Selective editing for the new Services Producer Price Indices (SPPIs) from indirect data sources

Diego Bellisai, Marco Lattanzio, Tiziana Pichiorri, Simona Rosati

Istat, Italian National Institute of Statistics

Unece Expert Meeting on Statistical Data Editing
7-9 October 2024, Vienna, Austria

Outline

SPPIs - Services Producer Price Indices (1/5)

- **Target variable**: the business-to-business (BtoB) production price, which is the quarterly average price of the service sold.
- **Data sources**: direct quarterly survey on enterprises (Oros) and administrative data (INPS, and ISA models).
- **NACE groups**:
  - 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.

SPPIs - Services Producer Price Indices (2/5)

Elementary chained-base indices:

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

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

SPPIs - Services Producer Price Indices (3/5)

Indices at the enterprise level: 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.

SPPIs - Services Producer Price Indices (4/5)

Aggregate indices (at the level of economic activity): a weighted mean of the enterprise indices:

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

the weighting coefficients $\omega_{i,0}$ are given by the share of revenues of the enterprise with respect to the total yearly revenues for a given offered service.

SPPIs - Services Producer Price Indices (5/5)

The aggregate indices can be re-written as:

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

This formula makes clearer the relationship with the hourly labour cost at the enterprise level and makes the use of SeleMix suitable. Indeed, $\boldsymbol{\pi_{i,0}}$ defines the **influence weight** to be used in the process of influential observations detection.

Outline

SeleMix - the method

- Approach based on **contamination normal models**.
- **Intermittent error mechanism**:
  - 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.
- The observed data distribution can be derived by combining two regression models:
  - true data distribution
  - the error mechanism

$\Longrightarrow$ it is possible to estimate the **magnitude of the error**; thus, to identify errors that have high impact on target estimates (influential errors).

SeleMix - R package

Three key functions:

1. ***ml.est*** $\longrightarrow$ the estimation of model parameters
2. ***pred.y*** $\longrightarrow$ the prediction of variable values
3. ***sel.edit*** $\longrightarrow$ the selection of observations affected by potential influential errors

An additional function ***sel.pairs*** provides valuable graphical tools.

The editing strategy (1/2)

Two distinct models of the type were specified:

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

where i=enterprise, K=economic activity.

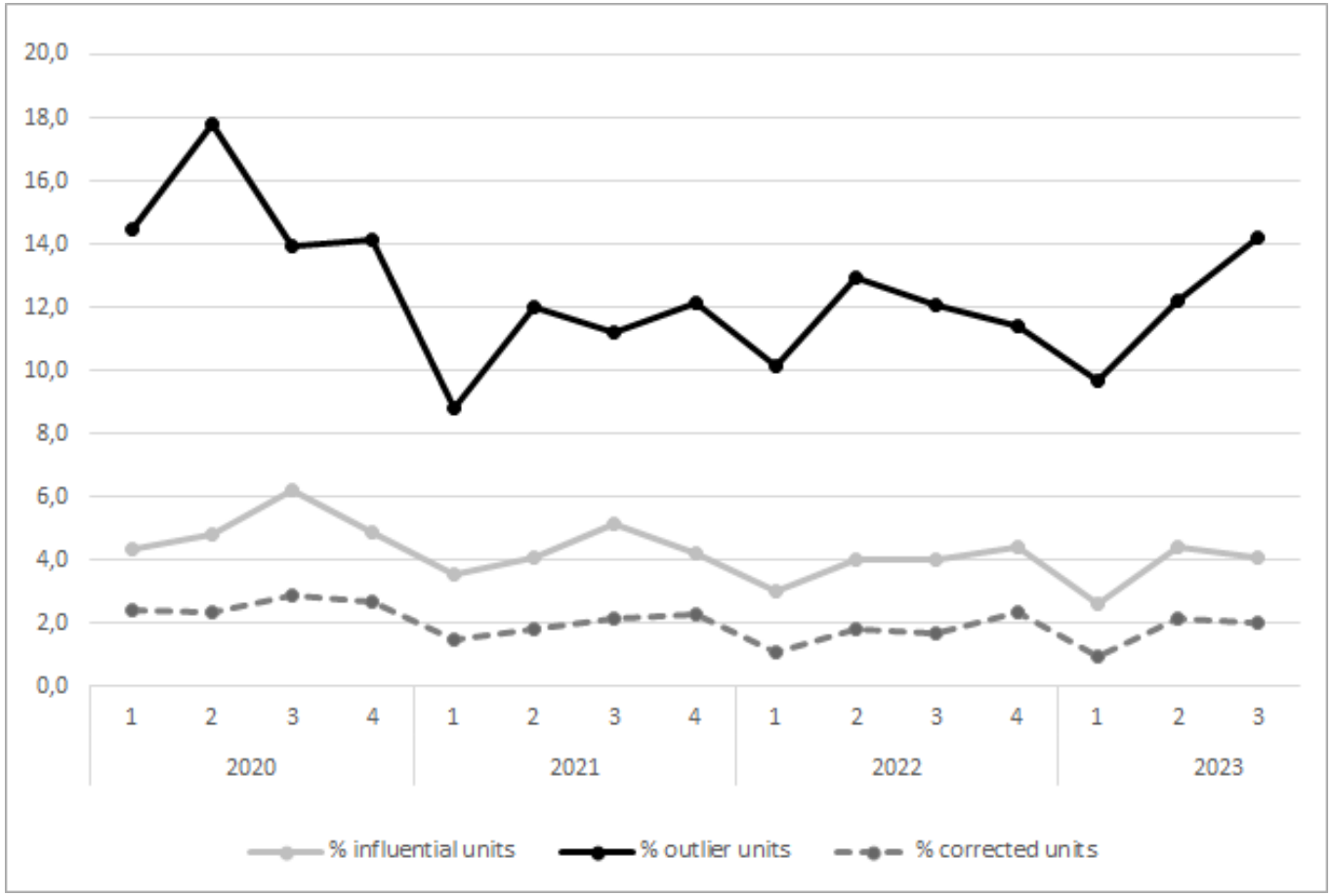
1. The **Model 1** with one covariate aims to identify influential units for the responding units (enterprises) for which the average labour cost at $t-4$ is available.
2. The **Model 2** with no covariates aims to identify influential units on all responding units.

The editing strategy (2/2)

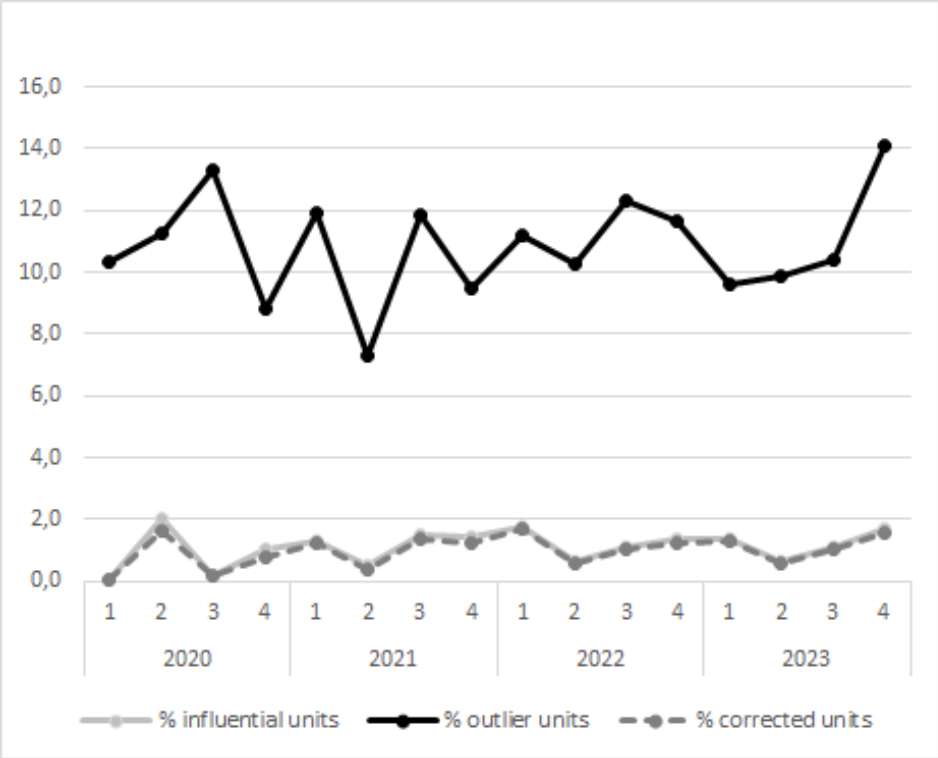
3. The entire data set was divided into two subsets based on the 99-th percentile of the distribution of the target variable ($clor_t$) and the share of per capita revenue.
4. Model 1 and Model 2 were applied to each of the two subsets, for $i = 1, 2, \ldots, n_K$.
5. The list of influential units is given by the union of the influential units identified by Model 1 and Model 2 (unless the influential units already identified by Model 1).

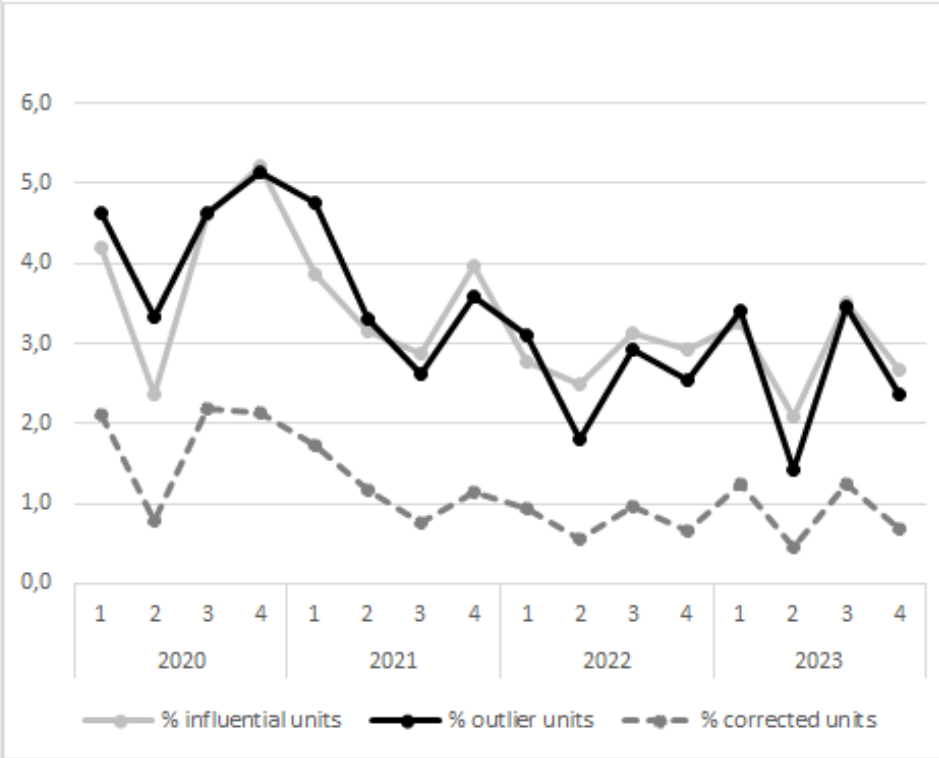
Outline

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

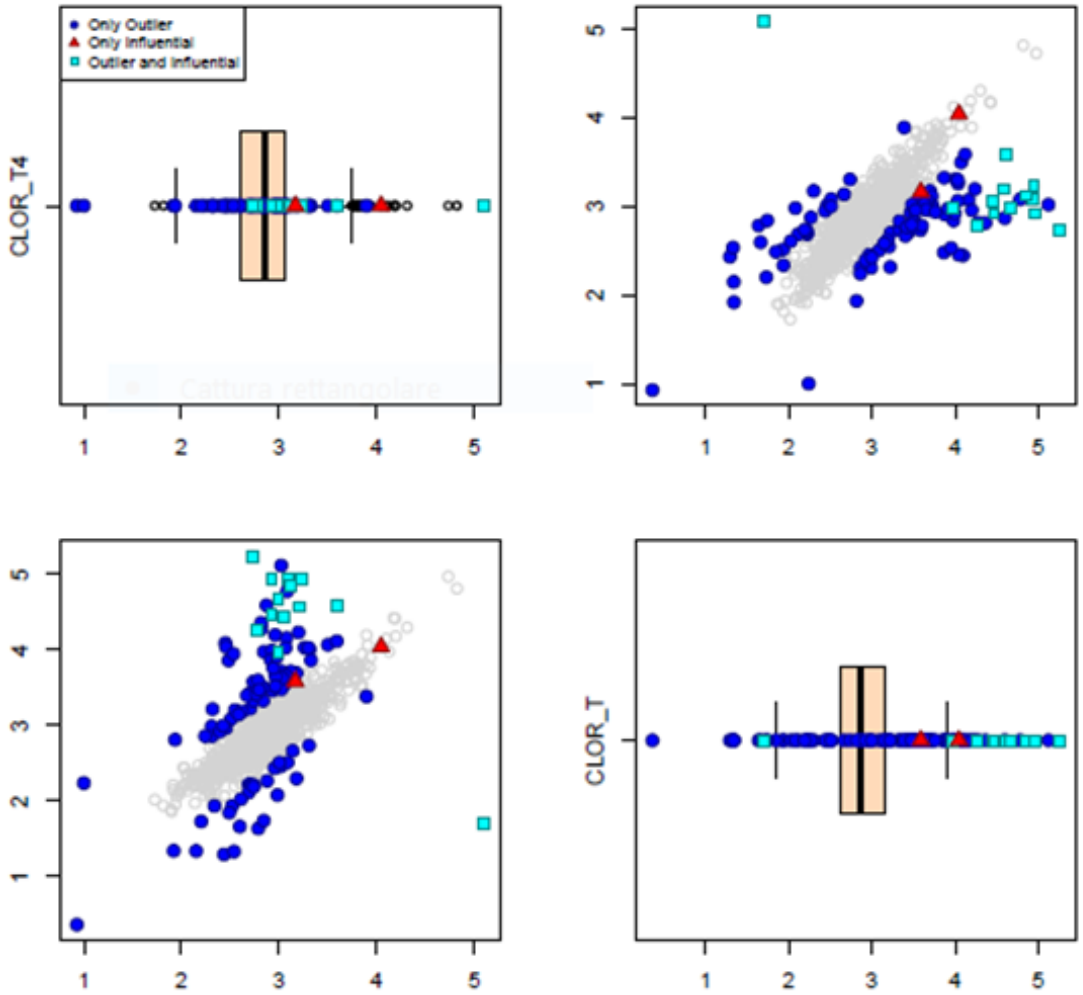


Percentage of outlier, influential and corrected units out of the total units, NACE groups as a whole - Model 1 and Model 2

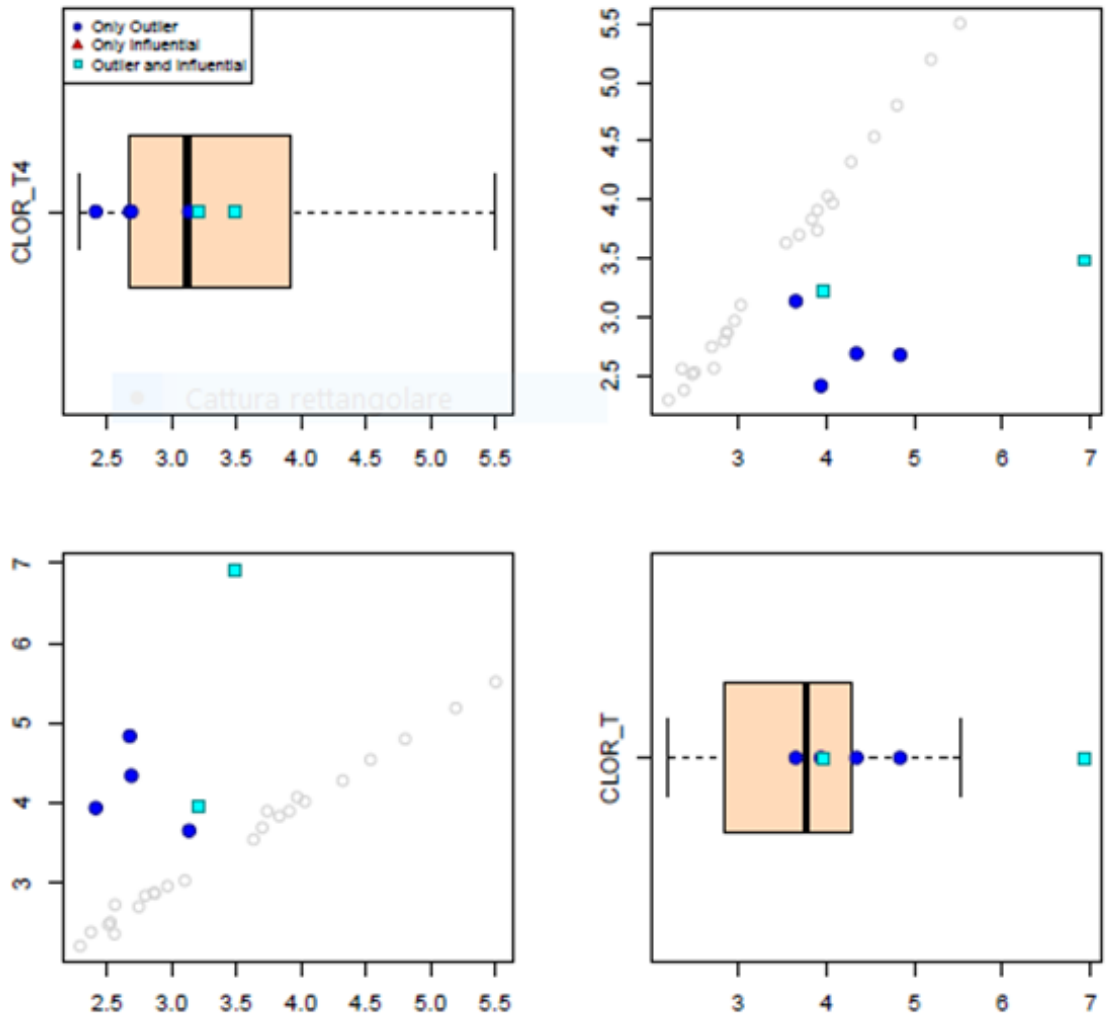




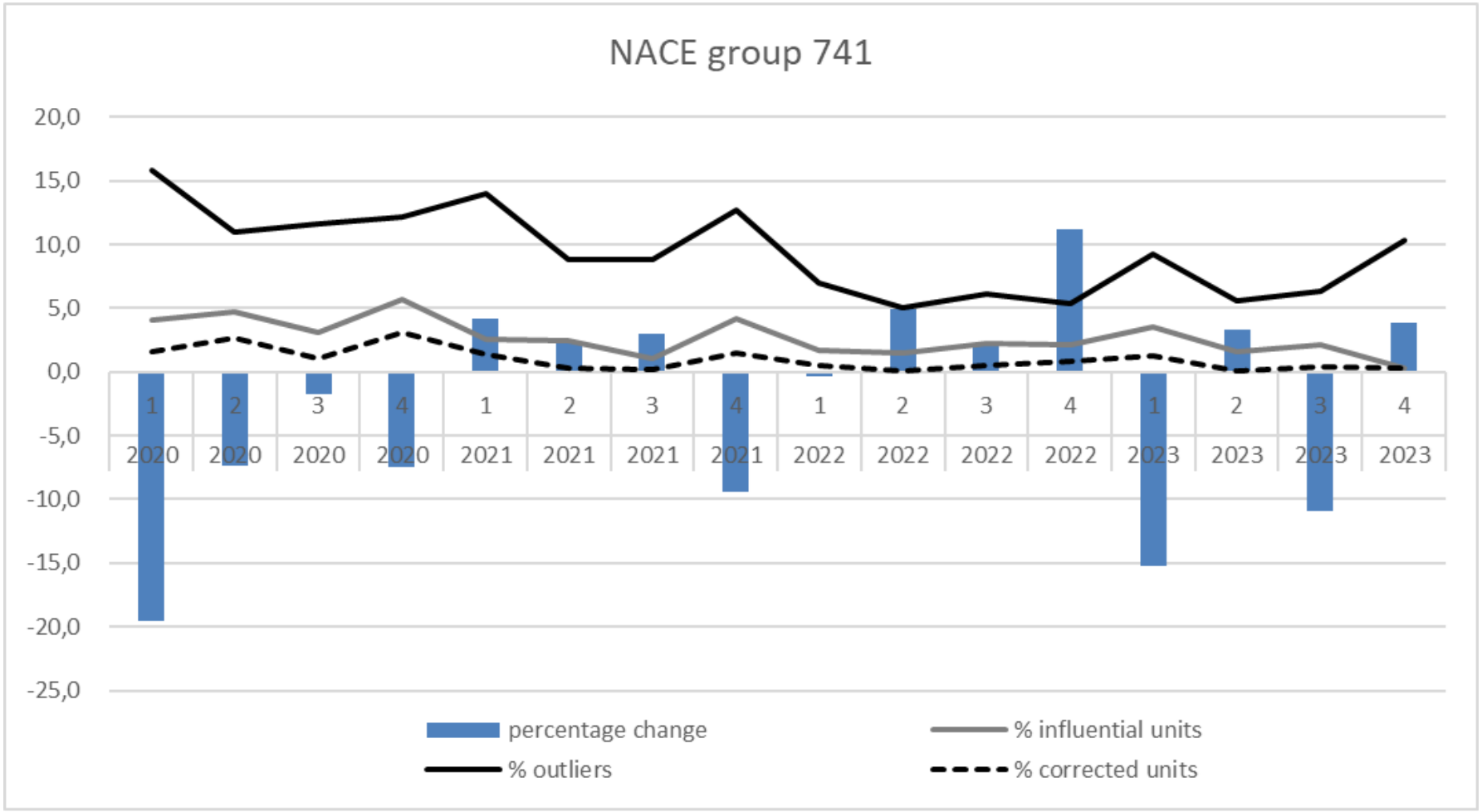
NACE group 741, Q1 2020, units below the 99-th percentile. Outliers and influential errors highlighted.



NACE group 741, Q1 2020, units exceeding the 99-th percentile. Outliers and influential errors highlighted.



NACE group 741: Percentage changes, and percentage of outliers, influential and corrected units out of the total



Outline

Conclusions and future work

- Different performance for Model 1 when focused on units above the 99-th percentile versus the same model on the rest of units: only by setting a high threshold were potential influential units identified.
  - Possible cause: strong relationship between the dependent variable and its covariate.
  - Possible solution: discarding highly similar observations between time (t) and $(t-4)$ may be better.
- Different methods should be found for subsets of units above the 99-th percentile for which SeleMix did not converge.
- All identified influential units should be manually reviewed by experts; it is reasonable to expect that fewer units will need to be corrected.

Questions, comments, suggestions?

sirosati@istat.it