allows to evaluate the goodness of fit of the initial complete model specified for the employment and unemployment variables, at Fua and City level. Table 4.6 instead reports the corresponding evaluation for the model actually used for the estimation, identified after a stepwise regression selection of the independent variables involved in the fixed part of the final mixed model. The data displayed across columns regards: the sample dimensions (**N_Obs**); the information criterion of Akaike (AIC), that is used for assessing and comparing statistical methods; the Bayesian information criterion (**BIC**); the log-likelihood (**LL**) of the model; the degrees of freedom of the model (**DF**); the residuals' standard deviation (**Sigma**) and the random effects (**Sigma_u**); the marginal coefficient of determination $R^{2,}$ that refers exclusively to the model's fixed component (**MarginalR2**); the conditional coefficient of determination (**ConditionalR2**), that takes into account the significance of the entire mixed-effect model, including the random domain effect. As expected, the marginal R^2 is very high (77%) for the employed persons estimate, thanks to the presence of variables that are strongly associated to the employment status such as administrative employment and work income. The value of the marginal R^2 is instead lower (14%) for the estimate of unemployed persons, as there are no variables strongly associated to the phenomenon. The random domain effect leads to a small increase of the R2 index (conditional R2). The selection of auxiliary variables leads towards a not significant improvement of the marginal R2 and of the conditional R2.

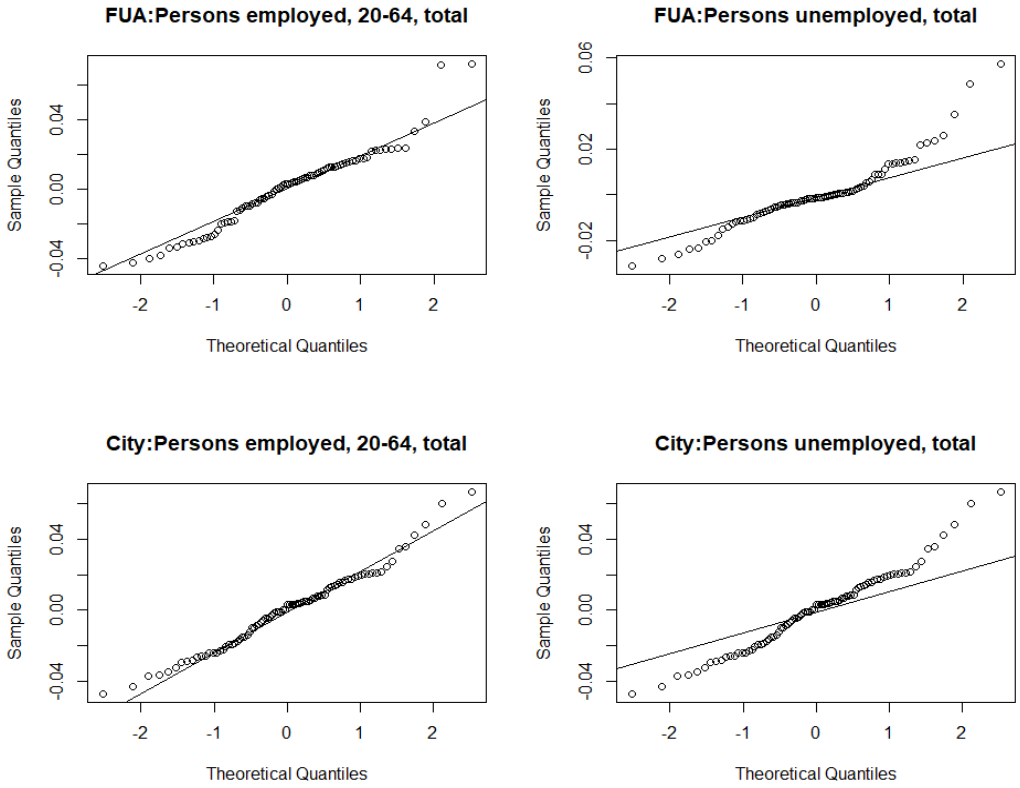
Table 4.5 Indicators of the goodness of fit of the initial model

area	Indicator	N_Obs	AIC	BIC	LL	LLDF	Sigma	MarginalR2	ConditionalR2	MarginalF2	ConditionalF2
FUA	Employed	231504	-17238.2	-16834.47	8658.10	39	0.23	0.773	0.775	3.409	3.452
FUA	Unemployed	231504	-73972.7	-73569.01	37025.38	39	0.20	0.122	0.127	0.139	0.146
City	Employed	144899	-9762.7	-9377.28	4920.37	39	0.23	0.771	0.773	3.361	3.404
City	Unemployed	144899	-38879.1	-38493.63	19478.55	39	0.21	0.127	0.132	0.145	0.152

Table 4.6 Indicators of the goodness of fit of the selected model

area	Indicatore	N_Obs	AIC	BIC	LL	LLDF	Sigma	MarginalR2	ConditionalR2	MarginalF2	ConditionalF2
FUA	Employed	231504	-17142.4	-16697.2	8614.20	43	0.233	0.773	0.775	3.407	3.450
FUA	Unemployed	231504	-76704.8	-76259.7	38395.42	43	0.205	0.132	0.137	0.152	0.159
City	Employed	144899	-9689.8	-9264.8	4887.92	43	0.233	0.771	0.773	3.359	3.403
City	Unemployed	144899	-40462.9	-40037.9	20274.49	43	0.210	0.136	0.142	0.158	0.165

Figure 4.2 Q-Q of the random effect of the domain in the selected model



Finally, in figure 4.2 the Q-Q plot of random effects highlights that there are more areas with extreme values in the residuals than expected, especially for the unemployed. In addition, the Q-Q plot of the unemployed shows a leptokurtic distribution of random effects.

4.2 Analysing small areas estimates

The objective of the present paragraph is to assess the quality of the small area estimates in terms of their correctness and variability with respect to the direct estimates. The following figures allows to compare the small area estimates computed with the MIND estimator with the correspondent direct estimates. From figures 4.3 and 4.4 can be notice that SAE estimates do not present evident systematic bias respect to direct ones. This applies in general to the whole set of parameters of interest as well as to the estimates of both Cities and Fuas.

Figure 4.3 The relationship between direct estimates and small area estimates – City

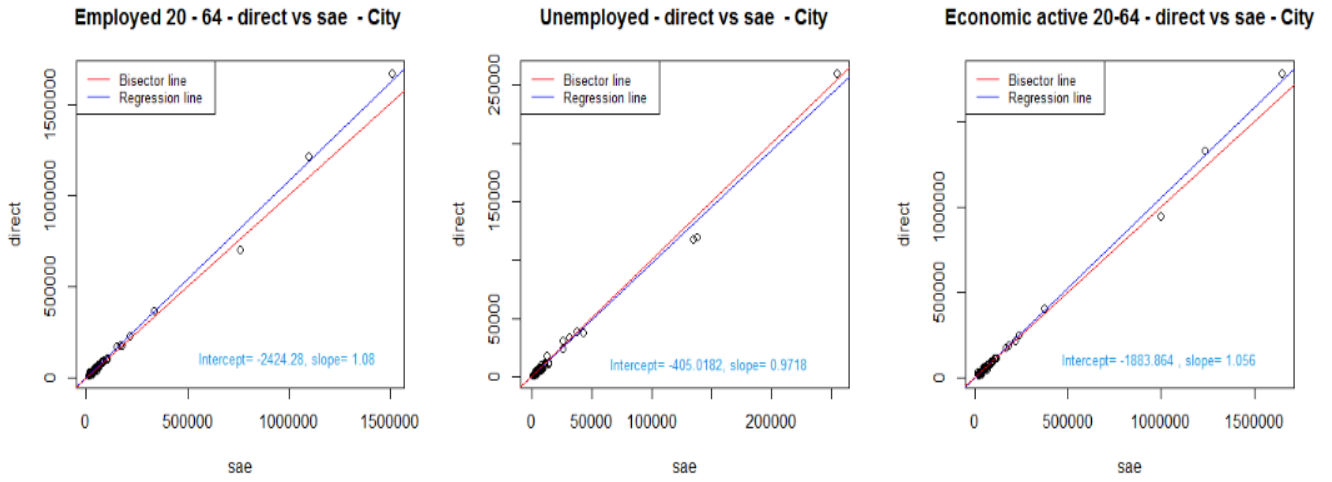
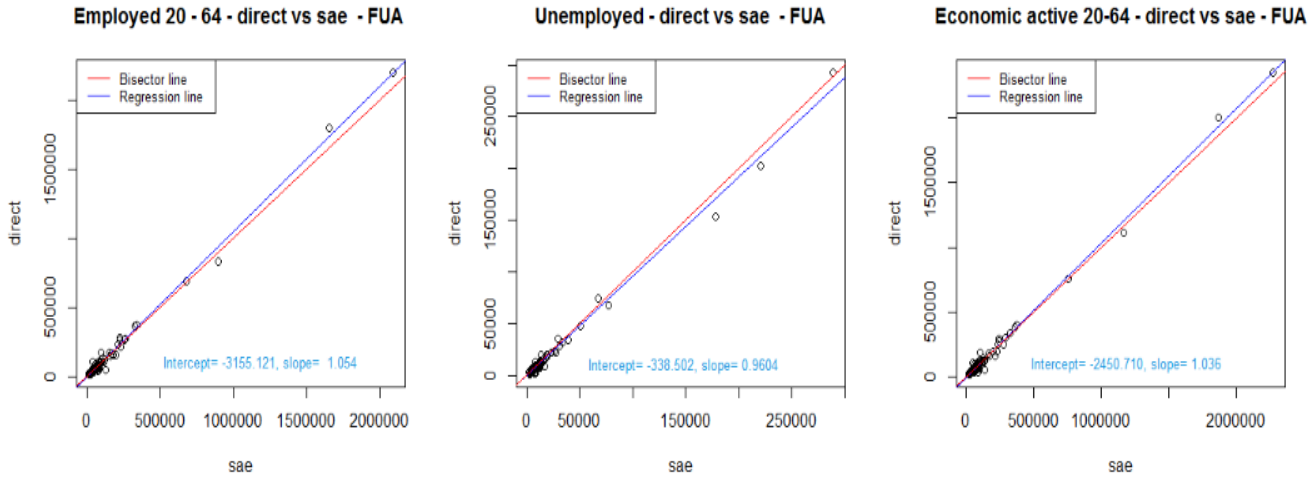
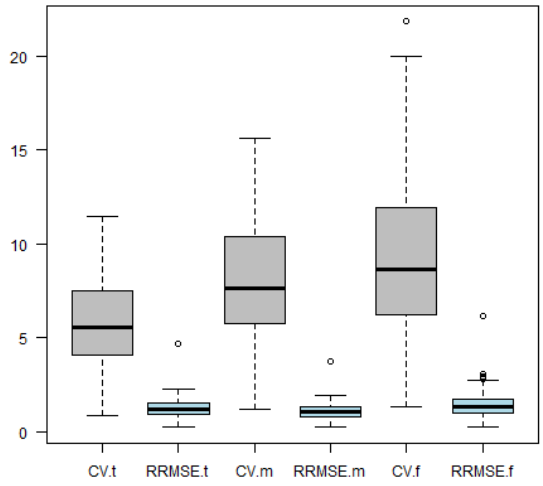


Figure 4.4 The relationship between direct estimates and small area estimates – Fua

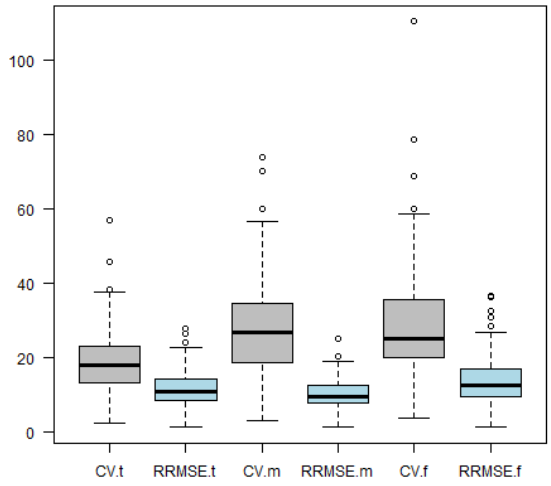


In figures 4.5 and 4.6 it is possible to observe the distribution of the coefficient of variation – as percentage – and the distribution of small area estimates and the related direct estimates, at City and Fua level. Since Cities and Fuas are not planned domains for the Labour Force Survey and the sampling size can be even very low in some of these areas, the coefficients of variation of direct estimates can be even very high. The small area estimation method that we developed allows to produce significant improvements in terms of efficiency compared with the direct estimator, for all the indicators and both the City and Fua estimates.

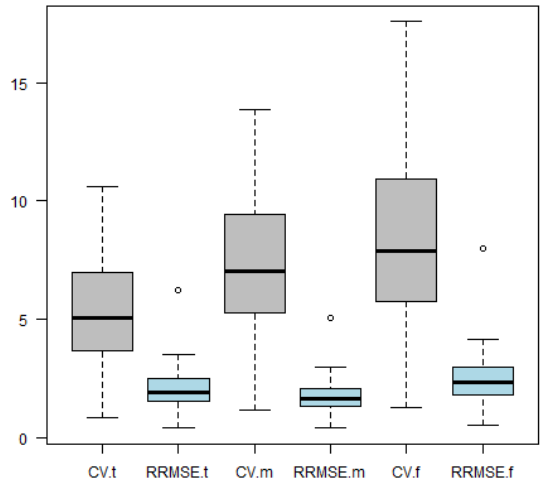
Figure 4.5 The distribution of CV% - City



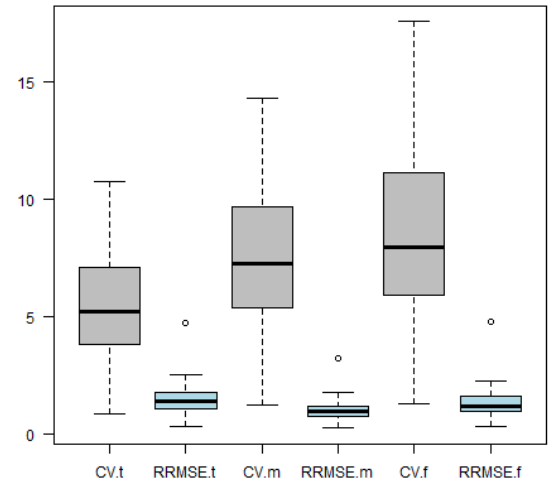
City - Person Employed 20-64 by sex



City - Person Unemployed 15+ by sex

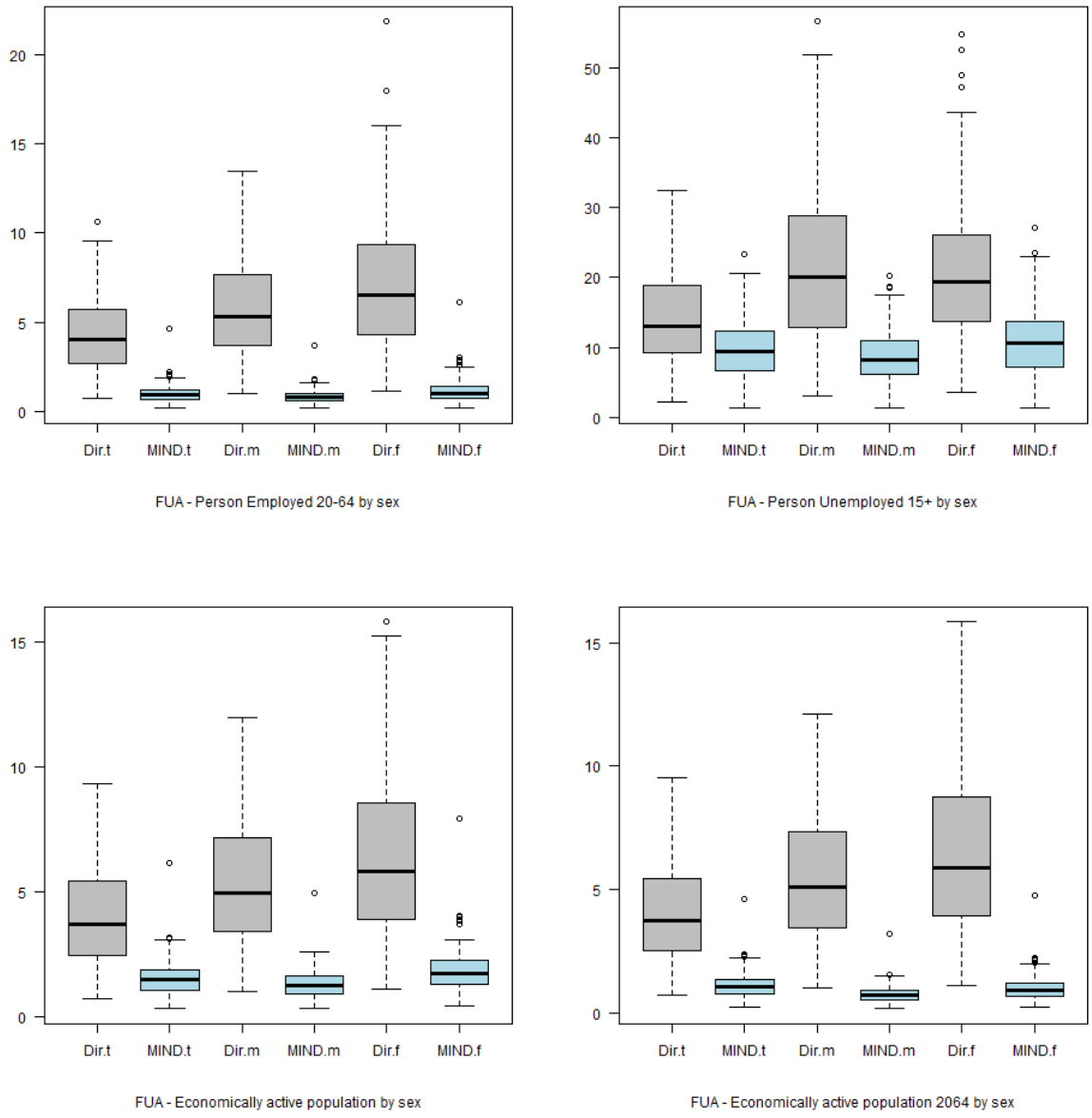


City - Economically active population by sex



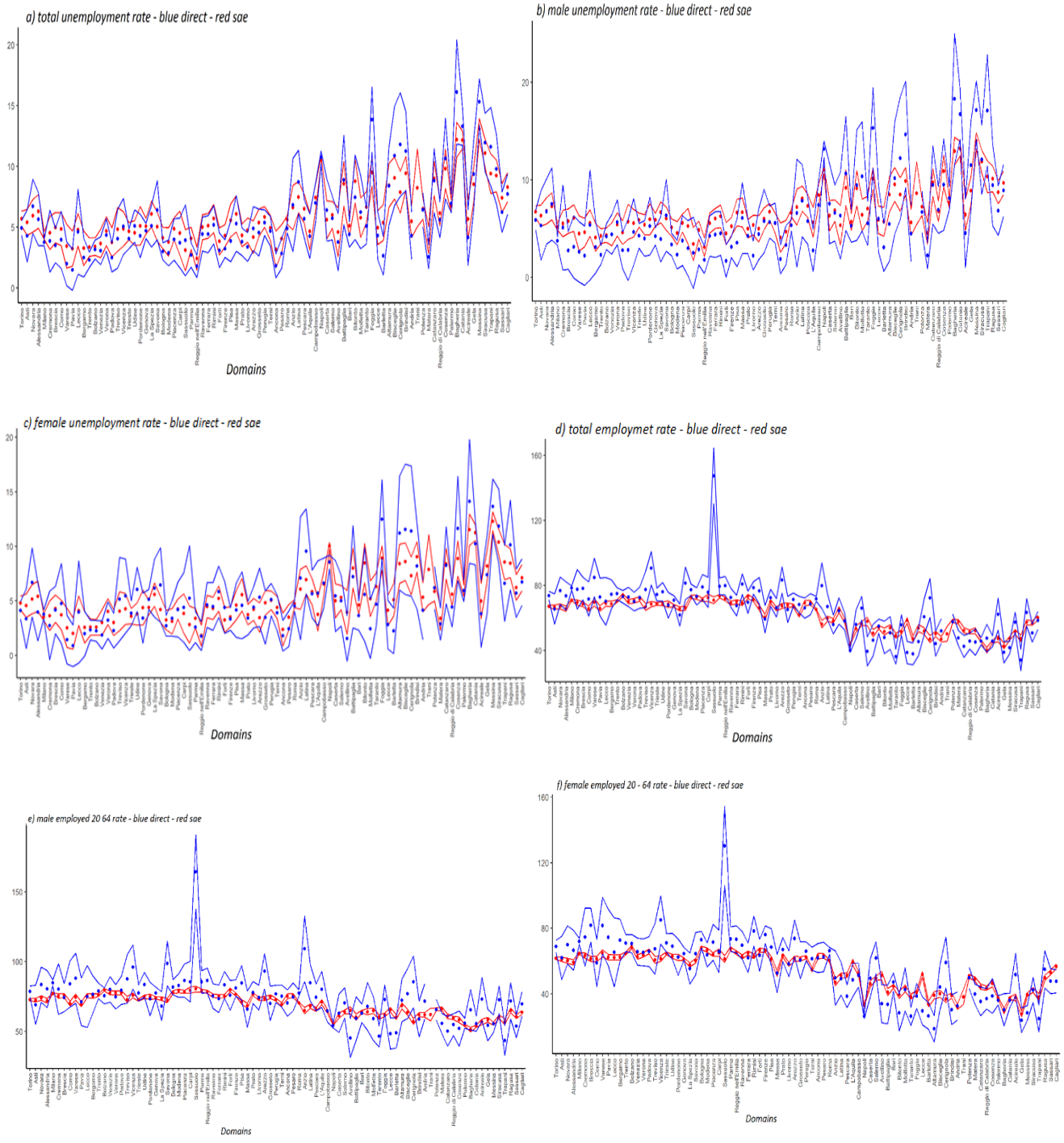
City - Economically active population 2064 by sex

Figure 4.6 The distribution of CV% - Fua

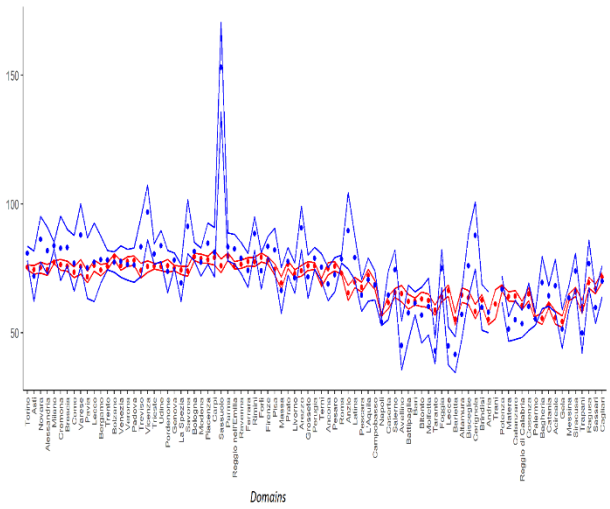


Moreover, in figure 4.7 it is possible to see that the small area estimates for the parameters of interest regarding Cities fall almost entirely in the confidence interval of the direct estimates, confirming the conclusions already drawn while considering the previous figures. In some cases, anomalous direct estimates were corrected. Similar results can be traced for the parameters estimates regarding Fuas.

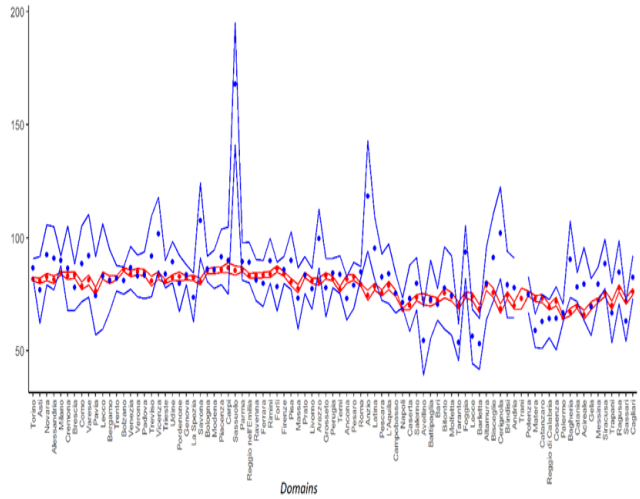
Figure 4.7 SAE estimates and direct estimates and related confidence intervals



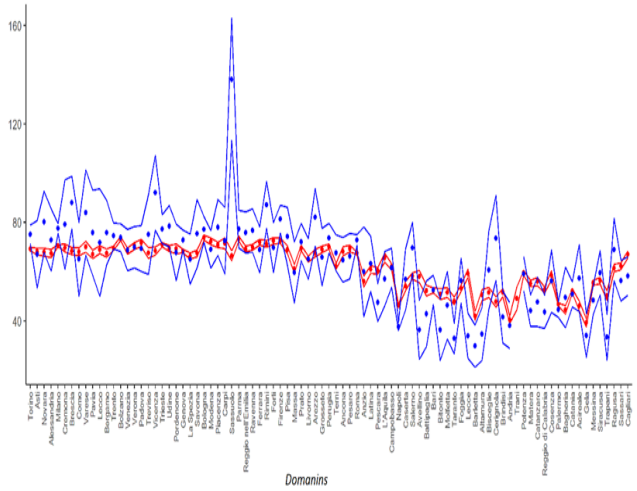
g) economic active people 20-64 rate - blue direct - red sae



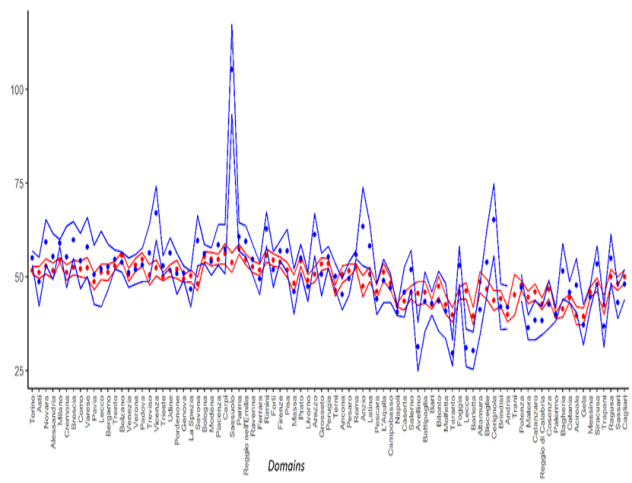
h) male economic active people 20 -64 rate - blue direct - red sae



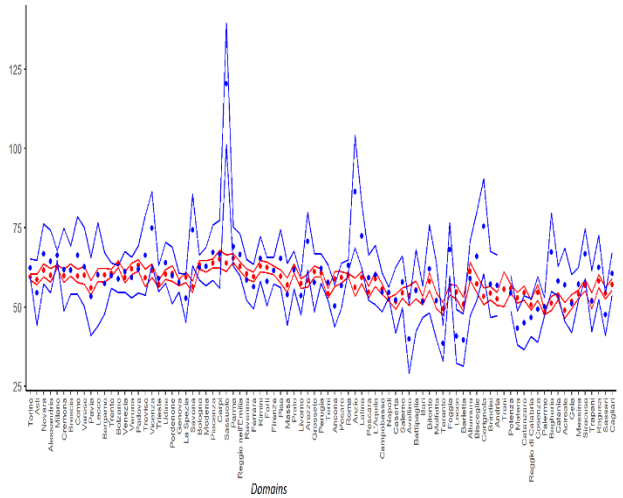
i) female economic active people 20-64 rate - blue direct - red sae



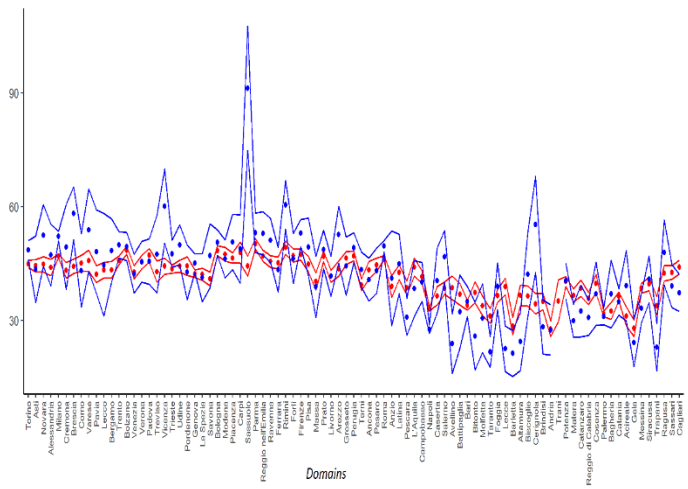
j) total economic active people rate - blue direct - red sae



k) male economic active people rate - blue direct - red sae

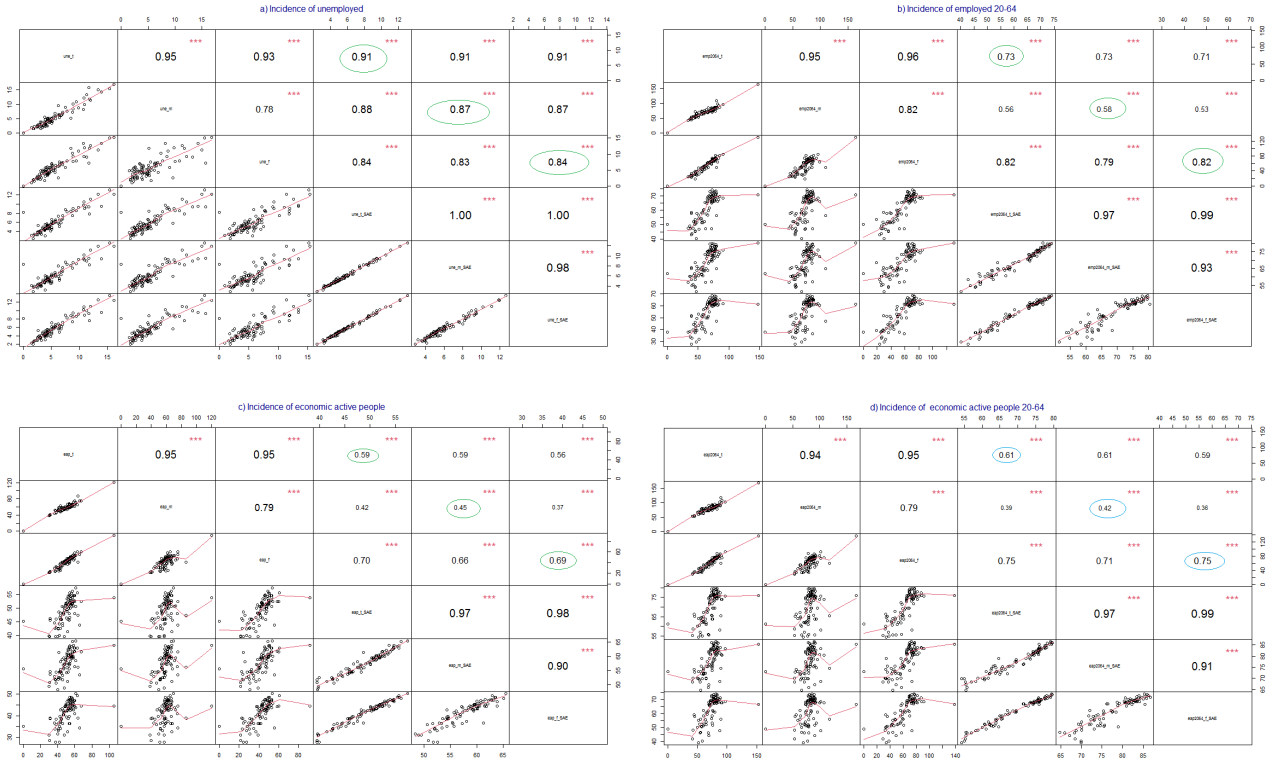


l) female economic active people rate - blue direct - red sae



The correlations between direct estimates and small area estimates of Cities are reported in figure 4.8. Similar results can be drawn for the estimates at Fua level. The graphs show that the small area estimates are significantly correlated to the related direct estimates, especially with regards to the incidence of the unemployed over the total population. The correlation for employed and economically active population is slightly lower. This is due to the fact that the covariates used for the model fitting are more predictive for employment and economically active population than for unemployment.

Figure 4.8 Correlation between direct estimates and SAE estimates at City level – incidence with respect to the population.



5. An overview on the set of small area estimates produced at City and FUA level for years 2018-2020

The aim of this paragraph is to present some analysis on the main results of small area estimates produced at City and Fua level for years 2018-2020. In particular, the analysis is focused on the changes of the estimates over the three years, comparing SAE and direct estimates.

In figure 5.1 percentage yearly variations (2020 vs 2019 compared with 2019 vs 2018) of SAE estimates of employed aged 20-64 and unemployed persons, over FUA and 13 greatest Cities are shown. Dots in the scatter plot are painted according to the precision of the corresponding direct estimates: green dots correspond to 25% of the areas in which the direct estimates' CVs are the lowest (high precision); red dots correspond to 25% of the areas in which the direct estimates' CVs are the highest (low precision); yellow dots correspond to the remaining 50% of the areas (medium precision). The graphs show that the range of yearly variations for the estimates of employed persons is about 8 percentage points in the comparison 2019 versus 2018 and about 10 percentage points

in the comparison 2020 versus 2019. As regards the estimates of unemployed persons, the range of yearly variations is higher (about 60 percentage points), mainly depending on the lower level of the estimates.

Figure 5.1 Percentage yearly variations (2020 vs 2019 compared with 2019 vs 2018) - SAE estimates of employed aged 20-64 and unemployed persons, over FUA and 13 greatest Cities

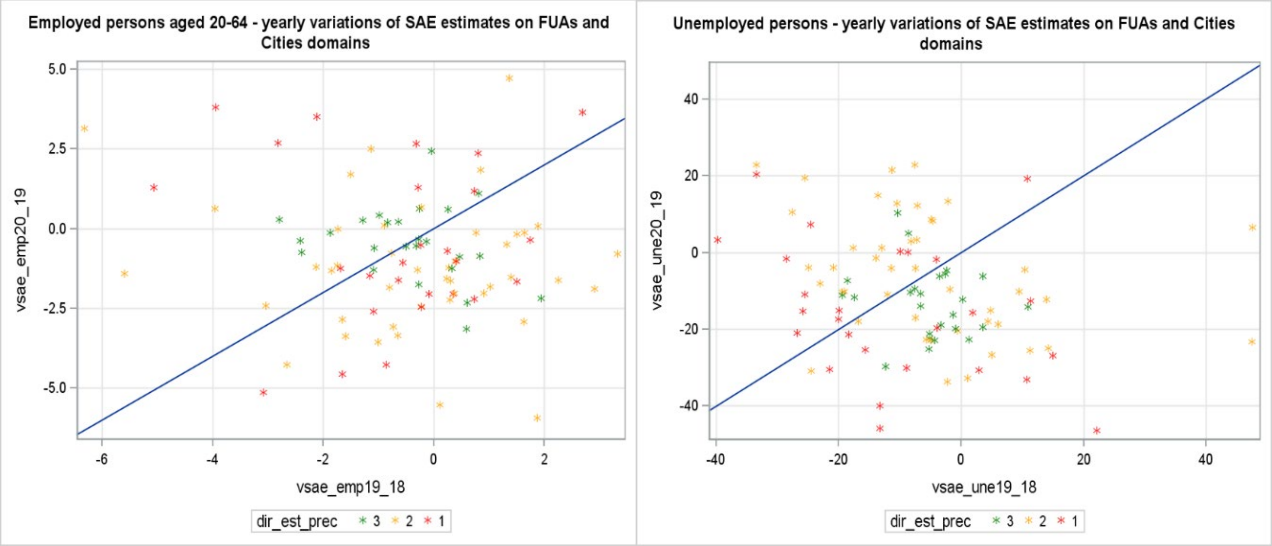


Figure 5.2 compares the percentage yearly variations (2020 vs 2019 and 2019 vs 2018) of SAE estimates with those of direct estimates; data refer to employed persons aged 20-64 over FUA and 13 greatest Cities. Graphs show that the yearly variations of SAE estimates are lower compared with the corresponding variations of direct estimates. This holds in particular for areas in which the precision of direct estimates is lower (yellow and red dots). SAE models exploiting a wide set of covariates, highly correlated with the employment status, allow to significantly improve the precision of the estimates compared with the direct ones, and this has an impact also in the reduction of yearly variations.

Figure 5.3 shows the same graphs referred to unemployed persons: the yearly variations of SAE estimates are rather close to those referred to direct estimates, even if some outliers of the latter are corrected by the former. The covariates used in the SAE models are not so correlated with unemployment, so SAE estimates are rather close to direct estimates, especially when they are more precise.

Figure 5.2 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)

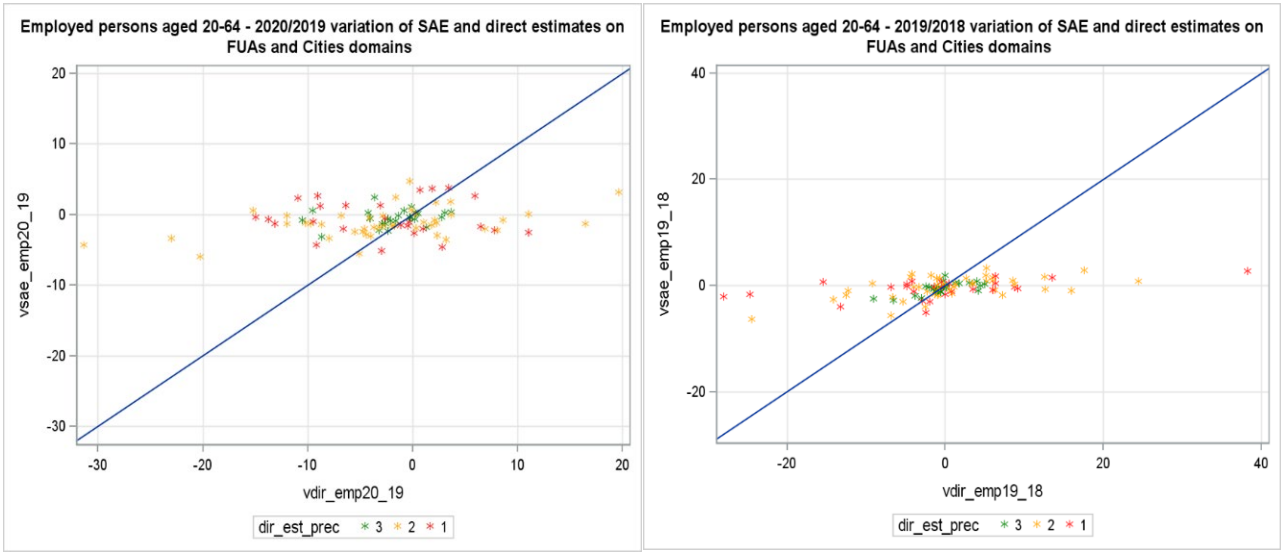
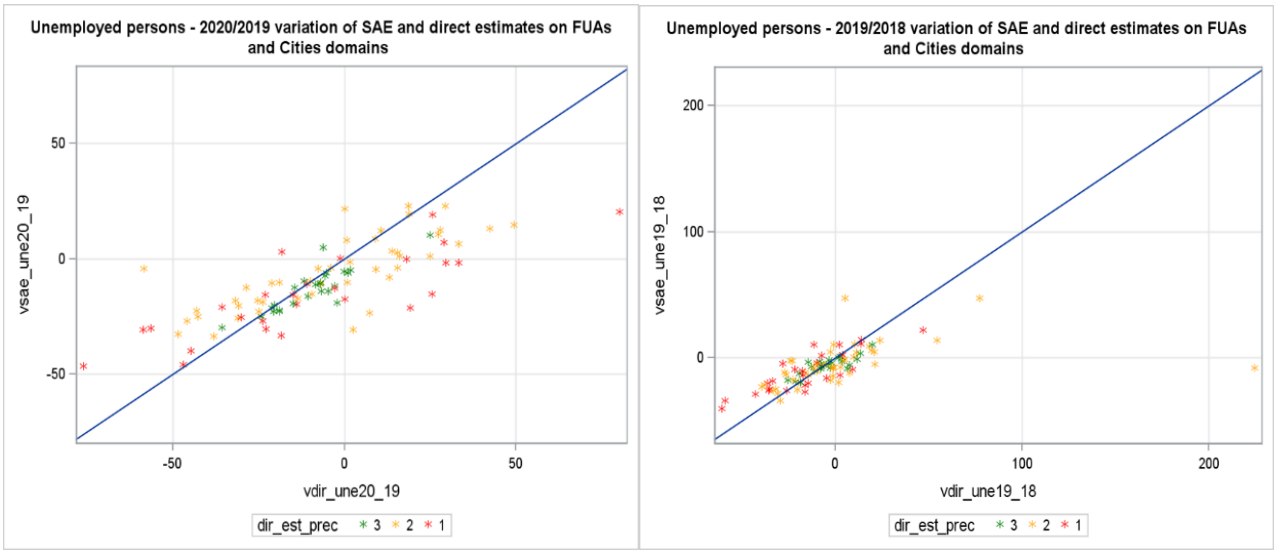
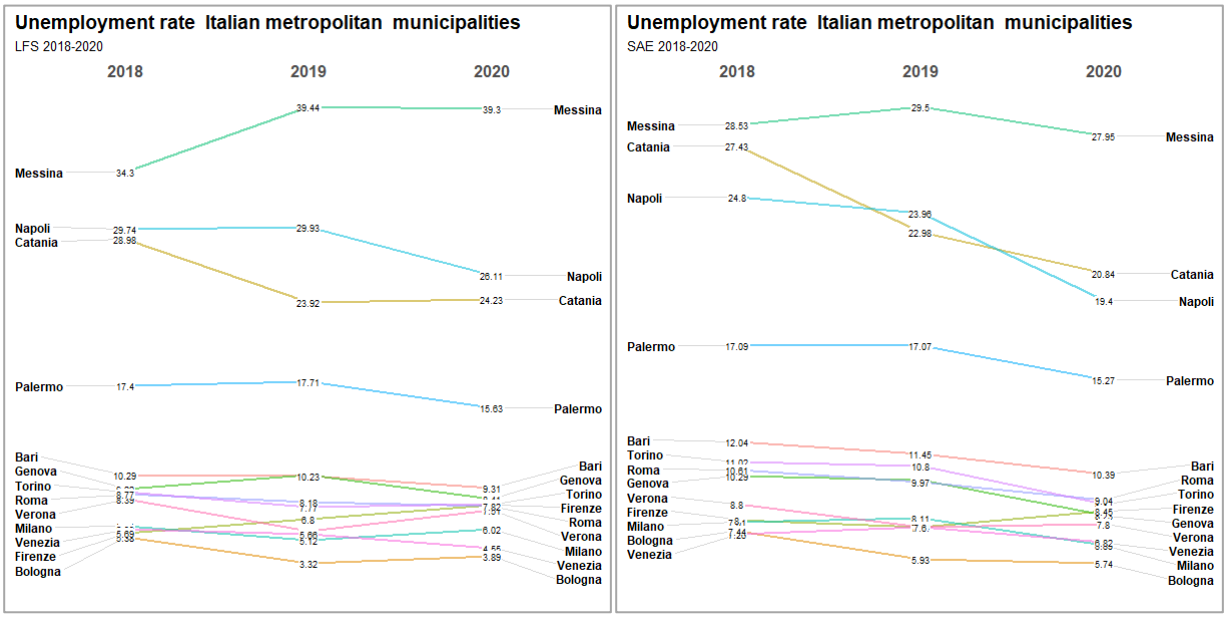


Figure 5.3 Percentage yearly variations of SAE estimates compared with direct estimates - unemployed persons over FUA and 13 greatest Cities (2020 vs 2019 and 2019 vs 2018)



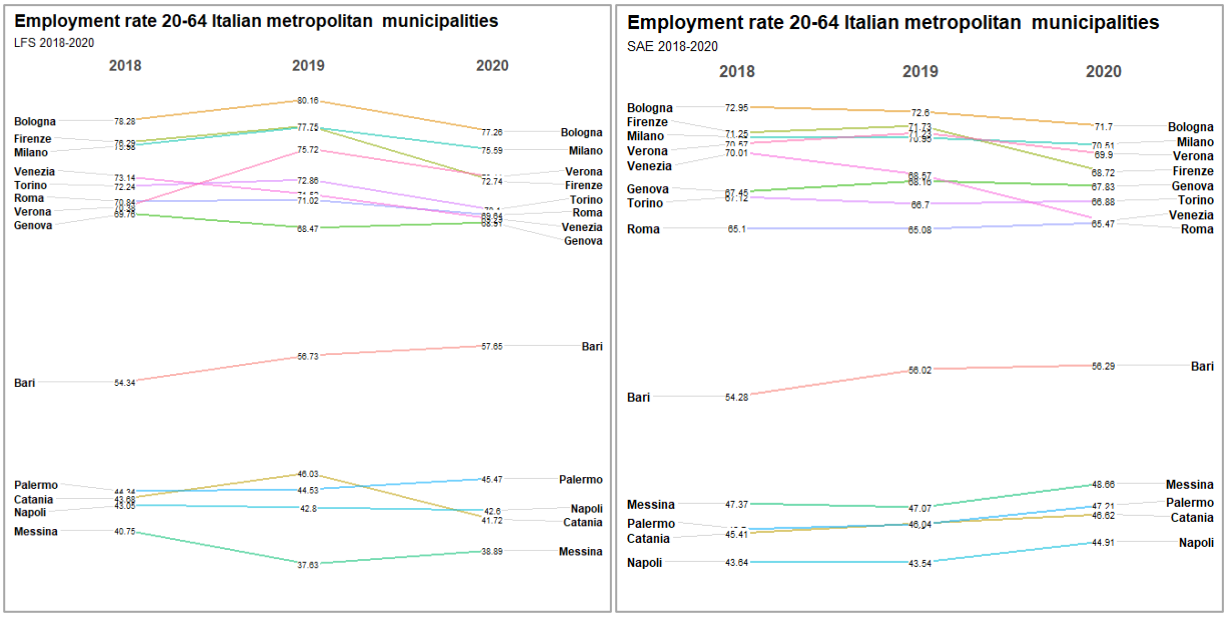
The following figures show a focus on the largest municipalities, the metropolitan ones, with the largest sample size. In particular, figure 5.4 compares the trend of the unemployment estimates calculated through the direct estimator with the corresponding trend of the SAE estimates. It can be seen that the SAE estimates are smoother than the direct ones, and that the variations in the SAE estimates over the three years are also less pronounced than those observable with direct estimates, showing that the improvement in the precision of the SAE estimates produces a reduction of the volatility of the series.

Figure 5.4 Temporal variation of direct and SAE estimates of unemployment rate for metropolitan municipalities



Similar conclusions can be drawn from figure 5.5, in which the results are instead referred to the direct and sae estimates of employment rate.

Figure 5.5 Temporal variation of direct and SAE estimates of employment rate for metropolitan municipalities



6. Conclusions

The aim of the report is to present the methodology developed by Istat to estimate a selection of labour market indicators at City and Fua level. In Italy, these LFS survey based indicators are released at regional and provincial level, but not for the domains of interest (Fuas and Cities), that refer to the OECD-EC city definition. The aim of the WP3 of the Grant Agreement "Sub-national statistics – Italy" between Istat and Eurostat is to fulfil the need for this statistical information. The estimates carried out refer to years 2018, 2019, 2020.

The main goal was the development of a proper small area method to estimate the set of selected indicators in order to improve the efficiency of direct estimates for the domains of interest. In this framework, given the high number of related target indicators to be realised, the main challenge was the production of coherent and consistent estimates. The estimator used is based on a multivariate mixed-effects linear model implemented through the R MIND function. In order to produce the labour market indicators at City and Fua level, the potential of the multivariate approach proposed by MIND has been exploited in two ways: to specify the dependent variable and to define the random effects. The dependent variable was in this case defined as a vector composed by three dichotomous variables representing the categories of the employed, unemployed and inactive individuals. In particular, two models were defined: one for the Fua estimates and another for the City estimates. The coherence of indicators across different domains was reached via a single cross-classification model that included all the domains of interest.

The choice of an estimator based on unit level model has been made in order to better exploit the relevant auxiliary information available from the new integrated system of statistical register. In particular, the Population and Labour Register are used as main sources of information for individuals' demographic characteristics and working condition. In addition, the dataset was integrated with information on social aspects, welfare benefits and type of income deriving from welfare agencies, from the Italian Ministry of Finance and from the national revenue agency.

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

Estimates based on univariate area models have also been computed; however, the use of a specific model for each variable does not guarantee directly (but only in retrospect) the numeric coherence among estimates. Moreover, estimates based on area level models generally allow less efficiency gains with respect to the estimates based on unit level model: in fact, by linking register and survey elementary data, more predictive unit level models can be specified. This is the main reason for which in this document only the results and the evaluation of unit level model based estimates are reported.

The estimates generated by the application of the chosen estimator were compared to the direct estimates carried out in the initial phase of the study. The results show that is possible to achieve significant gains in efficiency with respect to the direct estimates, with particular regards to the estimation of the unemployed persons (total and by sex), for which the sample errors were rather high.

The estimates are currently available online as part of Eurostat's *Cities database* and they can be downloaded selecting Italian Cities and Fuas in the Labour Market section.

References

Battese G.E., Harter R.M., Fuller W.A. (1988), "An Error-Components Model for Prediction of County Crop Areas Suing Survey and Satellite Data" in *Journal of the American Statistical Association, 83(401)*, pp. 28–36. doi:10.1080/01621459.1988.10478561

D'Alò M., Falorsi S., Fasulo A. (2021) "MIND R Package" on CRAN
https://cran.rproject.org/web/packages/mind/index.html

ESSnet on SAE (2012), *Guidelines for the application of the small area estimation methods in NSI sample surveys*, https://ec.europa.eu/eurostat/cros/system/files/WP6-Report.pdf

EURAREA Consortium (2004), "Enhancing Small Area Estimation Techniques to Meet European Needs", in *Project Reference Volume, Deliverable 7.1.4*.

EUROSTAT, *Cities database*, https://ec.europa.eu/eurostat/web/cities/data/database

EUROSTAT (2018), *Methodological manual of territorial typologies*,
https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Territorial_typologies_manual

Rao J.N.K., Molina I. (2015), *Small Area Estimation*, John Wiley & Sons Eds, doi:10.1002/ 9781118735855