UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert meeting on Statistical Data Editing

7-9 October 2024, Vienna

Selective editing for the processing of new Services Producer Price Indices (SPPIs) from indirect data sources[1]

Diego Bellisai, Marco Lattanzio, Tiziana Pichiorri, Simona Rosati
Istat - Italian National Institute of Statistics, Italy

bellisai@istat.it, lattanzio@istat.it, pichiorr@istat.it, sirosati@istat.it

I. INTRODUCTION

1. To comply with the Regulation EBS 2152/2019 on European business statistics, the Italian National Institute of Statistics (Istat) has recently started producing and disseminating quarterly producer price indicators for services pertaining to statistical units primarily involved in activities of divisions 74 and 82 of the Statistical Classification of Economic Activities, NACE Rev. 2. These indicators aim to measure quarterly fluctuations in business-to-business prices for services, focusing exclusively on transactions between businesses, thereby excluding sales of services to consumers [OECD/Eurostat, 2014].

2. The estimation of these indices relies on hourly labour cost data from Oros, the Quarterly Survey on Employment and Labour Cost [Istat-Oros, 2019], which in turn takes these data from the Italian National Social Security Institute (INPS) administrative archives.

3. However, administrative as well as survey data may contain errors, such as measurement errors, that can lead to biased estimates when data are used for statistical purposes. Given the nature of the variables involved (labour cost, regularly paid hours worked) and the required timeliness between data availability and their dissemination, we adopted a selective editing approach to identify outliers and influential errors. Specifically, we employed a method based on contamination normal models; this method is implemented in the *R* package named SeleMix, developed at Istat [Guarnera and Buglielli, 2013]. SeleMix aims to detect units with the most influential values, i.e. potential errors with the highest impact on the target estimates. Suspicious outliers and influential errors are then flagged for manual review by subject matter experts.

4. The paper is organized as follows. Section II introduces the context of the study and provides a description of the new Services Producer Price Indices (SPPIs). Section III illustrates the selective editing approach based on contamination normal models, which is implemented in the *R*-package SeleMix. The editing strategy for the SPPIs is presented in Section IV, while Section V reports the most relevant results derived from the analysis. Finally, Section VI concludes with some considerations on the main results and by outlining potential developments for future works.

[1]The views expressed in this paper belong to the authors and do not necessarily reflect the views or policies of the Italian National Institute of Statistics

II. The new Services Producer Price Indices (SPPIs)

A. Data source, field of observation, units of analysis and survey units

1. Istat produces quarterly indices of producer prices for business-to-business (BtoB) services covering economic activities H, J, L, M and N according to the NACE Rev.2 classification [Istat-PPS, 2024]. For most economic activities within the scope of observation, the indices are derived from data collected through a direct survey on enterprises, conducted quarterly. For the remaining sectors, indices are calculated using administrative data sources and existing databases already available at the Institute.

2. The unit of analysis is the service sold on both the domestic and the foreign markets by businesses to a clientele comprising enterprises and/or institutions belonging to the Public Administration. The survey units are represented by enterprises resident in Italy that provide services to other enterprises and/or to the Public Administration, both nationally and internationally. The target variable is the BtoB production price, which is the quarterly average price of the service sold. This price includes contributions received from the producer, discounts, rebates, and surcharges applied to the customer. However, it excludes VAT and similar directly deductible taxes related to turnover, as well as all taxes on invoiced services. The definition of the production price follows the guidelines of the Commission Implementing Regulation (EU) 2020/1197.

3. In the following, we will focus on some of the economic sectors covered by administrative sources. The field of observation includes services categorized under the following 3-digit NACE Rev.2 codes:

- 741 - Specialized design activities
- 742 - Photographic activities
- 743 - Translation and interpretation activities
- 749 - Other professional, scientific and technical activities n.e.c.
- 821 - Office administrative and support activities
- 829 - Business support service activities n.e.c.

4. Based on information obtained from the Frame-SBS Statistical Register, annually produced by Istat, these activities have been identified as those whose share of revenue (and thus, indirectly, production) given by labour cost is prevailing. This allows us to utilize available quarterly administrative data on hourly labour cost as a reliable approximation for producing price indicators.

5. The reference universe comprises enterprises that annually submit the ISA (Synthetic Reliability Indicators) models to the Revenue Agency and whose primary economic activities fall within the aforementioned categories. Since the compilation of the ISA models is mandatory only for enterprises with a turnover below a specified threshold for a particular activity, this set is supplemented with enterprises from the ASIA Statistical Business Register that have primary economic activities in the six described groups and turnover above the threshold. The reference universe also provides the annual weighting system of the indices through the revenue/turnover variable.

6. Samples are composed of enterprises identified by linking administrative data from the Italian Revenue Agency Register (ISA models) and the National Social Security Agency, INPS (Social Security Positions), as well as enterprises selected from the ASIA Enterprise Business Register. The sample size

for the aforementioned economic activities is of about 20,000 enterprises altogether. These samples are updated annually.

7. The quarterly data on hourly labour cost are derived from the statistical processing of INPS data conducted by the Oros survey. These data include information on wages, social contributions, employees, labour cost and paid hours worked, available at the level of individual social security contributions units for each enterprise.

B. **Indices formulas**

1. As far as the Service producer price quarterly indices calculation is concerned, the process starts from the social security contributions unit elementary chained-base indices:

$$I_{i,j,t} = \frac{clor_{i,j,t}}{clor_{i,j,0}} \ , \tag{1}$$

where $clor_{i,j,t}$ is the average hourly labour cost for the quarter t, i is the enterprise and j is the social security contributions unit, while $clor_{i,j,0}$ is the average hourly labour cost for the base quarter, i.e. the fourth quarter of the previous year.

2. Subsequently, the elementary indices are used to compute indices at the enterprise level as a weighted geometrical mean over the n_i social security contributions unit belonging to a given enterprise:

$$I_{i,t} = \left(\prod_{j=1}^{n_i} I_{i,j,t}^{w_j} \right)^{\frac{1}{\sum w_j}} \ . \tag{2}$$

where the weights w_j are given by the quarterly paid hours for the j-th social security contributions unit. Finally, the aggregate indices (at the level of economic activity) are obtained through a weighted mean of the enterprise indices:

$$I_t^K = \sum_{i=1}^{n_K} I_{i,t}\, \omega_{i,0} = \sum_{i=1}^{n_K} I_{i,t}\, \frac{ric_{i,0}}{\sum_{l=1}^{n_K} ric_{l,0}} \ , \tag{3}$$

where the weighting coefficients $\omega_{i,0}$ are given by the share of the enterprise yearly revenues for the K-th economic activity, related to the previous year with respect to the index.

3. The aggregate indices in the previous formula can also be expressed in a way which makes clearer their relationship with the hourly labour cost at the enterprise level and makes the use of SeleMix suitable, also to detect influential values for aggregate indices. Indeed, they can be re-written as:

$$I_t^K = \sum_{i=1}^{n_K} clor_{i,t}\, \pi_{i,0} = \sum_{i=1}^{n_K} clor_{i,t} \frac{1}{clor_{i,0}} \omega_{i,0} \ , \tag{4}$$

where $\pi_{i,0}$ defines the influence weight to be used in the process of influential observations detection and $clor_{i,0}$ is the average quarterly labour cost of the last quarter of the previous year.

III. A multivariate selective editing approach

A. The contamination model

1. According to the authors [Di Zio and Guarnera, 2013], true unobserved data, typically expressed in log-scale, can be represented by independent realizations of p random variables $(Y_{i1}, Y_{i2}, \ldots, Y_{ip})$ following a Gaussian distribution with mean vector μ_i and common covariance matrix Σ, where i indicates the i-th sampled unit, for $i = 1, 2, \ldots, n$. Furthermore, for each unit a set of q covariates can also be available, $x_{i1}, x_{i2}, \ldots, x_{iq},$. Therefore, the previous assumptions can be expressed by the model

$$Y^* = XB + U,$$

where Y^* is the $n \times p$ true data matrix, X is the $n \times q$ covariate matrix, B is the $q \times p$ coefficient matrix and U is the $n \times p$ matrix of normal residuals identically and independently distributed (i.i.d.) with zero mean and covariance matrix Σ.

2. It follows from the previous assumption that

$$f(y^*) = N(y_i^*; \mu_i, \Sigma), \quad f(u_i) = N(u_i; 0, \Sigma), \quad i = 1, 2, \ldots, n\ ,$$

where $f(y^*)$ and $f(u_i)$ are the marginal distributions of the true value and of the residual, respectively, for the i-th unit, and $N(y; \mu, \Sigma)$ denotes the Gaussian distribution with mean vector μ and covariance matrix Σ.

3. A particular feature of this method is that only a specific proportion of the set of units is considered erroneous. The dataset is therefore divided into two subsets: one consisting of the erroneous units and the other one consisting of the error-free units. The membership of each unit to either group is not known. This implies that data can be represented by a latent class model, where the latent variable is a binary variable indicating the presence or absence of error for each unit. In other words, the error mechanism is intermittent, which is a crucial aspect of this model. The error mechanism can be modeled by introducing Bernoullian random variables I_i with parameter π, where $I_i = 1$ if an error occurs on unit i and $I_i = 0$ otherwise, for $i = 1, 2, \ldots, n$.

4. After specifying the true data distribution and the error mechanism, the observed data distribution can be derived by combining two regression models that share the same coefficient matrix B and have proportional residual variance-covariance matrices. This distribution can be estimated by maximizing the likelihood from n sample units using an Expectation Conditional Maximization (ECM) algorithm [Meng and Rubin, 1993]. Since the distribution of the unobserved "true" data and the error mechanism are specified separately, it is possible to estimate the magnitude of the error. This allows us to identify errors that have high impact on target estimates (influential errors), which is the purpose of selective editing.

5. The contamination model can be used to obtain predictions or "anticipated values" for true unobserved data. Predictions are obtained from the distribution $f(y^* \mid y_i)$ of the true data conditional on the observed data (possibly including values of error-free covariates X). A straightforward application of the Bayes formula provides the posterior probabilities that a unit with observed values y_i belongs to the correct or erroneous data group.

6. Once the prediction formula is obtained, the expected error can be defined in terms of two components: the "risk component" and the "influence component". These components are incorporated into the score function definition. In practice, parameters involved in the expected error are unknown and have to be estimated [the algorithm is described in Di Zio and Guarnera, 2013].

7. Observations are then ordered by their global score, and all units exceeding a specified threshold are selected. The threshold should be set to ensure that the impact of errors in the unedited observations on the target estimates remains negligible. The threshold essentially represents the level of accuracy of the estimates of interest.

B. The SeleMix package

1. The SeleMix package consists of a set of *R* functions designed to implement the method illustrated in the previous paragraphs. It consists of three key functions: ***ml.est***, ***pred.y***, and ***sel.edit***, each serving a distinct purpose within the selective editing process. These functions facilitate the estimation of model parameters, the prediction of variable values, and the selection of a subset of observations affected by potential influential errors, leveraging both local and global scores. Additionally, the ***sel.pairs*** function provides valuable graphical tools to enhance analysis [for details see Guarnera and Buglielli, 2022, 2013]

IV. The editing strategy for the SPPIs

1. For each sector of economic activity, a list of individual security contributions units pertaining to active enterprises is defined annually. From this list, quarterly average data are extracted relating to labour costs, paid hours, social security contributions, wages, and number of employees. For most enterprises, data for quarter t are available, as well as data for quarter $t-4$ (same quarter of the previous year).

2. The target variable object of the editing and imputation process is the average quarterly labour cost at time t ($clor_t$). Selective editing procedures have been implemented through two distinct linear regression models of the form

$$clor_t^{i,K} = A_t^K + B_t^K clor_{t-4}^{i,K} + \varepsilon^i \tag{5}$$

where i stands for the enterprise, K for the economic activity and ε^i is a vector of normally distributed random residuals with zero mean. Specifically:

(1) the model with one covariate (Model 1) aims to identify influential units for the responding units (enterprises) for which the average labour cost at $t-4$ is available.
(2) the model with no covariates (Model 2) aims to identify influential units on all responding units.

3. Before applying the two models, the data set was divided into two subsets based on the 99-th percentile of the distribution of the target variable ($clor_t$) and the share of per capita revenue given by the ratio $\frac{\omega_{i,0}}{DITM_t^i}$, where $DITM_t^i$ represents the enterprise's quarterly average number of occupied positions. This aims to single out units representing a specific subpopulation with distinct characteristics compared to the rest of population. Then, selective editing was applied on both subsets of units, i.e. Model 1 and Model 2 were applied to the subset above the 99-th percentile, as well as to the rest of units below the 99-th percentile, for $i = 1, 2, \ldots, n_K$.

4. The definition of the list of influential units is given by the union of the influential units of Model 1 and Model 2 (in this case, discarding the influential units already identified by Model 1). Manual revision of all the identified influential units should be performed by subject matter experts, exploiting all the auxiliary information available, such as administrative data on wages, quarterly average employees, and social security contributions. Since the revision is ongoing, we assumed that all units

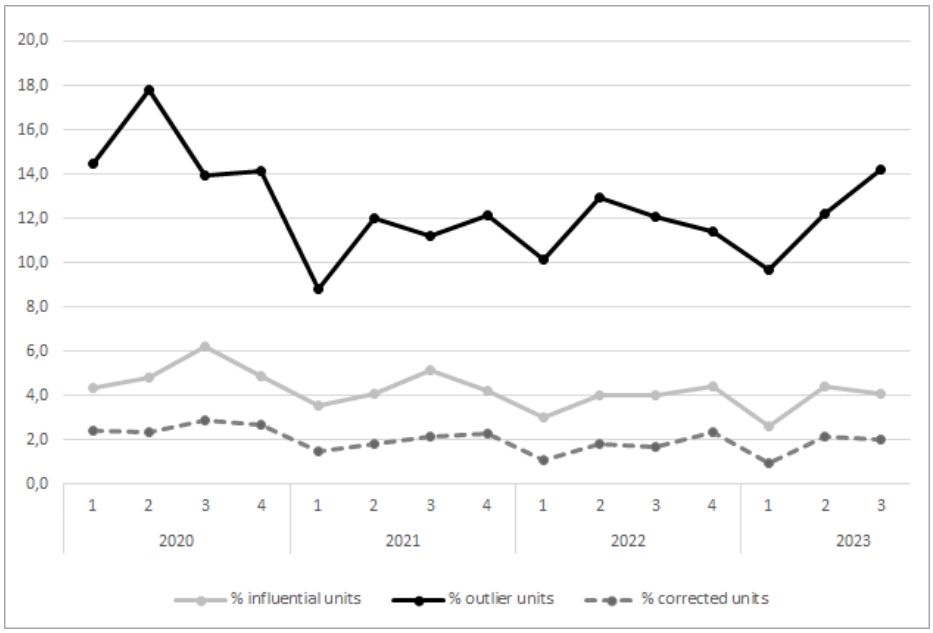

FIGURE 1. Percentage of outlier, influential and corrected units out of the total units, NACE groups as a whole.

identified as both influential and outliers were erroneous and replaced them with the corresponding predicted values.

V. **Results**

1. In this section, the main results of the application of SeleMix to the observed data are presented. As explained in Section II, data are related to NACE groups 741, 742, 743, and 749 from Section M, as well as groups 821 and 829 from Section N. Figure 1 illustrates the quarterly trends of influential, outlier, and corrected units across all NACE groups. The percentage of influential units identified by SeleMix fluctuates between 3% and 6% of the total, reaching its lowest point in the first quarter of each year. At the same time, the trend of corrected units is quite similar to that of influential units, but with lower percentages, ranging from 1% to 3%. In contrast, the percentage of outliers is significantly higher, averaging around 13%, showing a less coherent trend.

TABLE 1. Average number of units, percentage of influential, outlier and corrected units, by NACE group.

Group	**Average N. of units**	**% influential units**		**% outliers**		**% corrected units**	
		mean	**min-max**	**mean**	**min-max**	**mean**	**min-max**
M 741	1,534	2.6	0.3 - 5.7	9.1	5.1 - 15.8	1.0	0.1 - 3.1
M 742	274	3.9	0.4 - 11.0	24.7	7.7 - 39.1	2.9	0.0 - 8.9
M 743	200	1.0	0.0 - 4.1	12.8	3.0 - 27.5	0.6	0.0 - 2.4
M 749	1,928	4.0	0.4 - 6.6	8.5	3.7 - 14.2	1.4	0.4 - 2.3
N 821	1,486	3.1	0.2 - 6.5	13.5	3.8 - 22.9	1.8	0.0 - 3.8
N 829	11,674	4.7	3.3 - 6.3	13.3	8.9 - 18.0	2.3	1.2 - 3.0

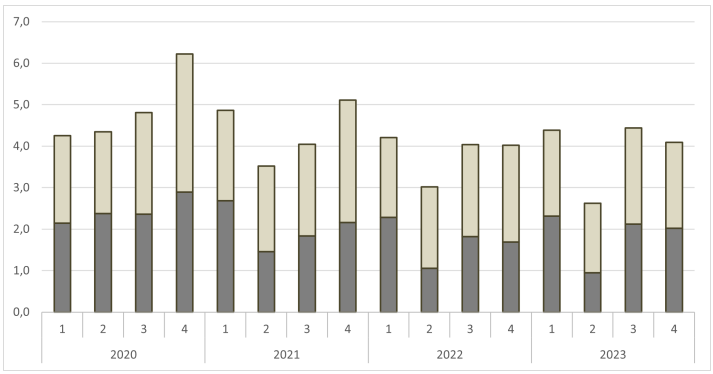

FIGURE 2. Percentage of influential units and corrected influential units (grey segment), NACE groups as a whole.

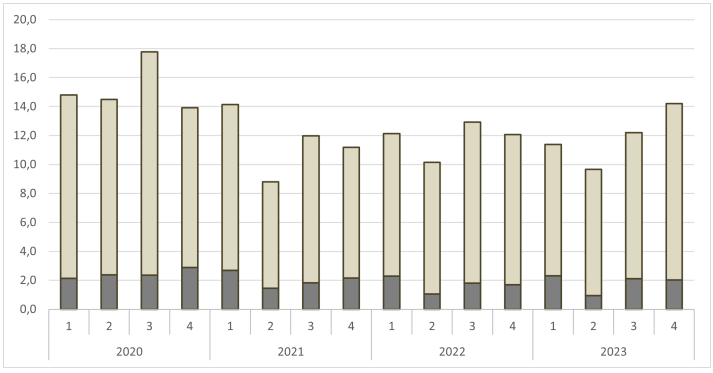

FIGURE 3. Percentage of outlier units and corrected outlier units (grey segment), NACE groups as a whole.

2. Table 1 presents the quarterly averages from 2020 to 2023 for each group, including the number of enterprises and the percentages of influential, outlier, and corrected units. These results revealed that the various groups exhibit significant heterogeneity in terms of average numerical composition, with group 829 comprising over 11,600 enterprises, while group 743 consists of only around 200 units. Furthermore, these groups vary not only in the quantities, but also in the levels and variability of outlier and influential units identified through the selective editing procedures.

3. As illustrated in Figures 2 and 3, about 50% of the units identified by SeleMix as influential are also outliers across all NACE groups in each quarter of the analyzed period. The corrected units, which are both influential and outliers, represent approximately 2% of the total units and account for an average of 16% of all outliers.

4. The analysis of the trends for influential, outlier, and corrected units showed distinct patterns between the two models. In the model with covariate (Figure 4, left side), the number of outliers is markedly high, while influential units are scarce. In contrast, the model without covariate (Figure 4, right side) showed comparable numbers of influential and outlier units.

5. Moreover, as shown in Figure 4, the first model identified a significant number of outliers across all NACE groups, whereas the number of outliers and influential units detected by the second model is quite similar. This is probably due to the strong linear relationship between the target variable value at time t, $clor_t$, and its covariate, i.e. the corresponding value from the same quarter of the previous

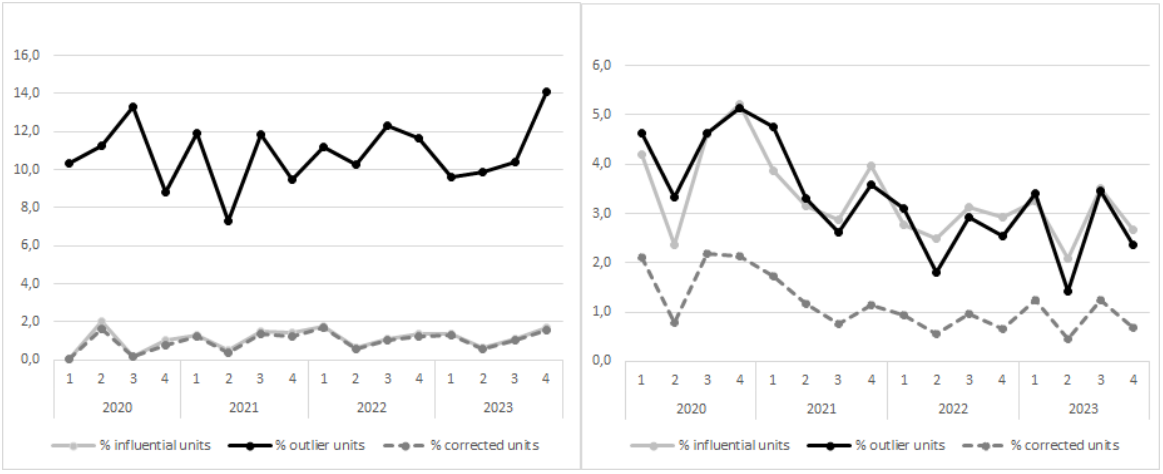

FIGURE 4. Model 1 and Model 2: Percentage of outlier, influential and corrected units out of the total units, NACE groups as a whole.

year, $clor_{t-4}$. As an example, Figures 5 and 6 illustrate the output from SeleMix for NACE group 741 in the first quarter of 2020.

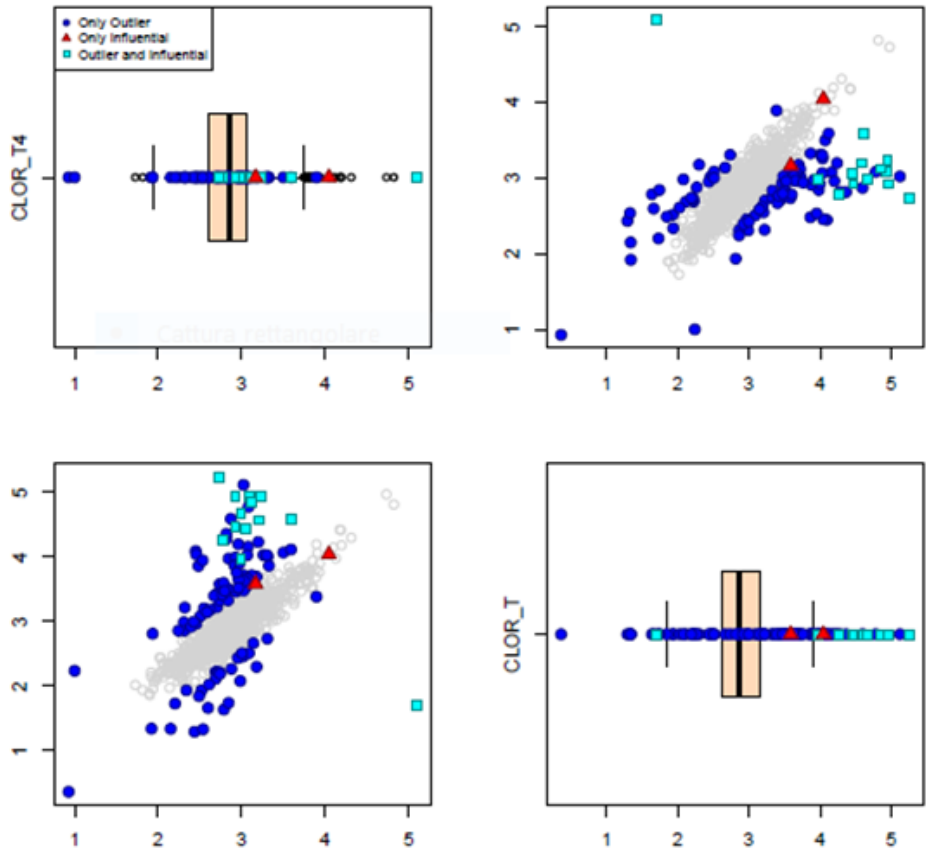

FIGURE 5. NACE group 741, Q1 2020, units with values below the 99-th percentile. Scatterplot with outliers and influential errors highlighted.

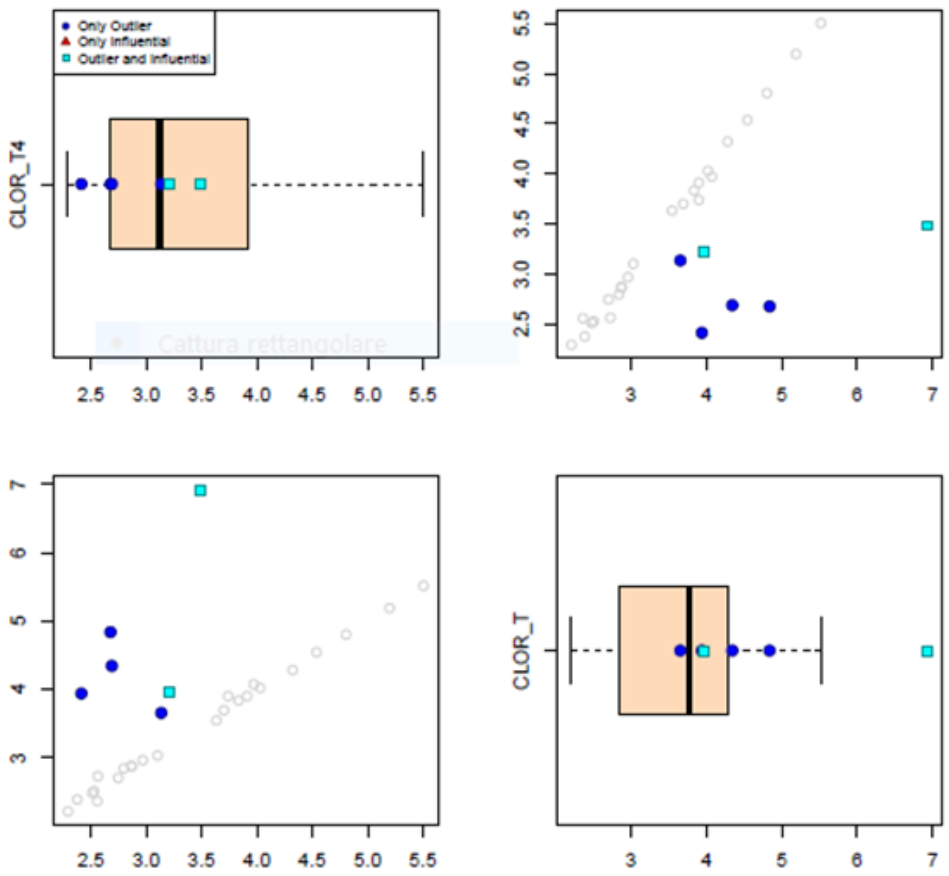

FIGURE 6. NACE group 741, Q1 2020, units with extreme values (exceeding the 99-th percentile). Scatterplot with outliers and influential errors highlighted.

VI. Conclusions and further developments

1. A selective editing approach using contamination normal models was applied to the new Services Producer Price Indices. Before defining the model, the data were analyzed to identify potential subpopulations with differing behaviors. Once these subgroups were identified, two models were specified: Model 1, which included one covariate, and Model 2, with no covariates.

2. The Model 1 that focused on units above the 99-th percentile performed differently from the same model that included all other units. In the latter case, indeed, an unusually high threshold has been set to identify potential influential units. As shown in the results (Figure 5), we observed a strong relationship between the dependent variable and its covariate, as most observations cluster closely around the line. This indicates that even minor variations in the expected error can lead to the presence of numerous outliers. Therefore, only by increasing the threshold could influential units be identified. On the other hand, SeleMix did not always achieve the convergence for specific quarters and NACE groups due to the limited number of units included. Therefore, a different solution should be found for these data.

3. To enhance the performance of Model 1, it is reasonable to consider discarding observations that are highly similar between time t and $t-4$. However, further analysis is needed to explore which characteristics distinguish the two subset of units, those below the 99-th percentile versus the remaining units, also with respect to their influential units.

4. For the sake of simplicity, the potential influential errors identified by the two models were replaced with the corresponding predicted values resulting from the ***pred.y*** function of SeleMix. Nevertheless, all identified influential units should be manually reviewed by experts; thus, it is reasonable to expect that fewer units will need to be corrected.

References

M. Di Zio and U. Guarnera. A Contamination Model for Selective Editing. *Journal of Official Statistics*, 29(4):539–555, 2013. doi: http://dx.doi.org/10.2478/jos-2013-0039.

U. Guarnera and M.T. Buglielli. SeleMix: an R Package for Selective Editing, 2013. URL https://www.istat.it/it/files/2014/03/SeleMix-vignette.pdf.

U. Guarnera and M.T. Buglielli. Package: SeleMix, 2022. URL https://cran.r-project.org/web/packages/SeleMix/SeleMix.pdf.

Istat-Oros. La rilevazione trimestrale Oros su occupazione e costo del lavoro: indicatori e metodologie, Collana Istat Metodi Letture Statistiche, Roma. ISBN 978-88-458-1973-5 , 2019. URL https://www.istat.it/wp-content/uploads/2019/03/La-rilevazione-trimestrale-oros.pdf.

Istat-PPS. Flash Statistics, First quarter 2024 Service Producer Prices, 2024. URL https://www.istat.it/wp-content/uploads/2024/07/Service-producer-prices-Q1-2024.pdf.

X.L. Meng and D.B. Rubin. Maximum Likelihood Estimation via the ECM Algorithm: a General Framework. *Biometrika*, 80:267–278, 1993.

OECD/Eurostat. Eurostat-OECD Methodological Guide for Developing Producer Price Indices for Services: Second Edition, OECD Publishing, Paris, 2014. URL https://doi.org/10.1787/9789264220676-en.