Working Papers
03/2022

Early Estimates of the Industrial Turnover Index using Statistical Learning Algorithms

S. Barragán
L. Barreñada
J.F. Calatrava
J.C. Gálvez Sáenz de Cueto
J.M. Martín del Moral
E. Rosa-Pérez
D. Salgado

First draft: December 2022

This draft: December 2022

Early Estimates of the Industrial Turnover Index using Statistical Learning Algorithms

Abstract

We use statistical learning algorithms to improve timeliness of the Spanish Industrial Turnover Index. The main idea is to use a gradient boosting algorithm to make a prediction for every single industrial turnover value not yet collected during the data collection, data editing and estimation phases. Regressors are constructed from the historical unit-level time series, current aggregated turnover moments and quantiles, and aggregated values of related industrial surveys. Accuracy indicators are also computed so that a quantitative trade-o˙ between accuracy and timeliness can be appraised. This mass imputation exercise provides us with a nowcasting proposal which can be readily extended to many similar design-based surveys.

Authors and Affiliations

S. Barragán
S.G. for Methodology and Sampling Design, Statistics Spain (INE), Spain
Dept. Methodology and Development of Statistical Production, Statistics Spain (INE), Spain
L. Barreñada
BCAM - Basque Center for Applied Mathematics, Spain
EMOS Student Intern, Statistics Spain (INE)
J.F. Calatrava
S.G. for Statistical Dissemination and Communication, Statistics Spain (INE), Spain
Dept. Methodology and Development of Statistical Production, Statistics Spain (INE), Spain
J.C. Gálvez Sáenz de Cueto
S.G. for Information and Communications Technologies, Statistics Spain (INE), Spain
Dept. Methodology and Development of Statistical Production, Statistics Spain (INE), Spain
J.M. Martín del Moral
S.G. for Short-Term Statistics, Statistics Spain (INE), Spain
S.G. for Industrial and Services Statistics, Statistics Spain (INE), Spain
E. Rosa-Pérez
S.G. for Short-Term Statistics, Statistics Spain (INE), Spain
S.G. for Industrial and Services Statistics, Statistics Spain (INE), Spain
D. Salgado
S.G. for Methodology and Sampling Design, Statistics Spain (INE), Spain
Dept. Methodology and Development of Statistical Production, Statistics Spain (INE), Spain

INE.WP: 03/2022

Early Estimates of the Industrial Turnover Index using Statistical Learning Algorithms

S. Barragán[*1], L. Barreñada[†3], J.F. Calatrava[‡1], J.C. Gálvez Sáenz de Cueto[§1], J.M. Martín del Moral[¶2], E. Rosa-Pérez[‖2], and D. Salgado[**1]

[1]Dept. Methodology and Development of Statistical Production, Statistics Spain (INE), Spain
[2]S.G. for Industrial and Services Statistics, Statistics Spain (INE), Spain
[3]EMOS Student Intern, Statistics Spain (INE)

2022-07-01

Abstract

We use statistical learning algorithms to improve timeliness of the Spanish Industrial Turnover Index. The main idea is to use a gradient boosting algorithm to make a prediction for every single industrial turnover value not yet collected during the data collection, data editing and estimation phases. Regressors are constructed from the historical unit-level time series, current aggregated turnover moments and quantiles, and aggregated values of related industrial surveys. Accuracy indicators are also computed so that a quantitative trade-off between accuracy and timeliness can be appraised. This mass imputation exercise provides us with a nowcasting proposal which can be readily extended to many similar design-based surveys.

Contents

[*]Present affiliation: S.G. for Methodology and Sampling Design, Statistics Spain (INE), Spain.
[†]Present affiliation: BCAM - Basque Center for Applied Mathematics, Spain.
[‡]Present affiliation: S.G. for Statistical Dissemination and Communication, Statistics Spain (INE), Spain.
[§]Present affiliation: S.G. for Information and Communications Technologies, Statistics Spain (INE), Spain.
[¶]Present affiliation: S.G. for Short-Term Statistics, Statistics Spain (INE), Spain.
[‖]Present affiliation: S.G. for Short-Term Statistics, Statistics Spain (INE), Spain.
[**]Present affiliation: S.G. for Methodology and Sampling Design, Statistics Spain (INE), Spain.

1 Introduction

The modernization of the production of official statistics is rooted on three basic pillars, namely, (i) industrial standardization, (ii) new data sources, and (iii) new statistical methods. The industrial standardization comprises the adoption of international production standard models such as the Generic Statistical Business Process Model, GSBPM for short, (UNECE, 2019b), the Generic Statistical Information Model, GSIM for short, (UNECE, 2013), the Generic Statistical Activity Model for Statistical Offices, GAMSO for short, (UNECE, 2019a) and some others (see multiple references and projects by UNECE (2021)). The core aspect of this industrial standardization and these models is the adoption of a modular approach to statistical production and an enterprise architecture (see e.g. Eurostat (2021) for a European context). The incorporation of new data sources comprises the use of traditional survey data together with administrative registers and new digital data (Big Data included) of any nature. This incorporation entails a fairly large amount of issues to be solved (see e.g. Kitchin, 2015; Hand, 2018; Salgado and Oancea, 2020, and multiple references therein), which is intimately linked to the need of using new statistical methods and a refurbished production framework.

One of the potentially ensuing advantages of this modernization process is the generalized improvement of quality, understood in its multidimensional conception (see e.g. Wand and Wang, 1996; ESS, 2014). According to most appraisals, the huge availability of data is expected to remarkably improve the timeliness quality dimension, bringing press releases closer to the point in time when phenomena take place. This timeliness improvement is usually associated to new data sources, especially, new digital data sources. The initial

success of the example of the detection of influenza epidemics using search engine query data, i.e. of the so-called Google Flu Trends (Ginsberg et al., 2009) spread the feeling that digital data could bring immediacy to the production of statistics. However, the lack of accuracy was soon proved to lurk this kind of analysis (D. Lazer et al., 2014). The faintness of many of these correlations between digital data and the target variables in Official Statistics arises mostly due to the lack of statistical structural medatata (concepts and definitions) in all these new data sources.

In this working paper, we propose an early estimation of the Industrial Turnover Index (ITI), which is a monthly Short-Term Business Statistics produced by National Statistics Institutes (NSIs) in the European Statistical System (ESS) under Regulation (Eurostat, 2019, 2010). This early estimation of the ITI clearly improves the timeliness of this statistic using only the traditional survey data providing also an accuracy measure. The work is organised as follows. In section 2 we briefly discuss about the concepts of advanced indices, nowcasting and early estimates. In section 3 we describe the traditional production process of the ITI at Statistics Spain (INE). In section 4 we present the statistical methodology to compute the early estimates. In section 5 we describe the production process for the advanced index executed in this pilot study. In section 6 we present the results for the periods between January, 2016 to May, 2021. In section 7 we analyse the results. Finally, in section 8 we include some conclusions and remarks.

2 Early estimates and nowcasting

The provision of timely and accurate information is crucial to have knowledge about the present state of the economy and of society, in general (Broe et al., 2021). This is fundamental for a more efficient policy-making and decision-taking at all levels. In this sense, the subject *nowcasting/early estimates* has received much attention in the past years (see e.g. Giannone et al., 2013; Bok et al., 2017; Eurostat, 2017, and multiple references therein).

Terms to refer to this type of estimates abound in the literature: early estimates, rapid estimates, flash estimates, nowcasting, forecasting, leading indicators, coincident indicators, advance estimates,... with subtle differences between them (sometimes none). To put the current working paper into context we use the 8-dimensional description by Mazzi and Cannata (2017) providing details about each axis in their comprehensive account of this sort of estimates. See table 1 for details. It is important to underline that this is a pilot experience and these early estimates are not currently implemented in production.

This work is embedded in the context of the *production of official statistics*, not of the *use and analysis of official statistics* (let alone about *data management*). Thus, our attempt to improve timeliness is not to be understood as a resource to advanced statistical and econometric techniques, but as an effort to modernise the official statistical production system (Eurostat, 2017). Timeliness in this sort of short-term business statistics is hard to improve due to several factors. Firstly, since this is flow data, it is conceptually impossible to **measure** any indicator relative to a reference period until the reference period is over. Secondly, when the reference period is over and the clock starts ticking away, we face both external and internal restrictions. On the one hand, it takes some time to collect data from respondents with traditional data collection modes (physical/electronic questionnaires, telephone, email). With automatic data collection modes (e.g. with an automatic XML reporting), we could possibly accelerate the process, but integration with accounting systems for industrial establishments of all kind is a major issue. Furthermore, data may not be immediately ready for collection after the reference month ends because of entangled accounting processes within the industrial establishments themselves. On the other hand, the execution of statistical data editing strategies (see e.g. UNECE, 2019c) requires also some time consuming tasks, often leading to recontacts and follow-ups, especially in the error treatment tasks. It is important to underline that a minimal amount of interactive (manual) editing tasks is necessary to guarantee the accuracy and the quality of the data editing phase. Probably, improvements in the editing during collection (e.g. computing accurate individual data validation intervals and developing customized respondent/NSI interfaces) could possibly reduce post-collection editing activities. Finally, higher-frequency indicators could also possibly improve the utility of official statistics and bring them closer to user needs. However, so far official short-term business statistics only consider monthly

Axis	Description
Uniqueness of an official release versus the potential multiplicity of producers	These early estimates are to be produced by an NSI or equivalent office producing the official ITI.
Target variable	The early estimates of the ITIs are to be produced using the historic time series of same hard microdata and paradata used to produced the regular estimates. These hard microdata and paradata are also combined with related statistics such as the Industrial Production Index and the Industrial Price Index.
Revisions in the estimate	These estimates are intended to provide a reliable estimate of the first official release of the ITIs in parallel to the data collection and data editing production phases. Currently, they are considered as a pilot exercise before considering them as an experimental statistics.
Adherence to the regular production process	The methodology essentially amounts to jointly making a mass imputation exercise and predicting final validated values (measurement-error free) on not yet collected values. The computation of the early indices follows exactly the same methodology as the regular estimates.
The information set	The information set is constituted by an incomplete observation set for the reference period under estimation, including paradata for the same survey and aggregated data from the Industrial Production Index and Industrial Price Index surveys.
Model versus parameter uncertainty	Both the model and the parameters are chosen to compute the predictions of the target variable of each not yet collected value.
Appropriate release time	Early estimates are produced before the end of the reference period (a prediction, indeed) and after the reference period ends when each microdata batch is processed at the Central Office (currently, at $m+20d$, $m+29d$, and $m+38d$). The first official release is at $m+51d$.
Reporting and reference period	These are flow data, hence the coincidence between the reference and the report period.

Table 1: 8-dimensional description of the early estimation of the ITI (see Mazzi and Cannata, 2017, for definitions of the axes).

periods as reference periods in their legal regulations. More research is needed in this line (e.g. to propose weekly or daily indices).

In connection with the issue of timeliness, we believe that it is important to underline that NSIs should provide a **measurement** of the economy and the society, in general, through population aggregates, indices, and indicators. In another words, they should avoid, to the extent feasible, **predictions** or **estimates** based on implicit or explicit assumptions and judgements of statistical officers. In this line of thought, early estimates must be based on data-intensive algorithms, techniques, and methods with as few assumptions as possible.

This work presents a proposal with a pilot study to reduce the ITI provision delay by using statistical learning algorithms on the available data at times $m+\Delta_i < m+\Delta_{\text{release}}$, $i=1,2,3$, to reconstruct the values of the whole sample. There exist two fundamental ideas behind our approach:

1. This is a bottom-up approach so that early estimates of the ITI are produced by an

exercise of mass imputation of those missing values at each early time $m + \Delta_i$ at the **statistical unit level**. No prediction method is applied to any aggregate or index. Aggregates and indices emerge from the imputed sample, as in their final validated released versions.

2. Despite the fact of focusing on statistical units for the construction of aggregates and indices, only aggregate regressors of the reference time period are used together with regressors from past reference time periods at the statistical unit level.

The combination of these two decisions invites us to express the prediction exercise as a pattern recognition of individual microdata (the turnover) against both present aggregated values and past individual values. It is, thus, a predictive exercise, where predictions are also provided with a measure of uncertainty.

By dropping out regressors for the present reference time period, we can also compute predictions before any data from the reference month is collected. This helps us assess the importance of using current data from the reference time period in providing reliable official statistics.

We shall use the term *early estimate* of the ITI making reference to table 2 as we progress in the different production steps to reach the final value of the index.

3 The statistical production process of the ITI

The ITI survey is a European short-term business statistics produced by the ESS in compliance with the Regulation (EU) no. 2019/2152 and Comision Implementing Regulation 2020/1197, (Eurostat, 2019, 2010). The ITIs have the objective of measuring the evolution of the activity of establishments included in the industrial sector through their turnover. Thus, the core target variable is the *industrial turnover*, i.e. "the value of the invoicing of the establishment, in the reference month, for the sales of industrial goods and provision of industrial services, considering both those carried out by the establishment itself, and those performed through subcontracting with third parties. It therefore includes the income from the sales of finished products, of semi-finished products, of subproducts, of waste and recovered materials, of packages and packaging and of merchandise (goods acquired for resale in the same state as that in which they were acquired), as well as the income from the provision of services related to the normal activity of the establishment" (see INE, 2018, for details).

The ITI survey comprises all industrial establishments whose main economic activity is included in sections B "Extractive industries" (except division 09 in the case of Spain) and C "Manufacturing industry" of the Spanish National Classification of Economic Activities (CNAE-2009), adapted from the international NACE Rev. 2. The survey has monthly periodicity and provides data at national and NUTS2 geographical levels (not including Ceuta and Melilla). The population frame is built from the Industrial Products Survey[1] and the Structural Business Statistics (Industrial Sector) [2]. Sampling units are selected according to a cut-off sampling design with a sample size of around 12000 units. The sample is revised yearly. The indices follow a fixed-base Laspeyres formula in two steps:

- Computation of the elementary indices.- The sample of units is partitioned into strata determined by their NUTS2 geographical variable and some groupings of CNAE-2009 economy activity codes (divisions and groupings of sections) (INE, 2018). The elementary index for stratum U_d at reference month m of year y with base year y_B (currently 2015) is then computed recursively as

$$ {}_{y_B}I_{U_d}^{my} = \frac{\sum_{k\in s_d^{my}\cap s_d^{m-1y}} z_k^{my}}{\sum_{k\in s_d^{my}\cap s_d^{m-1y}} z_k^{m-1y}} \times \cdots \times \frac{\sum_{k\in s_d^{2y_B}\cap s_d^{1y_B}} z_k^{2y_B}}{\sum_{k\in s_d^{2y_B}\cap s_d^{1y_B}} z_k^{1y_B}} \times \frac{\sum_{k\in s_d^{1y_B}} z_k^{1y_B}}{\frac{1}{12}\sum_{m=1}^{12}\sum_{k\in s_d^{my_B}\cap s_d^{m-1y_B}} z_k^{my_B}} \times 100, \tag{1} $$

where s_d^{my} denotes the sample for the stratum U_d at reference month m of year y (by convention $s_d^{-1y_B} = U_d$) and z_k^{my} stands for the target variable value of establishment k at reference month m of year y (the total turnover of the industrial establishment).

[1] https://www.ine.es/dyngs/IOE/en/operacion.htm?id=1259931057259.
[2] https://www.ine.es/dyngs/IOE/en/operacion.htm?id=1259931057090.

- Computation of composite indices.- After computing the weight $w_d^{y_B}$ of each stratum U_d for the base period y_B using data from the Structural Business Statistics (Industrial Sector), we can compute the composite index for a functional aggregate $U_A = \bigcup_{d \in A} U_d$ just as a weighted arithmetic mean of elementary indices:

$$ {}_{y_B}I_{U_A}^{my} = \sum_{d \in A} w_d^{y_B} \times {}_{y_B}I_{U_d}^{my} \tag{2}$$

The main steps in the production pipeline in relation with our pilot study (see INE, 2018, for more details) are:

- Data collection for reference month m starts at $m + 1d$ (the immediate day after the reference period ends).
- Data editing during collection takes place all along the collection period by Statistics Spain's provincial delegations.
- Data are processed at Statistics Spain's central office by the survey managers in three data batches constituted at $m + 20d$, $m + 29d$, and $m + 37d$ from the provincial delegations.
- Post-collection data editing is conducted upon these batches.
- Computation of final indices and variation rates is conducted in the last week prior to the press release date.
- Press release takes place at $m + 51d$.

The proposed early estimates of the indices aim at producing the same output as the press release as soon as data are available for processing at Statistics Spain's central office (with the exception of the breakdown per market and the seasonal and calendar-adjusted indices).

4 Statistical methodology

4.1 Design-based and model-based inference

As stated in the preceding section, the sample is selected according to a cut-off sampling design with cut-off values set every year. At all times the cut-off sample is selected for the whole year by determining the cut-off thresholds from external data, so that we can always write $s_d^{my} = U_{c,d}^{my}$, showing the dependence on the cut-off values (subscript c). The sample can change from one month to another depending on new units appearing in the industry market and/or units changing or ceasing their industrial activity. At time t, a concrete subsample of respondents has provided their responses so that we write $r_d(t) \subset s_d(t)$, where we have dropped out the superscript my for ease of notation (time is already implicitly denoted by t). Notice that this sampling design is not probabilistic and provides zero design-based variance for the traditional estimator $\widehat{Z}_{U_d}^{my} = \sum_{k \in s_d^{my}} \frac{z_k^{my}}{\pi_k^{my}} = \sum_{k \in U_c^{my}} z_k^{my}$.

4.2 Estimates and estimators of population totals

4.2.1 Target variable: edited and validated values

The core computation in the series of indices is the population total of the industrial turnover, i.e. $Z_{U_d}^{my} = \sum_{k \in s_d} z_k^{my}$. As described above, editing tasks are carried out from the data collection activity itself to the final estimation phase. Indeed, as a consequence of these editing tasks (recontacts and follow-ups), the value z_k^{my} of a given industrial establishment k can change several times during this editing phase. We shall denote by $z_k^{my,\mathrm{val}}$ the final validated value of variable z for unit k entering into the computation of the first official release of the ITI for reference month m and year y. Similarly, we shall denote by $z_k^{my,\mathrm{ed}}(t)$ the value of variable z for unit k at the time t of the editing strategy for the reference month m and year y. If t_{release} is the number of days after the reference month ends, under this notation we have $z_k^{my,\mathrm{ed}}(t_{\mathrm{release}}) = z_k^{my,\mathrm{val}}$.

According to the execution of the collection and editing processes, the value of the turnover z_k^{my} for a given unit k may change in the three different batches processed at the central office depending on the result of the validating and error treatment activities.

4.2.2 Estimators

In the traditional production process we estimate the population total $Z_{U_d}^{my}$ for each reference month m and year y after collecting and editing the whole sample as

$$\widehat{Z}_{U_d} = Z_{U_{c,d}} = \sum_{k \in U_{c,d}} z_k^{\text{val}}, \tag{3a}$$

where we have dropped out the reference time period dependence for ease of notation (every month the process is similarly repeated). This can be decomposed at any time t as

$$Z_{U_{c,d}} = \sum_{k \in r_d(t)} z_k^{\text{val}} + \sum_{k \in U_{c,d} - r_d(t)} z_k^{\text{val}}. \tag{3b}$$

Notice that this decomposition can only be actually computed after finishing these two production phases (collection and editing), since we need the final validated values z_k^{val}. The goal is not to wait until all data collection and all data editing are both concluded to produce an early estimation of the ITI with the ongoing collected and edited information. Taking into account the values we already know and predicting what we do not know yet, we decompose this estimate as follows:

$$Z_{U_{c,d}} = \sum_{k \in r_d(t)} \left[z_k^{\text{ed}}(t) - e_k^{\text{meas}}(t) \right] + \sum_{k \in U_{c,d} - r_d(t)} \left[\widehat{z}_k^{\text{val}}(t) - e_k^{\text{pred}}(t) \right], \tag{3c}$$

where $e_k^{\text{meas}}(t)$ denotes the measurement error $e_k^{\text{meas}}(t) = z_k^{\text{ed}}(t) - z_k^{\text{val}}$ and $e_k^{\text{pred}}(t)$ denotes the prediction error $e_k^{\text{pred}}(t) = \widehat{z}_k^{\text{val}}(t) - z_k^{\text{val}}$.

Notice that so far these all are real numbers and there are some of them still unknown (the predicted values and the measurement and prediction errors). Based on these decompositions of the population total, we shall propose estimators taken into account the information we have at time t. The key concept is to take into account the distinction between random variables and their realizations according to the time instant in which we are computing the estimate. We propose two estimators:

4.2.3 Estimator 1

A first estimator for the population total $Z_{U_d}(t)$ with data collected up to time t is given by

$$\widehat{Z}_{U_d}(t) = \sum_{k \in r_d(t)} z_k^{\text{ed}}(t) + \sum_{k \in U_{c,d} - r_d(t)} \widehat{Z}_k^{\text{val},\xi_p}(t), \tag{4a}$$

where $\widehat{Z}_k^{\text{val},\xi_p}(t)$ is the random variable representing the prediction for value $\widehat{z}_k^{\text{val}}(t)$ according to prediction model ξ_p. This estimator, when applied at time t, produces estimates of the form:

$$z_{U_d}(t) = \sum_{k \in r_d(t)} z_k^{\text{ed}}(t) + \sum_{k \in U_{c,d} - r_d(t)} \widehat{z}_k^{\text{val},\xi_p}(t). \tag{4b}$$

Notice that, when compared with decomposition (3c), this amounts to neglecting measurement errors $e_k^{\text{meas}}(t)$ and considering $e_k^{\text{pred}}(t) \approx 0$. In this way, we just need to build only one prediction model ξ_p.

4.2.4 Estimator 2

A second estimator for the population total Z_{U_d} with sample data collected up to time t is given by

$$\widehat{Z}_{U_d}(t) = \sum_{k \in r_d(t)} \left[z_k^{\text{ed}}(t) - \widehat{E}_k^{\xi_m}(t) \right] + \sum_{k \in U_{c,d} - r_d(t)} \left[\widehat{Z}_k^{\text{val},\xi_p}(t) - \widehat{E}_k^{\xi_e}(t) \right], \tag{5a}$$

where ξ_p stands for a prediction model for z_k^{val}, ξ_m stands for a measurement error model for $e_k^{\text{meas}}(t)$ and ξ_e denotes a model for the prediction error $e_k^{\text{pred}}(t)$. This estimator, when applied at time t, produces estimates of the form:

$$z_{U_d}(t) = \sum_{k \in r_d(t)} \left[z_k^{\text{ed}}(t) - \widehat{e}_k^{\xi_m}(t) \right] + \sum_{k \in U_{c,d} - r_d(t)} \left[\widehat{z}_k^{\text{val},\xi_p} - \widehat{e}_k^{\xi_e}(t) \right]. \tag{5b}$$

Notice that for the estimator 2 we need now three models ξ_p, ξ_m, and ξ_e. For simplicity's sake, so far we have only explored the first option.

4.2.5 Remarks

Notice that the estimators are proposed having in mind that the ITI survey follows a cut-off sampling design. This is important because the generalization to other sampling designs needs a modification, e.g. for the HT estimator, for the ratio estimator, for the GREG estimator, or for the Sanguiao-Zhang estimator (Sanguiao and Zhang, 2021), where in all these cases a purely probabilistic sampling design is in place.

The other important difference with the Sanguiao-Zhang estimator (which generalizes all the rest) is that since the subsample $r_d(t)$ is not selected by the survey manager and we do not have data from $U_{c,d} - r_d(t)$ (because of the definition itself of $r_d(t)$), we cannot make use of the bias-correcting term proposed by Sanguiao and Zhang (2021).

Let us see the extreme case in which all data have been collected from the complete selected sample $s = U_c$. Estimator and estimate (4) reduce, respectively, to

$$\widehat{Z}_{U_d}(t) = \sum_{k \in U_{c,d}} Z_{kt}^{\text{val}}, \tag{6a}$$

$$z_{U_d}(t) = \sum_{k \in U_{c,d}} z_{kt}^{\text{val}}, \tag{6b}$$

as expected, thus boiling down to the traditional estimator and estimate.

When only part of the sample data has been collected, estimators (4) and (5) compute the estimate for Z_{U_d} by using the collected values z_{kt} for units $k \in r_d(t)$ and by using predicted values $\widehat{z}_k^{\text{val}}(t)$ for the rest of units $k \in U_{c,d} - r_d(t)$ not yet collected, predicted with a model trained with collected values up to time t and using complementary models and variables to estimate both measurement and prediction errors.

Notice that we can provide an alternative estimator closer to the original Sanguiao-Zhang estimator (except for the bias-correcting term):

$$\widehat{Z}_{U_d}(t) = \sum_{k \in r_d(t)} z_k^{\text{ed}} + \sum_{k \in U_d - r_d(t)} \widehat{Z}_k + (\text{Bias-Correcting Terms}), \tag{7}$$

where the fraction of the target population below the cut-off values $k \in U_d - U_{c,d}$ is also taken into account in the predictive part. This would imply that we had the same auxiliary information (i.e. the model regressors) for these units to build the prediction model, but currently this is not the case. Thus, we shall concentrate on estimator (4).

4.3 Regressors

We detail the construction of each regressor x_k used in the prediction model ξ_p. All these regressors $\{x_k^{(p)}\}_{p=1,2,\ldots,P}$ have been constructed computing on survey microdata and/or paradata from the ITI survey itself, with the exception of some aggregates from the Industrial Price Index and Industrial Production Index surveys (see below). Neither administrative data nor a new digital data source has been used at all. This is an important remark, since statistical offices do not need access to new data sources to start using new statistical methods for their production.

4.3.1 Metadata of regressors

We provide a minimal metadata set for the regressors in the appendix. These metadata are specified by attribute-value pairs providing the following information:

1. **Definition**, which provides a verbal definition of the regressor.
2. **Statistical Type**, which denotes whether the regressor is *categorical* or *numerical*.
3. **Values**, which denotes the range of possible values of the regressor.
4. **Example**, which provides an example of the value of the regressor.
5. **Source**, which denotes whether the regressor is a *primary* variable or a *derived* variable from either the survey itself (*internal*) or other data source (*external*).
6. **Formula**, which provides details for the derivation of the regressor with a mathematical formula or algorithmic procedure, in general.
7. **Statistical Program Reference**, which provides reference information for the statistical program(s) providing the data. We use *Spanish-IOE* for the Spanish Inventory of Statistical Programs (acronym in Spanish).
8. **Unit/Aggr**, which denotes whether the regressor value is computed using information from an individual statistical *unit* or a group of statistical units (thus, being an *aggr*egate).
9. **Time Periods**, which denotes the reference time periods involved in the computation of the regressor with the following relative notation: 0 denotes the current reference time period, $-i$ denotes the ith period before the current reference period.
10. **Long/Cross**, which indicates whether the regressor value is computed using only *long*itudinal information (time series), only *cross*-sectional information (e.g. an arithmetic mean by NACE group over a given time period), or both (*long + cross*).
11. **Cross-Domain Vars**, which denotes the name of the variables defining the population domains over which the cross-sectional computations are carried out.
12. **Encoding**, which indicates the type of encoding for categorical variables (*one-hot*, *dummy*, *mean*, etc.).

Finally, we classify them also in tables in terms of their semantic content: (a) geographical variables, (b) time variables, (c) economic activity variables, (d) target-related variables, (e) external survey variables, and (f) cross-categories therein. See the appendix for a detailed list of regressors.

4.3.2 Construction of regressors

Primary regressors are taken directly from the collection, editing, and estimation paradata, microdata, and aggregates of each survey. In turn, derived regressors are computed for each batch and each reference time period following their metadata specifications for all units $k \in U_c$ in the cut-off population.

We underline the fundamental idea that regressors are chosen so that the prediction model ξ_p can be applied to units both in $r(t)$ and in $U_c - r(t)$, thus discarding regressors at the unit level for the current reference time period. Instead, only aggregated quantities such as quantiles, means, standard deviations, etc. are used as regressors with reference to this time period.

Notice also that we can build a preliminary prediction model $\xi_{\bar{p}}$ in which regressors from the current reference time period are not included. This will allow us to predict turnover values $\hat{z}_k^{\xi_{\bar{p}}}$ before any data from the current reference time period is collected (a genuine prediction exercise). Thus, we shall be able to assess the relevance of data from the current reference time period, even from a fraction of the sample, to produce a reliable early estimate.

As stated, there exist two types of units, namely, those included in the corresponding batch $k \in r(t)$ and those not $k \in U_c - r(t)$. For those regressors whose computation involves data from the current reference time period (i.e. value 0 is contained in the regressor metadata attribute named Time Periods), this computation cannot be achieved for the latter units $k \in U_c - r(t)$, i.e. we have not yet collected values. These regressors are aggregated variables, so that we can straightforwardly construct these by the corresponding value for the cross-domain variable(s) specified in the metadata. For example, the regressor `mean_trnovr_ed_NACE2div` is computed as the mean of the edited turnover values by NACE Rev.2 division using only units in the sample $r(t)$ of respondents at time t (complete-cases

analysis according to Little and Rubin (2002)):

$$\bar{z}_d = \frac{1}{N_{r_d(t)}} \sum_{k \in r_d(t)} z_k^{\text{ed}}, \qquad d \in \text{NACE2div.}$$

For those units $k \in U_{c,d} - r_d(t)$ (thus not yet with a value for the edited turnover), the value is constructed as $x_k(t) = \bar{z}_d(t)$, which is the value for the corresponding domain $U_{c,d}$. Note that when we use NACE Rev. 2 refers here to the Spanish National Classification of Economic Activities (CNAE-2009), adapted from the international NACE Rev. 2.

4.3.3 Treatment of missing values

We resort to the nested structure of official statistical classifications to propose the general imputing rule for any missing value in numerical regressors: any missing value in a numerical regressor is imputed by the mean of the same regressor in the immediately hierarchically superior category. For example, if regressor `mean_cnae3_est` (mean of establishment turnover by NACE Rev. 2 group) is missing, we impute it with the value `mean_cnae2_est` (mean of establishment turnover by NACE Rev. 2 division).

This general rule is complemented with a similar rule for regressors computed with past values. For example, missing values in the regressor `quantile_MA12subdiv_3` (value of the 12-month moving average ecdf $F^*_{\text{MA12subdiv}}$ at validated value $z_k^{m-3y,\text{val}}$) are imputed by `quantile_MA12subdiv_1` (the immediately most recent computed value). This is also applied to geographical categories. This general rule is indeed applied in real production conditions.

In remaining highly unlikely missing values in categorical variables, they would be treated as a category itself.

4.3.4 Treatment of outliers

It is important to keep in mind that industrial turnover data have non-negligible representative outliers (Chambers, 1986). These are readily explained by the standard industrial activity of some notably large firms. As a matter of fact, there exist industrial establishments with a strong influence on both the released indices and annual variation rates. It is essential, though not easy, to be able to predict these values with some degree of accuracy. To this end, for the time being we shall not process them in a special way but we shall use this fact as an orientation to make an adequate choice of regressors and the prediction model.

Regarding the regressors, we have defined variables reflecting the outlying behaviour of turnover values. In particular, we have defined regressors such as 95th percentiles, indicator values thereof, maximum values per domain, values of empirical cumulative distribution functions, etc. (see appendix). In this way, the outlying behaviour in the target variable values in preceding time periods is also contained in multiple regressors.

Regarding the prediction model, as we shall see in the next section, we have chosen a gradient boosting algorithm, which makes intensive use of residuals, so that outliers will be taken into account from the onset in the model building algorithm.

4.4 The prediction model

4.4.1 The model

Nowadays there exist multiple choices to compute predicted values for imputation in a data-rich environment from simple polynomial regression models to neural networks over random forests and boosted regression trees, to name a few. There already exist excellent initiatives to compare different imputation methods based on diverse statistical learning algorithms to assess their performance (Dagdoug et al., 2021) and a definitive recommendation about the models cannot be provided yet.

In our case, we have prioritized the construction of an end-to-end prototyping production process with every single aspect needed for a further implementation in realistic conditions, leaving for later improvements the optimization and choice of the model. However, we have taken different factors into account. Firstly, we pursue versatility so that any type of regressor can be used in the model. Furthermore, non-linear dependence between the

target variable and the regressors must be allowed. Thus, our first choice has been to use Random Forests (see e.g. Hastie et al., 2009; Murphy, 2013). Secondly, in order to improve accuracy and, especially, to deal with outliers which will be rapidly detected through their residuals, we have chosen to use boosting (see e.g. Watt et al., 2020). In this way, the trained boosted regression tree will naturally incorporate the effect of outliers, which will also be predicted with a reasonable accuracy. Finally, among the different choices within the boosting algorithm family we focus on the gradient boosting algorithm (Friedman, 2001) and, in particular, on the LightGBM version (Ke et al., 2017) (see also Microsoft Corporation, 2022). This choice is basically motivated by speed without a compromise of accuracy (see Bentéjac et al., 2021, for a comparison of gradient boosting algorithms). We have used the R API in the form of the R package `lightgbm` (Shi et al., 2021). Parameters and hyperparameters are specified below.

4.4.2 Learning scenario: train, validation, test

It is mandatory to describe the learning scenario in which our prediction model is built and used. By and large, depending on data availability there exist multiple scenarios (supervised learning, unsupervised learning, semi-supervised learning, transductive inference, online learning, reinforcement learning, ...) (see e.g. Mohri et al., 2018). In our case, the calendar time and the monthly periodicity plays a central role in data availability and data use.

To predict turnover values at times t_i in reference month m and year y, we must be aware that validated values from the preceding reference month $m-1$ are already available at these time instants. Thus, we follow for each and every month m the following procedure:

- We train the model for a set of multiple alternative hyperparameter sets $h = 1, 2, \ldots$ with data up to reference month $m-2$.
- We apply each trained model to the data set with corresponding reference month $m-1$ obtaining, thus, the predicted values $\widehat{z}_k^{m-1y}(t)$ for each unit k.
- We compute for each trained model ξ_h the absolute error of the total turnover $AE_h = \left| \sum_{k \in U_c^{my}} \widehat{z}_k^{m-1y,\xi_h}(t) - \sum_{k \in U_c^{my}} z_k^{m-1y,\text{val}}(t) \right|$. We select the model m_{h^*} with optimal value of AE_h.
- We train again the same model with data up to reference month $m-1$ with hyperparameters h^*.
- We apply the trained model to data collected up to time t of reference month m and year y. Thus, we obtain the predicted values $\widehat{y}_k^{my,\xi_{h^*}}(t)$ to be plugged in the estimator (4).

A cautious reader may point out that test data sets are not actually used, only train and validation sets. Firstly, in the pilot study we simulate with real data under the same operational procedure that we would be following in usual production conditions. The role of the test data set to assess performance is played by the data set at running time t every simulated month, where the performance will indeed be assessed by direct comparison with actually released index values. Secondly, since we repeat the same procedure every single month over a sequence of over 60 months incorporating successively new data, we can be sure that no overfitting is detected.

4.4.3 Hyperparameters and model selection

As in many other statistical learning algorithms, a set of hyperparameters must be tuned in advance to train the model to be used for prediction. A priori, the optimal values of these hyperparameters have not been exhaustively searched in this pilot experience, since we have prioritised the construction of an end-to-end prototyping process, which can be improved later on and be continuously evolving according to changing production conditions.

We have considered a minimal hyperparameter grid for two of the multiple parameters of the algorithm (see table 2), namely `nrounds` (the number of training rounds) and `eta` (the shrinkage rate in the gradient boosting algorithm).

Three more hyperparameters are customised in the core function `lightgbm` according to the following values to train the models:

The rest of parameters takes on their default values (see Microsoft Corporation, 2022, for details). Missing values treatment and encoding are not part of hyperparameter optimization. Needless to say, our choices are clearly suboptimal and a more exhaustive search should be

nrounds	eta
300	0.05
1000	0.05
300	0.01
1000	0.01

Table 2: Minimal hyperparameter grid search

Parameter	Definition	Value
'objective'	Type of regression application	'regression'
'metric'	Metric to be applied on the evaluation set(s)	'mae' (absolute loss)
'boosting'	Algorithm variant	'gbdt' (traditional Gradient Boosting Decision Tree)

Table 3: Customised hyperparameters for function `lightgbm`

accomplished to find the best combination of hyperparameters, missing values treatment and encoding, especially those regarding a trade-off between accuracy and speed (see Parameter Tuning section in Microsoft Corporation, 2022). Despite this, results prove that an evolving end-to-end process can be effectively designed and implemented to provide reasonably accurate early estimates.

4.5 Accuracy and model assessment

As stated above, we assess the performance of the prediction model by direct comparison between each predicted set of indices for the complete range of breakdowns published in the press release (see section 7) with their truly released versions under the traditional production process with the whole data collection and data editing phases fully accomplished.

Thus, if we denote by $\widehat{I}_{U_A}^{my}(t)$ the predicted ITI for domain U_A and reference period my at day t, we compute $\widehat{I}_{U_A}^{my}(t) - I_{U_A}^{my}$. We proceed similarly with the monthly and annual rates:

$$\widehat{\Delta}_{m,U_A}^{my}(t) = \frac{\widehat{I}_{U_A}^{my}(t) - \widehat{I}_{U_A}^{m-1y}(t)}{\widehat{I}_{U_A}^{m-1y}(t)}, \qquad \widehat{\Delta}_{y,U_A}^{my}(t) = \frac{\widehat{I}_{U_A}^{my}(t) - \widehat{I}_{U_A}^{my-1}(t)}{\widehat{I}_{U_A}^{my-1}(t)}.$$

Besides this empirical approach, it is highly convenient to make a theoretical analysis of accuracy to understand those factors impinging on it. We detach this analysis on the properties of bias and mean squared error of the proposed estimator.

4.5.1 Bias

We distinguish between the conditional bias $\mathbb{E}\left[\widehat{Y}_U(t) - Y_U(t) | S = U_c, R(t) = r_d(t), \mathbf{X}(t) = \mathbf{x}(t)\right]$ and the unconditional bias $\mathbb{E}\left[\widehat{Y}_U(t) - Y_U(t)\right]$. The former will provide the bias for the concrete cut-off sample U_c, concrete subsample $r_d(t)$ collected up to time t, and specific regressor values $\mathbf{x}(t) = \{\mathbf{x}_k(t)\}_{k \in r_d(t)}$ for all units k collected up to time t whereas the latter will provide the bias for any cut-off sample s, any subsample $r_d(t)$ collected up to time t (in the usual meaning of bias in probabilistic sampling designs), and any regressor values $\mathbf{x}(t)$.

The former is more meaningful for production and early release purposes, thus we shall focus on it. For estimator (4), we have

$$\begin{aligned}\mathbb{B}(\widehat{Y}_{U_d}(t)\big|U_{d,c}, r_d(t), \mathbf{x}(t)) &= \mathbb{E}\left[\widehat{Y}_{U_d}(t) - Y_{U_d}(t)\big|U_{d,c}, r_d(t), \mathbf{x}(t)\right] \\ &= \sum_{k\in r_d(t)} \mathbb{E}\left[E_k^{(m)}(t)\big|U_{d,c}, r_d(t), \mathbf{x}(t)\right] + \sum_{k\in U_{d,c}-r_d(t)} \mathbb{E}\left[E_k^{(p)}(t)\big|U_{d,c}, r_d(t), \mathbf{x}(t)\right] \\ &= \sum_{k\in U_{d,c}} \mathbb{E}\left[\tilde{E}_k(t)\big|U_{d,c}, r_d(t), \mathbf{x}(t)\right], \end{aligned} \tag{8a}$$

where $E_k^{(m)}(t) = Z_k^{\text{ed}}(t) - Z_k^{\text{val}}$ and $E_k^{(p)} = \widehat{Z}_k(t) - Z_k^{\text{val}}$ denote the random variables for the measurement error and the prediction error for model ξ_p, respectively, and where we have defined

$$\tilde{E}_k(t) = \begin{cases} E_k^{(m)}(t) & \text{if } k \in r_d(t), \\ E_k^{(p)}(t) & \text{if } k \in U_{d,c} - r_d(t). \end{cases} \tag{8b}$$

The expression (8a) for the conditional bias of estimator (4) basically states that there exist two contributions, namely (i) the sum of measurement errors for those units whose data have already been collected at time t and (ii) the sum of prediction errors for those units whose data have not been collected yet. The contribution from the measurement errors, as we shall see, is not negligible in most cases, thus inviting to further predict them with a complementary model (not accomplished in this pilot study). The contribution from the prediction errors is intimately linked to the quality of the model. If the model is accurate enough and no underfitting/overfitting is incurred, we may expect

$$\mathbb{B}(\widehat{Y}_{U_d}(t)\big|U_{d,c}, r_d(t), \mathbf{x}(t)) \approx \sum_{k\in r_d(t)} \mathbb{E}\left[E_k^{(m)}(t)\big|U_{d,c}, r_d(t), \mathbf{x}(t)\right],$$

which clearly invites us to model and predict also the measurement errors.

To estimate the conditionally expected generalised error $\mathbb{E}\left[\tilde{E}_k(t)\big|U_{d,c}, r_d(t), \mathbf{x}(t)\right]$ we may proceed following different assumptions. As a first option, we may estimate this expectation value as an empirical mean, as usual, which even allows us to use robust options such as the trimmed mean or the winsorised mean (Maronna et al., 2009). We can assume *exchangeability* among the industrial establishments at each predicted data set t for reference month m and year y and use

$$\widehat{\mathbb{E}}\left[\tilde{E}_k^{my}(t)\big|U_{d,c}, r_d(t), \mathbf{x}(t)\right] = \frac{1}{N_{r(t)}} \sum_{k\in r(t)} \tilde{E}_k^{my}(t)$$

as the estimated value. However, notice that this is too strong an assumption, since the conditioning values $U_{d,c}, r_d(t), \mathbf{x}(t)$ are not taken into account and, especially, the same pattern is assumed for every single statistical unit k, which is at odds with their different actual values and relevance in the indices.

Let us not assume exchangeability but, on the other hand, assume a stationary behaviour of the predicting capacity of the model, i.e. that the generalised errors $\tilde{E}_k^{my}(t)$ have the same behaviour for different reference time periods my and the same unit k. Then, we can estimate the expected value of $\tilde{E}_k^{m^*y^*}(t)$ by

$$\widehat{\mathbb{E}}\left[\tilde{E}_k^{m^*y^*}(t)\big|U_{d,c}, r_d(t), \mathbf{x}(t)\right] = \frac{1}{n_k} \sum_{my<m^*y^*} \tilde{E}_k^{my}(t), \tag{9}$$

where n_k denotes the number of past data values for unit k. We shall provide some insight on these estimators in section 7. A priori, since $k \in r_d$, the bias will boil down to the expected measurement error of each unit k (hence, again, the natural invitation to also model and predict the measurement errors).

4.5.2 Mean squared error

We reason similarly regarding the mean squared error of $\widehat{Y}_U(t)$. We shall focus on the conditional mean squared error so that, taking advantage of the notation introduced above, we can write

$$\mathbb{MSE}(\widehat{Y}_U(t)|r_d(t), U_c, \mathbf{x}(t)) = \sum_{k \in U_c} \mathbb{E}\left[\left(\tilde{E}_{kt}\right)^2 \left| r_d(t), U_c, \mathbf{x}(t)\right.\right], \tag{10}$$

where we use the corresponding definition for $\tilde{E}_{kt}$ given in equation (8b). As before, we can estimate the expected value of $\tilde{\mathbb{E}}\left[\left(\tilde{E}_{kt}^{my}\right)^2 \left| r_d(t), U_c, \mathbf{x}(t)\right.\right]$ for each statistical unit k, assuming time stationarity for the generalized errors and not assuming exchangeability, as:

$$\widehat{\mathbb{E}}\left[\left(\tilde{E}_{kt}^{m^*y^*}\right)^2 \left| r_d(t), U_c, \mathbf{x}(t)\right.\right] = \frac{1}{n_k} \sum_{my < m^*y^*} \left(\tilde{E}_k^{my}(t)\right)^2. \tag{11}$$

In section 6 we present the results and in section 7 we analyse these estimating proposals.

5 The production process for the early estimates of the ITI

5.1 Modularity in official statistics

As it is stated in the introduction, the industrial standardization comprises the adoption of international production standard models. In this chapter we focus on two of them: the Generic Statistical Business Process Model, GSBPM for short, (UNECE, 2019b) and the Generic Statistical Information Model, GSIM for short, (UNECE, 2013). GSBPM describes statistical production in a general and process-oriented way while GSIM is a reference framework for information objects. The production process for the early estimates of the ITI has been designed by following these models and the approach about the use of functional modularity stated in the working paper by Esteban et al. (2018).

The main idea can be observed in the diagram of figure 1 obtained from a seminar by Mark van der Loo (Statistics Netherlands) given at Statistics Spain (INE) in May 2019. This figure shows the two flows of inputs and outputs of a modular process: data flow in horizontal direction and metadata flow in vertical direction. Statistical processes, activities and tasks must be described in terms of the GSBPM and inputs and outputs must be described as GSIM information objects.

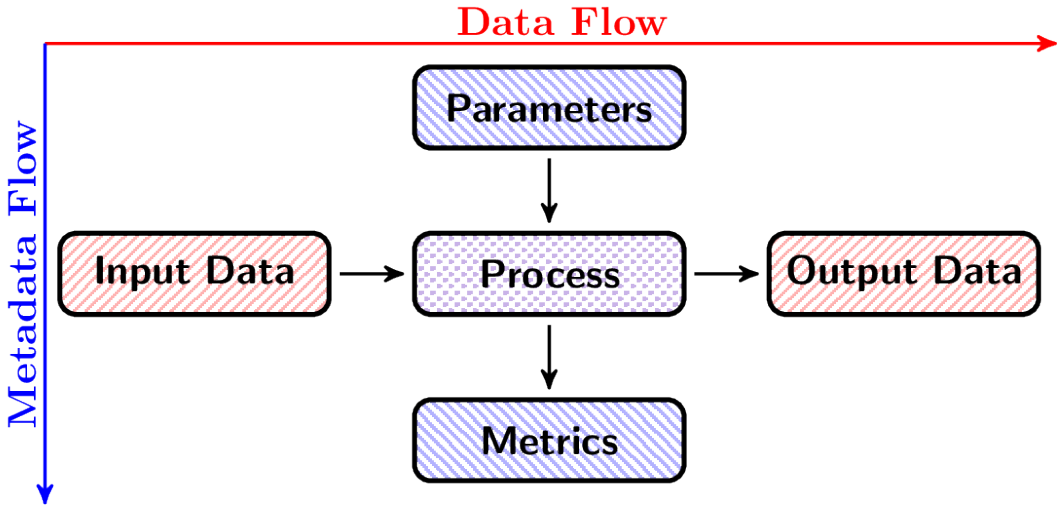

Figure 1: Modular process diagram

A process can consist of multiple steps where each step has its own inputs and outputs for both the data and metadata flows, which can be intermediate or final objects. There is a typical simile of a process step with a Lego piece so that, if each piece (step) is well designed, it can be fitted with the rest of pieces (steps) to form the whole structure (process).

5.2 The structure of the production process

In this section we describe the specific structure of the production process for the proposed advanced ITI. First, we show the pipeline of the process, i.e. the process design. Next, we show the specific computational implementation, i.e. how the execution is organised.

5.2.1 The workflow and dataflow of the process

The whole process from the collection of the data to the final dissemination of the aggregates is completely modular. It consists of the following steps:

- 00. Collect and Validate Data
- 01. Build Regressors
- 02. Build and Evaluate Model
  - 02.01. Train Model
  - 02.11. Predict
  - 02.21. Evaluate Predictions
- 03. Compute Aggregates
- 04. Visualize Output

The workflow and the flow of data can be observed in figure 2. In that diagram we see the process steps (logical flow) in blue colour and the data flow (physical flow) indicated with black arrows. There exist two types of data storage, namely a central repository (a central remote server) and smaller independent intermediate data folders (specific local folders for each step). The repository is a central standardised data repository of Statistics Spain (INE) containing microdata and paradata of multiple surveys and statistics from different stages of the production process (collection, editing, estimation). Intermediate data are not shared throughout the process because each step needs its own specific data storage in fulfillment of data abstraction, layering, and hierarchy. Only, the final output of each step (in yellow in the diagram) is shared with the rest of the process.

Now, we see each step in detail.

00. Collect and Validate data

This is the traditional data collection from respondents, including data editing during collection. This information is stored in the central repository in an internal standard format as key-value pair files named FD (edited file[3] by specialised clerks at provincial delegations) and FP (paradata file[4]). The file `FD[mm][yyyy]_D[t]` contains the values denoted by $z_k^{my,\mathrm{ed}}(t)$. Integrated GSBPM business functions in this step come from Phase 4 (Collect) and activity 5.3 (Review and Validate). See (UNECE, 2022) for a description of GSIM objects specifications for these generic processes.

01. Build Regressors

In this step, the available information is used to build the regressors used to predict the target variable. The microdata and paradata of the reference time period are taken from the repository as well as the final file of the previous periods (named FF[5].). The file `FF[mm-1][aaaa]_D[last]` contains the final validated values denoted by $z_k^{my,\mathrm{ed}}(t_{\mathrm{release}}) = z_k^{my,\mathrm{val}}$ at the press release moment. These values can undergo changes after the press release moment due to revisions and updates, then the last version of the file is taken. The main GSBPM business function is activity 5.5 (Derive new variables and results). GSIM objects specifications follow the general description by UNECE (2022).

02.01. Train Model

Once the data set is fully built, the model is also built. First, a preprocessing is conducted to deal with missing values and to do the encoding in each categorical variable. Next, a grid search is executed to choose the best hyperparameters of the model and then, the model is trained to render it ready for the application. The final output object comprises the fitted model and the training data used in this step. In terms of GSPBM business functions we agree activities 5.4 (edit and impute) and 5.5 (Derive new variables and results) to comprise missing variable treatment and all variable encodings in a statistical learning algorithm. No international agreement is known to these authors in this respect; however we find it the most appropriate GSBPM activity to describe this mathematical operation. However,

[3] In Spanish *F*ichero *D*epurado.
[4] In Spanish *F*ichero de *P*aradatos.
[5] In Spanish *F*ichero *F*inal

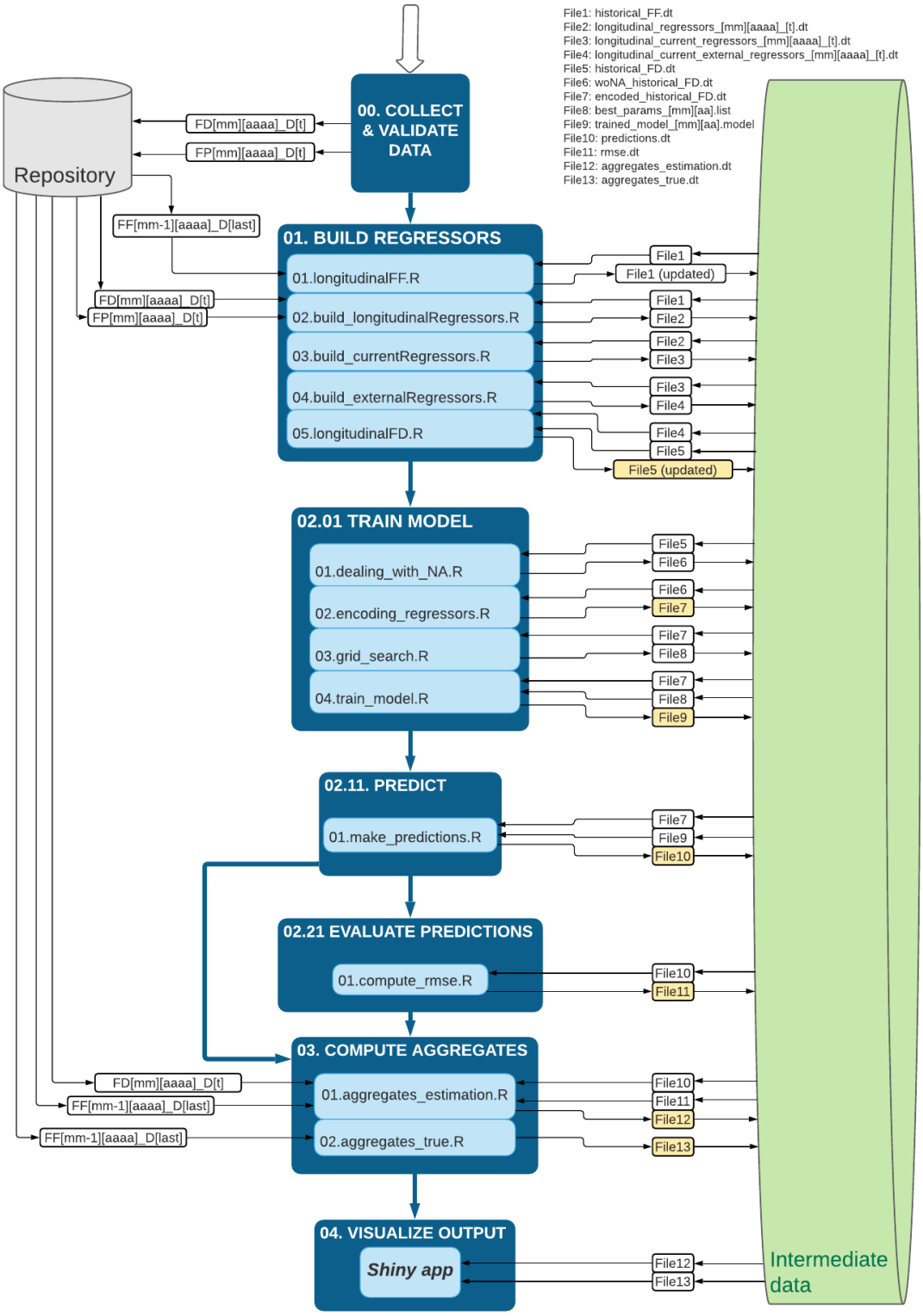

Figure 2: Advanced ITI statistical process diagram

we consider difficult to select GSBPM activities to describe the grid search and the model training, since we cannot envisage similar mathematical procedures in traditional survey methodology around which the GSBPM is built. An international agreement is needed; nonetheless at this point the identification of concrete mathematical procedures common to similar machine learning-based processes should now be prioritised.

02.11. Predict

Next, the turnover of industrial establishments not yet collected (i.e. $k \in U_c - r_d(t)$) is predicted for the current reference period using the model trained in the preceding step. These predicted values constitute the output of this step. In terms of the GSBPM, this can be described using the production activity 5.5 (Derive new variables and results). GSIM objects specifications follow the general description by UNECE (2022).

02.21. Evaluate Predictions

In this step, accuracy quality indicators are computed using the mean squared error, as explained in section 4.5. Mean squared errors of predicted values are estimated, which constitute the main output of this step. Since these values are still computed for each industrial establishment k, this is also described as a production activity 5.5 (Derive new variables and results) and GSIM objects specifications follow the general description by UNECE (2022).

03. Compute Aggregates

Once all variables have been computed for each statistical unit (industrial establishment), different aggregates are calculated. Totals of turnover, elementary indices and compound indices for all population domains are then computed. For the pilot study and the assessment of the present proposal, this is also computed using final validated true values similarly as with the predicted values. Root mean squared errors for the aggregates are computed using those for each industrial establishment. In this way, we shall be able to assess the quality of the early estimates.

In terms of the GSBPM, these tasks can be described as production activities 5.7 (Calculate aggregates) and GSIM objects specifications following the general description by UNECE (2022).

04. Visualize Output

Finally, the multiple time series of aggregates obtained in the preceding step (indices and errors) are visualized in interactive plots through a dashboard developed in Shiny (see section 6). This is a standard production activity described in terms of subprocesses in phase 7 (Disseminate) and the corresponding GSIM information objects. Notice, however, that this output is not intended for production purposes, only for analytical reasons (no press release is composed, no dissemination product beyond the interactive visualization itself).

5.2.2 The computational organization

The organization of the files and folders that contain the implementation of the process is fully based on the modularity approach and the diagram of figure 1. First there exists a general organization for the whole process where each step is a folder and there are general data storage (`data` and `repo`) and functions (`src`) that can be used in several steps. Each step is internally organised with the following set of folders: i) `code` (the implementation of the process itself), ii) `data` (storage for the intermediate inputs and outputs), iii) `metrics` (output of the metadata flow) and iv) `param` (input of the metadata flow).

The **code** folder has `script` and `src` folders to separate the scripts and the source functions. The **metrics** folder could have different kind of outputs such as graphics, reports or logs. In this case, just the `logs` are saved. The **param** folder has two types: `local` parameters and `global` parameters. In the global parameters there is information about the paths, the packages and functions to be loaded. In the local parameters there is information needed to run each part of the process: files_info (names, version of dictionary...), time (reference periods), execution (parameters of the model, name of the target variable, validation parameters ...).

If there is no survey-specific content, the files are valid for any survey. This happens in most of the functions and scripts. However, if there is some survey-specific information, this is saved in folders with the code of the survey (in our case E30052). The corresponding folders begin their names with the same code. Then, survey-specific information is located

just in the parameters files and in the specific functions inside the folders with the survey code, so the scripts and some general functions are free from survey-specific content therefore they are valid for any survey.

The source code for the whole process can be accessed at https://github.com/david-salgado/AdvITI.

6 Results

The main results of this pilot study comprise the series of early estimates of the ITI brokendown according to usual production conditions as well as their corresponding yearly and monthly variation rates for the three batches processed by the survey managers. These quantities are computed together with their respective root mean squared error. To assess the quality of these results we also compute these series for the prediction of the ITI without regressors from the current reference time period and for the true released value at $m + 51$.

The series comprise 60 consecutive months from May 2016 to April 2021. For each reference month we compute five values, namely the initial prediction without current data, the early estimates for the three batches, and the final validated value. The early estimates are computed together with their conditional root mean squared error. We have reconstituted the 7 types of breakdown for this index for each of their respective categories (see table 4).

Breakdown	**No. Categories**
General	1
NUTS2	17
MIGS	5
MIGS2	4
NACE Rev. 2 Section	2
NACE Rev. 2 Division	28
NACE Rev. 2 Division-Group[6]	38
Total	95

Table 4: Number of categories per index breakdown.

In figure 3 we represent an example comprising the three index versions (initial, batches, final) from January 2020 to April 2021.

In figure 4 we represent the corresponding annual variation rates for these same time periods.

The 2×95 time series (we focus on the index and the annual variation rate) cannot be graphically represented on a single working paper for all 60 analised time periods and breakdown categories. Thus, we have developed interactive Shiny applications (Chang et al., 2021) in the following URLs:

- Index Time Series: https://sandra-ba.shinyapps.io/Advanced_ITI_indices_v1/
- Yearly Time Series: https://sandra-ba.shinyapps.io/Advanced_ITI_annualRates_v1/

Statistical disclosure control has been applied so that not all 2×95 time series can be found.

7 Analysis of Results

To analyse the results we shall try to be as comprehensive as possible focusing both on the aggregate level and on the microdata level. An analysis from the perspective of the subject matter is also included.

7.1 Analysis of indices and rates

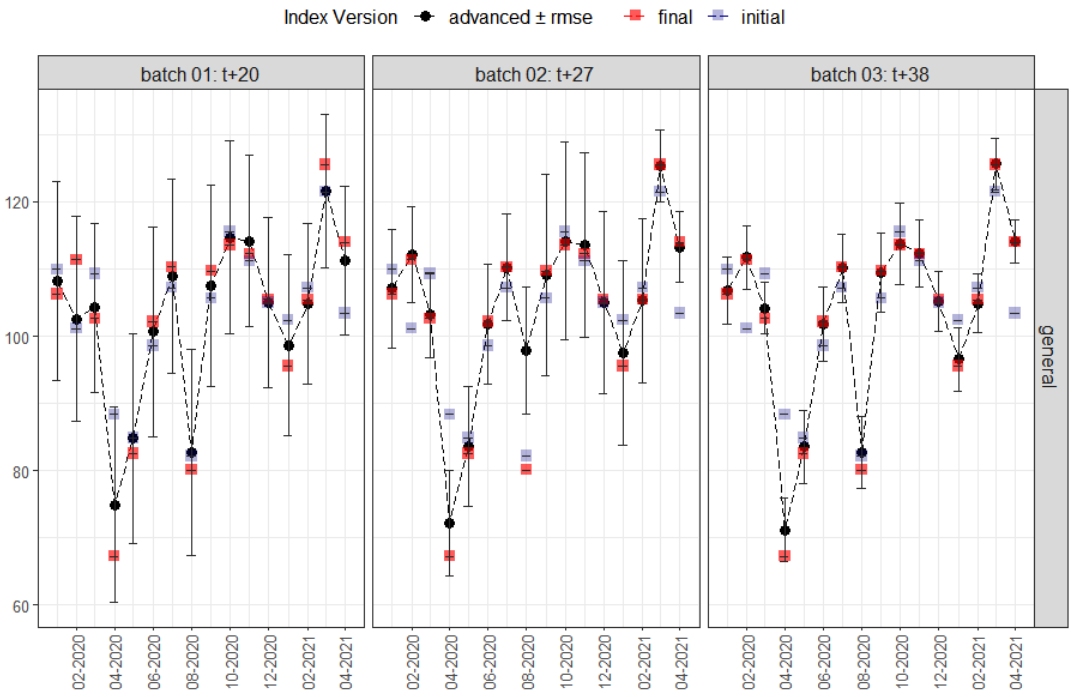

Figure 3: General Index Series from Jan, 2020 to April, 2021.

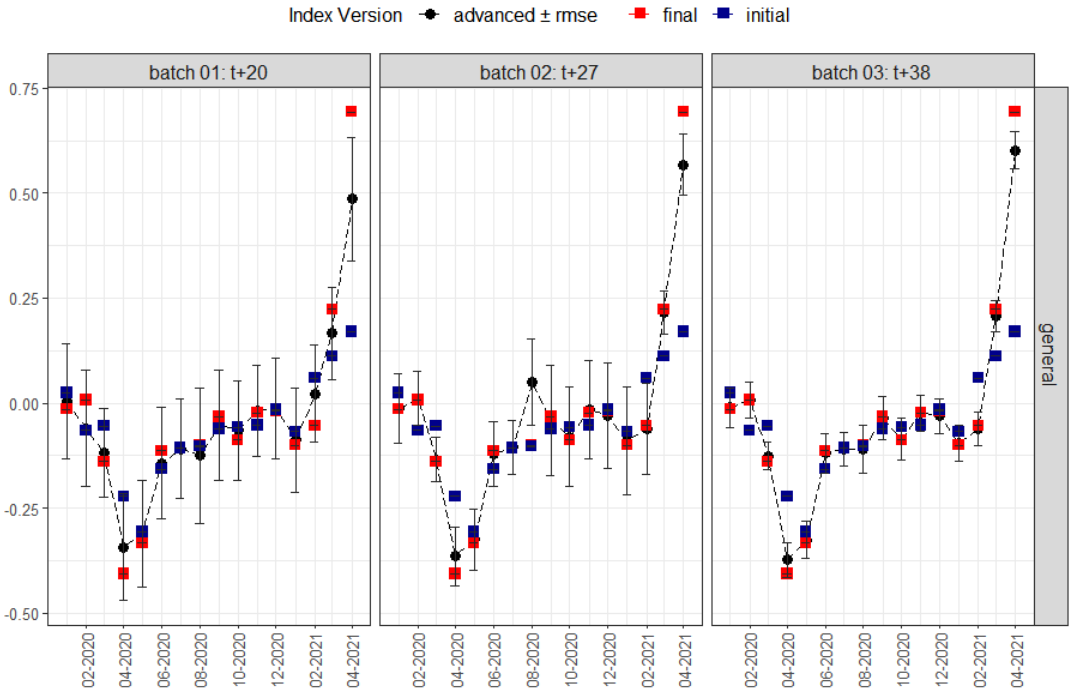

Figure 4: Annual Variation Rates Series from Jan, 2020 to April, 2021.

7.1.1 Indices

Firstly, we compute the relative error $\epsilon_d(t)$ of every single early index estimate of domain d and time instant t using

$$\epsilon_d(t) = \frac{\widehat{I}_{U_d}(t) - I_{U_d}^{true}}{I_{U_d}^{true}}.$$

In tables 5, 6, and 7 we show the first, second (median), and third quartiles of these relative errors $\epsilon_d(t)$ by breakdown categories according to the index dissemination plan. These are computed for the 60 time periods per batch.

Table 5: Quartile Q1 for the relative error for early index estimates by breakdown (in %).

breakdown	**initial**	**batch 01**	**batch 02**	**batch 03**
general	-2.46	-1.17	-0.20	-0.09
section	-1.10	-1.19	-0.30	-0.16
MIG	-3.01	-1.29	-0.41	-0.20
MIG2	-3.45	-1.43	-0.32	-0.14
NUTS2	-2.80	-1.59	-0.51	-0.27
division	-2.67	-1.42	-0.42	-0.23
division-group	-2.94	-1.55	-0.50	-0.27

Table 6: Median (quartile Q2) for the relative error for early index estimates by breakdown (in %).

breakdown	**initial**	**batch 01**	**batch 02**	**batch 03**
general	-0.68	-0.35	0.23	0.12
section	1.26	-0.02	0.24	0.10
MIG	-0.12	0.07	0.05	0.06
MIG2	-0.23	0.03	0.05	0.05
NUTS2	0.05	0.30	0.24	0.17
division	0.34	0.28	0.14	0.08
division-group	0.33	0.33	0.14	0.07

Table 7: Quartile Q3 for the relative error for early index estimates by breakdown (in %).

breakdown	**initial**	**batch 01**	**batch 02**	**batch 03**
general	1.91	1.30	0.63	0.45
section	4.52	1.99	1.34	0.84
MIG	2.70	1.34	0.75	0.53
MIG2	2.36	1.28	0.76	0.51
NUTS2	2.97	2.28	1.32	0.93
division	3.94	2.81	1.55	1.08
division-group	3.69	2.78	1.47	1.03

We clearly observe how the error gets smaller as more data are collected and used for training the model. It is important to notice how the use of data from the current period evidently improves the accuracy. The whole empirical distributions are represented in figures 5 and 6 (zoomed in).

To have an insight of the fraction of early index estimates above a given relative error for each breakdown we include figure 7. We obtain that around 85% (77% for batch 01, 86% for batch 02, and 88% for batch 03) of early estimates are below a relative error of around

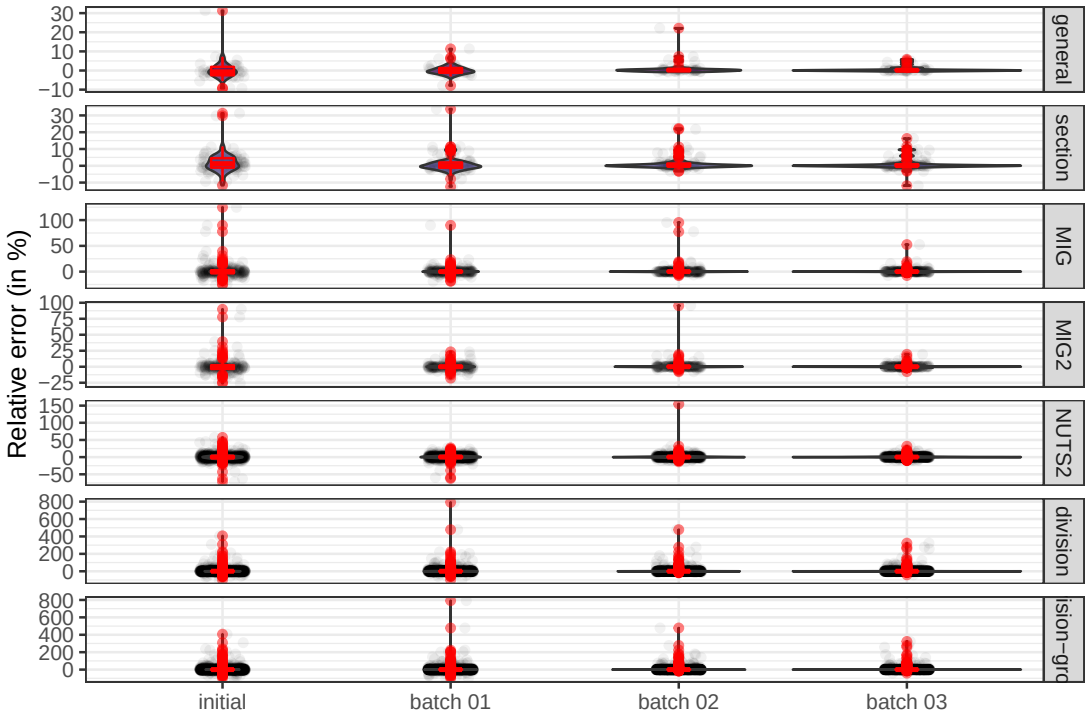

Figure 5: Empirical distributions of relative error (in percentage).

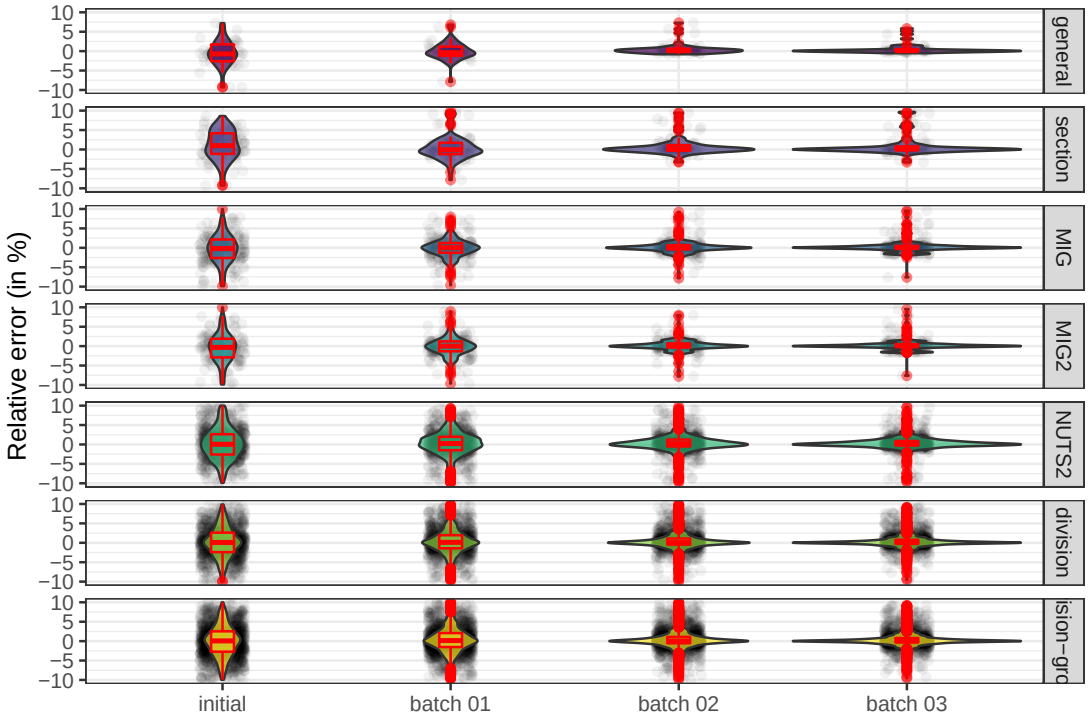

Figure 6: Empirical distributions of relative error (in percentage) (zoomed in).

10%, except for the initial prediction without data from the current reference month (which falls down to around 45%).

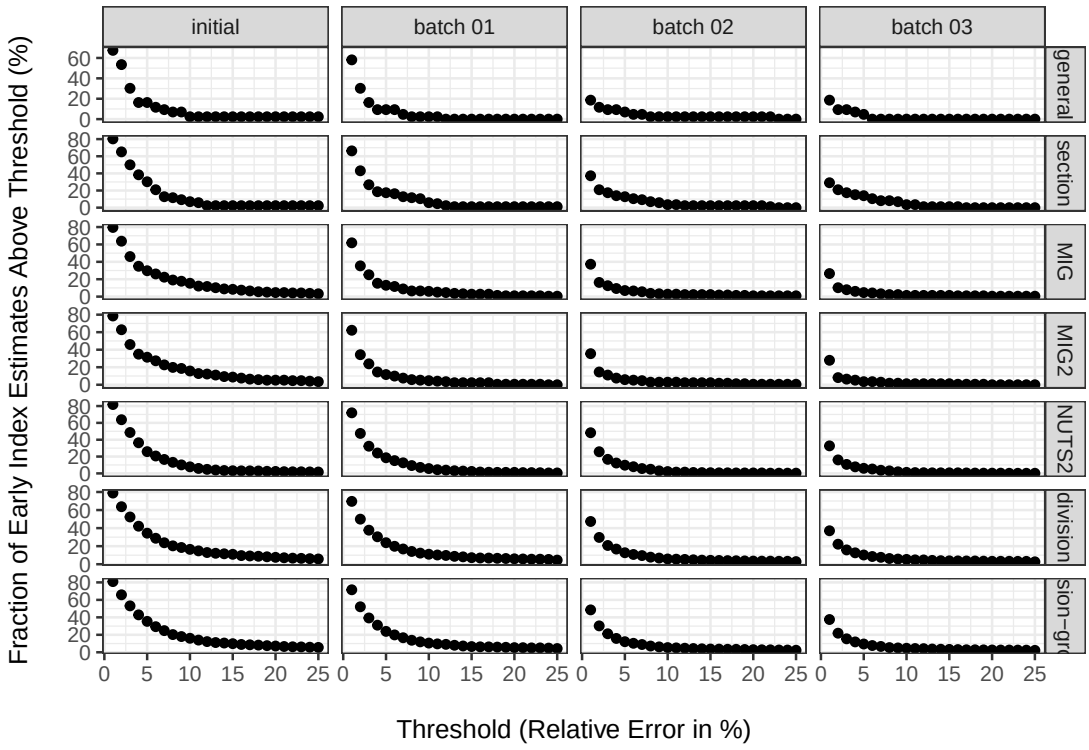

Figure 7: Fraction of early index estimates above a given relative error.

We can follow a complementary view on the accuracy by analysing the coverage rate of the uncertainty intervals based on the estimated root mean squared error. We define these uncertainty intervals as $I_\xi = \xi * \widehat{rmse}$ and compute the fraction of early estimates within the corresponding interval. This is represented in figure 8 for different factors ξ. On average, already for covering factors $\xi \geq 1.15$ we get that more than 95% of early estimates lie within the uncertainty intervals.

Another interesting result can be obtained analysing the behaviour of the absolute value of the relative index error $\epsilon_d(t)$ with respect to the number of statistical units in the breakdown category. To do this, we compute the median values of $\epsilon_d(t)$ and the sample size $n_d(t)$ for each batch using all reference time periods (60) and conduct a robust linear regression to contrast $\epsilon_d = O(n_d^{-\beta})$. In figure 9 we show that (i) $\beta > 0$ (except for very small-sized categories), thus with an expected dependence on the sample size and (ii) $\beta \to 0$ counterclockwisely as we use more collected data to construct the early estimate. Eventually $\beta > 0$ beacause of the presence of non-treated measurement errors. General and section breakdown categories have not been considered in this analysis because they have an irrelevant number of cases: 1 and 2, respectively. In our view, item (i) arises because of a larger number of random fluctuations cancel out with a larger sample size and item (ii) is a natural consequence of using a lesser number of predicted values in the index computation, although with remaining measurement errors.

We can repeat this analysis in terms of the volume of measurement error per breakdown category. We construct a summarization figure of merit for the relative measurement error as follows. Firstly, we compute the relative measurement error

$$\epsilon_k^{(m)}(t) = \frac{z_k^{\mathrm{ed}}(t) - z_k^{\mathrm{val}}}{z_k^{\mathrm{val}}}$$

for each unit k. Second, we compute the median of $\epsilon_k^{(m)}(t)$ per reference month, batch and breakdown category (general, section, MIG2, MIG, NUTS2, division, and division-group).

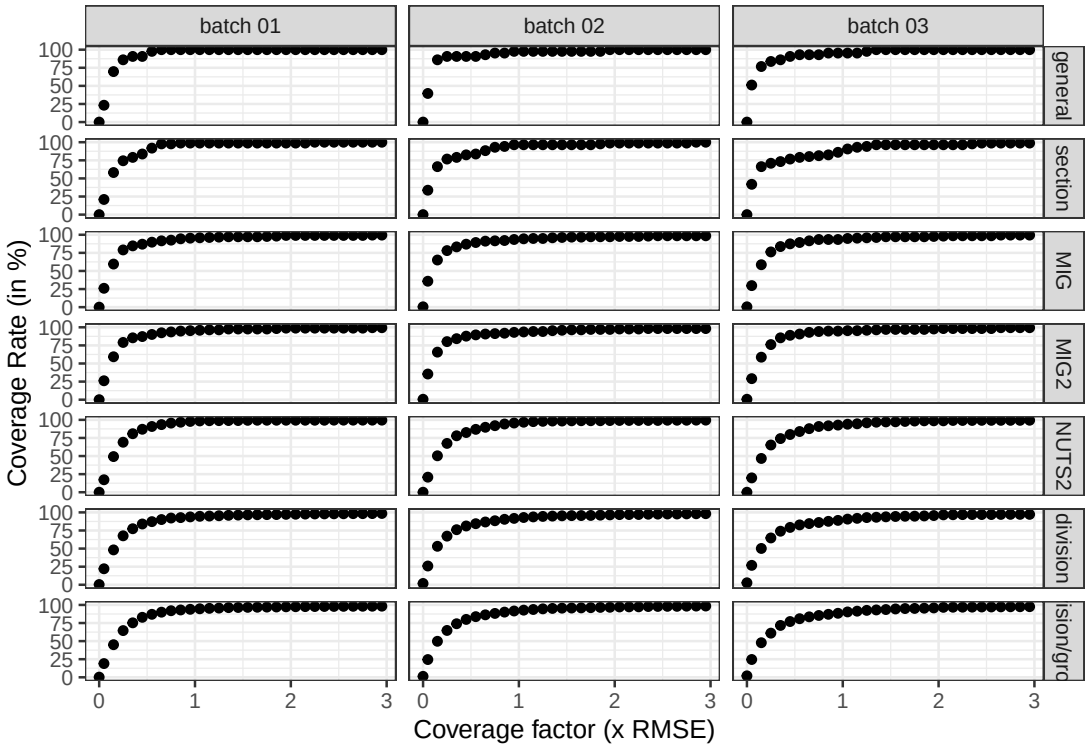

Figure 8: Coverage rate of rmse-based uncertainty intervals for the early index estimates.

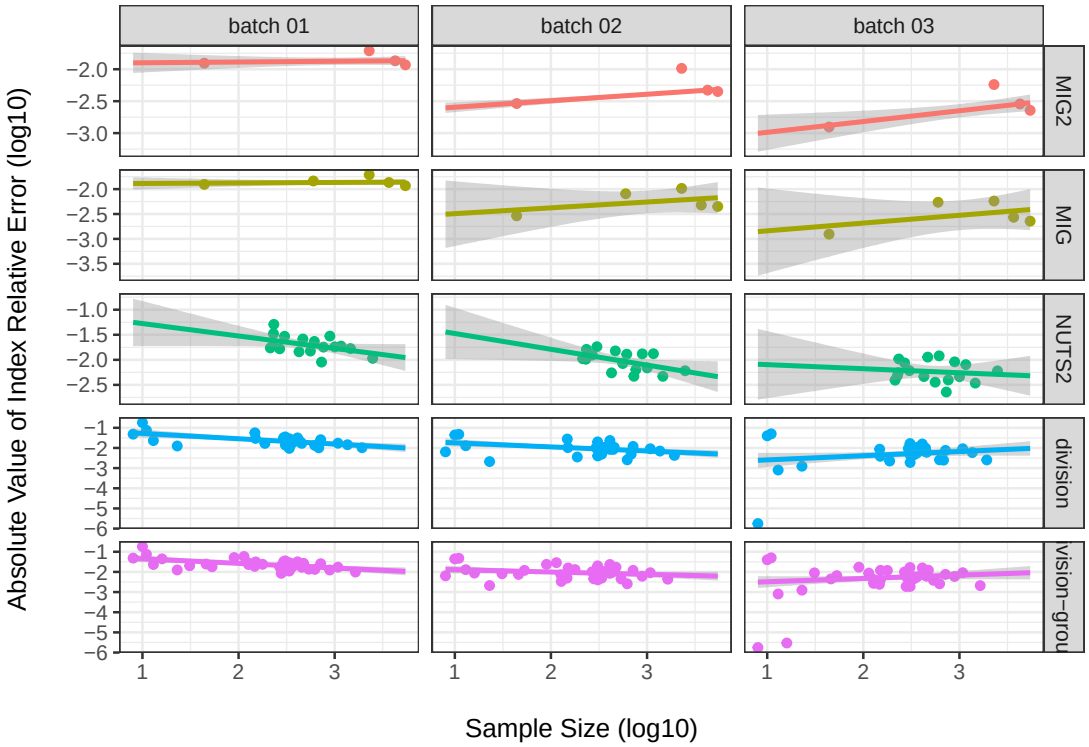

Figure 9: Absolute index error vs. sample size by breakdown and data batch.

Finally, these medians $\epsilon_k^{(m)}(t)$ are averaged over the whole sequence of time periods (60). Figure 10 (containing again robust linear regression models) shows a positive correlation (except for breakdown categories with very few values). As stated in preceding sections, this clearly invites us to complete the computation of the early estimates with a prediction of measurement errors: $\epsilon_d = O\left((\epsilon_d^{(m)})^\beta\right)$ with $\beta > 0$ and $\beta \nrightarrow 0$ as we have more collected values (measurement errors persist).

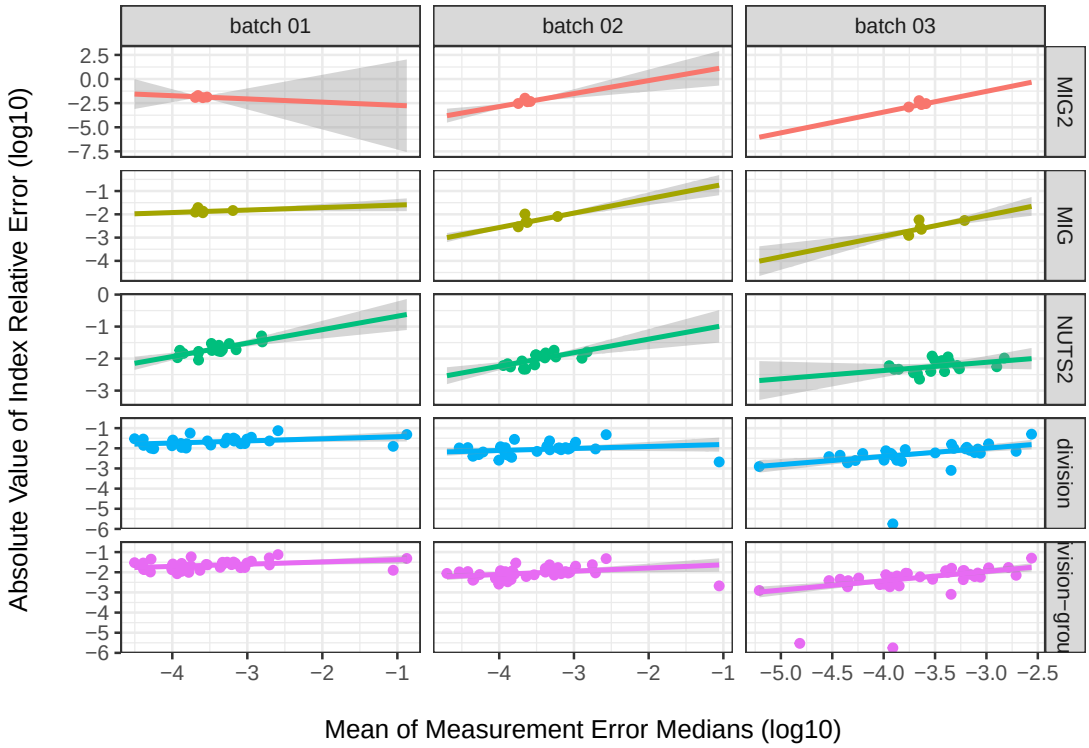

Figure 10: Absolute index error vs. measurement error by breakdown and data batch.

7.1.2 Variation rates

Annual variation rates of indices bear equal or greater importance, thus we should repeat this analysis for them. Instead of focusing on relative errors, we shall centre on absolute errors (number of percentual points). In tables 8, 9, and 10 we compute the third, second (median), and third quartiles of the error (in number of percentual points) of the early estimates of the annual variation rates.

Table 8: Quartile Q1 of the error of early estimates of annual variation of indices by breakdown (in percentual points).

breakdown	**initial**	**batch 01**	**batch 02**	**batch 03**
general	-2.73	-1.28	-0.25	-0.11
section	-1.27	-1.32	-0.32	-0.18
MIG	-3.36	-1.49	-0.47	-0.24
MIG2	-3.75	-1.70	-0.37	-0.16
NUTS2	-3.21	-1.82	-0.58	-0.30
division	-2.81	-1.61	-0.47	-0.26
division-group	-3.27	-1.77	-0.54	-0.31

Table 9: Median (quartile Q2) of the error of early estimates of annual variation of indices by breakdown (in percentual points).

breakdown	initial	batch 01	batch 02	batch 03
general	-0.77	-0.36	0.27	0.14
section	1.44	-0.02	0.27	0.13
MIG	-0.15	0.06	0.04	0.07
MIG2	-0.27	0.03	0.05	0.07
ccaa	0.06	0.33	0.27	0.19
division	0.33	0.29	0.16	0.08
division-group	0.32	0.33	0.16	0.08

Table 10: Quartile Q3 of the error of early estimates of annual variation of indices by breakdown (in percentual points).

breakdown	initial	batch 01	batch 02	batch 03
general	2.04	1.44	0.71	0.49
section	5.03	2.14	1.32	0.97
MIG	2.61	1.47	0.81	0.58
MIG2	2.48	1.38	0.82	0.57
ccaa	3.02	2.29	1.38	0.95
division	3.58	2.50	1.41	1.00
division-group	3.51	2.66	1.42	1.00

In figure 11 we represent the evolution of the coverage rates of the rmse-based uncertainty intervals for the early estimates of annual variation rates by breakdown category. On average, already for covering factors $\xi \geq 1.15$ we get that more than 93% of early estimates of annual variation rates lie within the uncertainty intervals.

7.2 Analysis of microdata predictions

7.2.1 Point predictions

Now we focus on the individual turnover predictions for all three batches and reference time periods. We shall assess the final synthetic prediction, which is the model prediction $\widehat{y}_k(t)$ when $k \notin r_t$ and the raw value $y_k^{\text{raw}}(t)$ when $k \in r_t$, by comparing it with the final turnover validated value (see figure 12). We clearly observe three clusters, namely, (i) the expected predicting behaviour (points around the line $y = x$), (ii) values predicted with (quasi) zero value but not having final zero turnover, and (iii) values predicted with non-zero values but having final zero turnover. These two last cases require special treatment according to the subject matter (units ceasing their activity, units having turnover decreased below the cut-off value,...); this has not been taken into account into the prediction model.

We can also compute the quantiles of the residuals (see table 11). We observe that for batch 01, 80% of them lie in the interval $[-75000, 180000]$ while for batches 02 and 03 this interval narrows down to $[-3, 4]$, i.e. basically no error[7].

batch	min	$c_{0.05}$	$c_{0.10}$	$c_{0.25}$	$c_{0.50}$	$c_{0.75}$	$c_{0.90}$	$c_{0.95}$	max
01	-707594876	-291139	-74983	-3	-2	4	179936	474030	886281465
02	-707594876	-16367	-3	-3	-2	0	4	172131	5660465879
03	-289082043	-3	-3	-3	-2	0	4	61156	881703000

Table 11: Quantiles of residuals for the final predicted values of the turnover.

[7] This low numbers arise because of the substitution of zero values in national, euro, and extra-EU market turnover to avoid numerical instabilities.

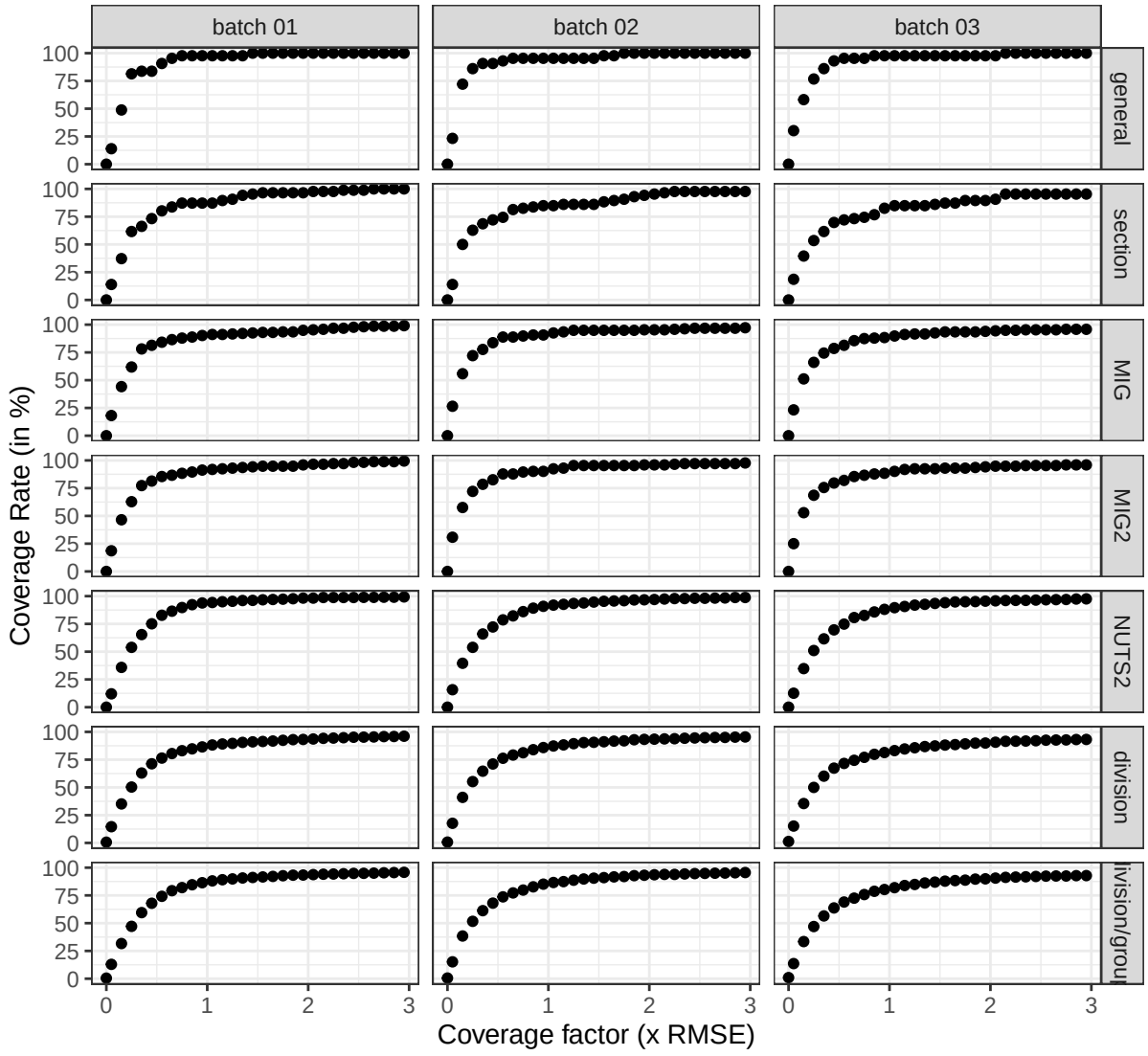

Figure 11: Coverage rate of rmse-based uncertainty intervals for the early estimates of the annual variation rates.

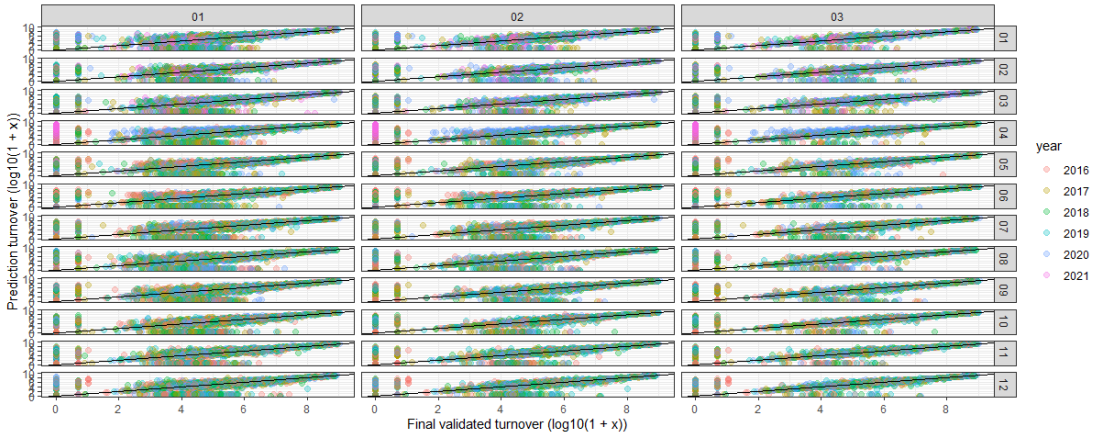

Figure 12: Scatterplot of the final turnover model predicted values vs. final turnover validated values

Outliers are clearly present and need further treatment (not undertaken in this pilot study). This can be clearly observed in the density plots by batch and year (see figure 13). Thus, the residuals distribution seems to be a mixture again of three populations, namely (i) those centred around zero, (ii) those with overestimation (non-zero predicted value but final validated zero value), and (iii) those with underestimation (zero-predicted value but final validated non-zero value). As data collection and data editing progresses in the field work (batch number increases), these two last populations decrease (except for the year 2021 when multiple industrial establishments needed to be removed from the sample because their turnover fell below the cut-off threshold).

By plotting the standardised residuals against the predicted values we can assess the homoskedastic/heteroskedastic nature of the model (see figure 14). We detect a different behaviour for small-sized industrial establishments (some heteroskedastic behaviour, especially affecting some underestimation) in contrast to large-sized establishments (random patterns in the standardised residuals, thus showing homoskedasticity). This suggests the previous use of an appropriate Box-Cox transformation (e.g. log). In this way we would improve the estimation of the prediction accuracy. As stated before, we prioritize the design of an end-to-end modular process susceptible of multiple incremental improvements over the optimal choice of some aspects of the model construction. This is clearly a feature to be improved in an enhanced version for implementation.

7.2.2 Uncertainty intervals

To assess the performance of the uncertainty intervals based on the root mean squared error, we shall proceed as before. We define uncertainty intervals for each unit k and time instant t as $I_{k,\xi}(t) = \xi * \widehat{rmse}_k(t)$ and compute the fraction of point estimates within the corresponding interval. This is represented in figure 15 for all dissemination cells. On average, already for covering factors in the range $[1, 1.3]$ we get that 86% of point predictions lie within the uncertainty intervals. This is an acceptable coverage rate, thus the point predictions are fairly accurate enough.

7.3 Subject matter analysis

Both the timeliness and the accuracy are two important quality dimensions requested by the users and that Official Statistics should always keep in mind. A balance between them both is required at any time and for any statistical operation. This proposal shows how it can be achieved with statistical learning algorithms.

From the business manager point of view, this new way of producing early estimates of the indices make it possible to have reliable information in a shorter period of time. The timeliness can be reduced in about 30 days keeping under control the accuracy assessment.

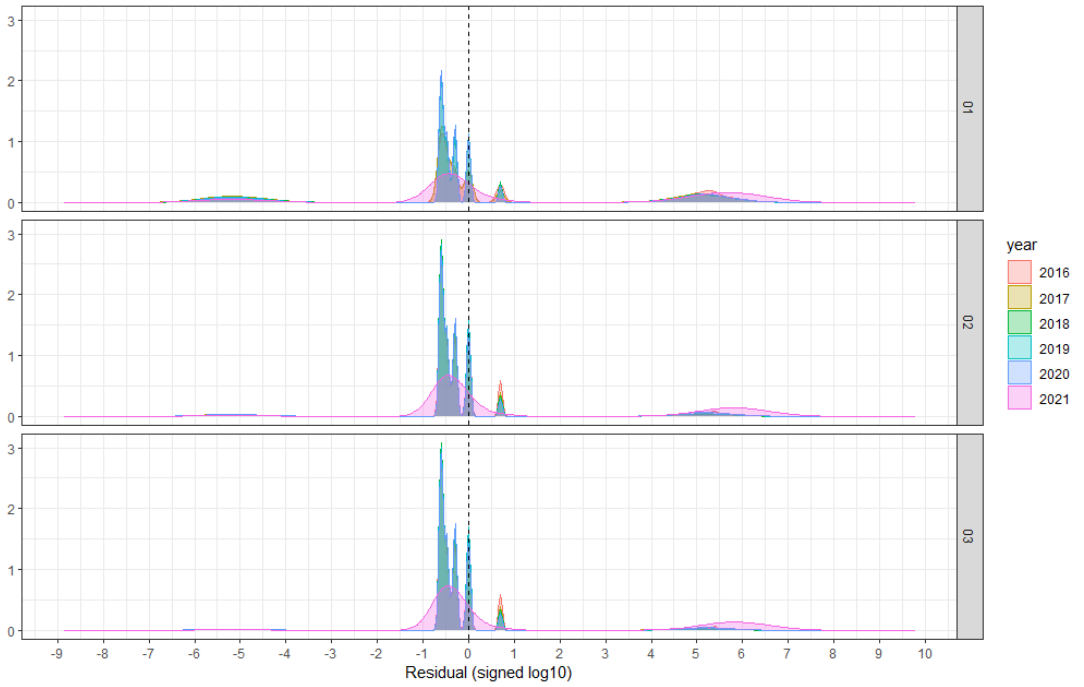

Figure 13: Density plots of the residuals of the final predicted values. Notice the signed log10 scale.

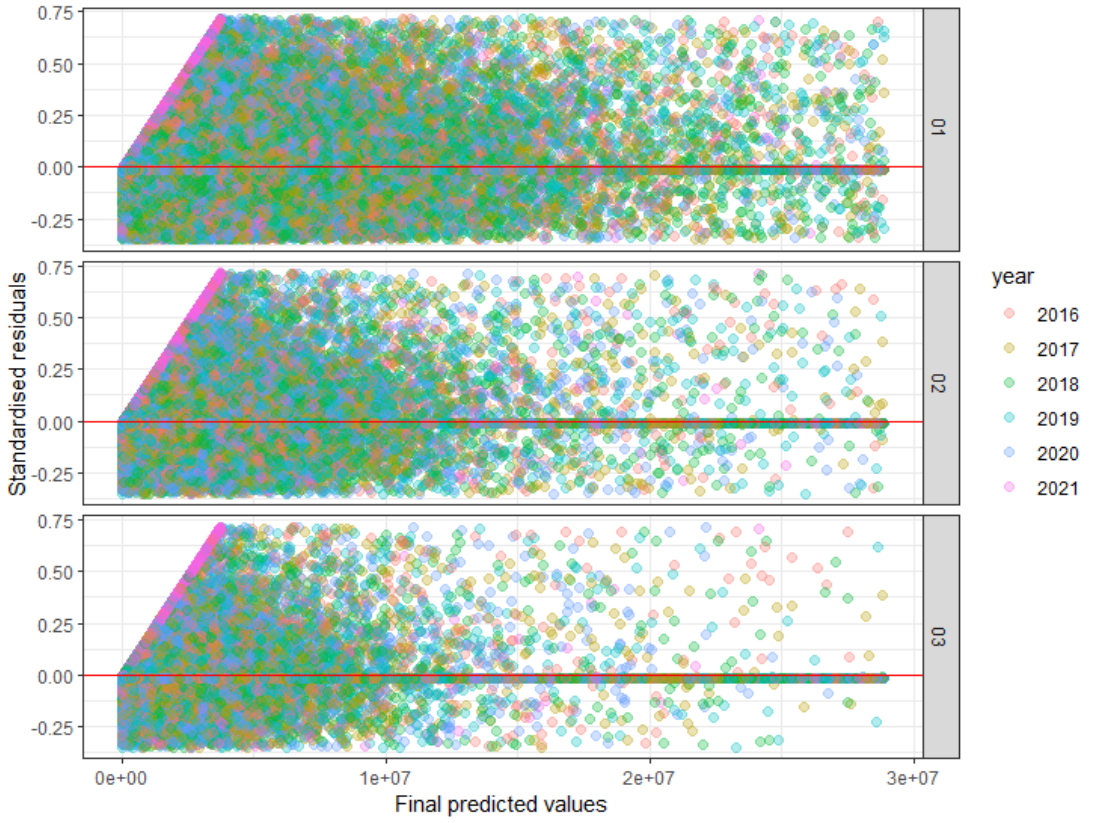

Figure 14: Scatterplot of standardised residuals vs. final predicted turnover values. Outlying residuals (above quantile 0.995 and below quantile 0.005) are not included in the plot.

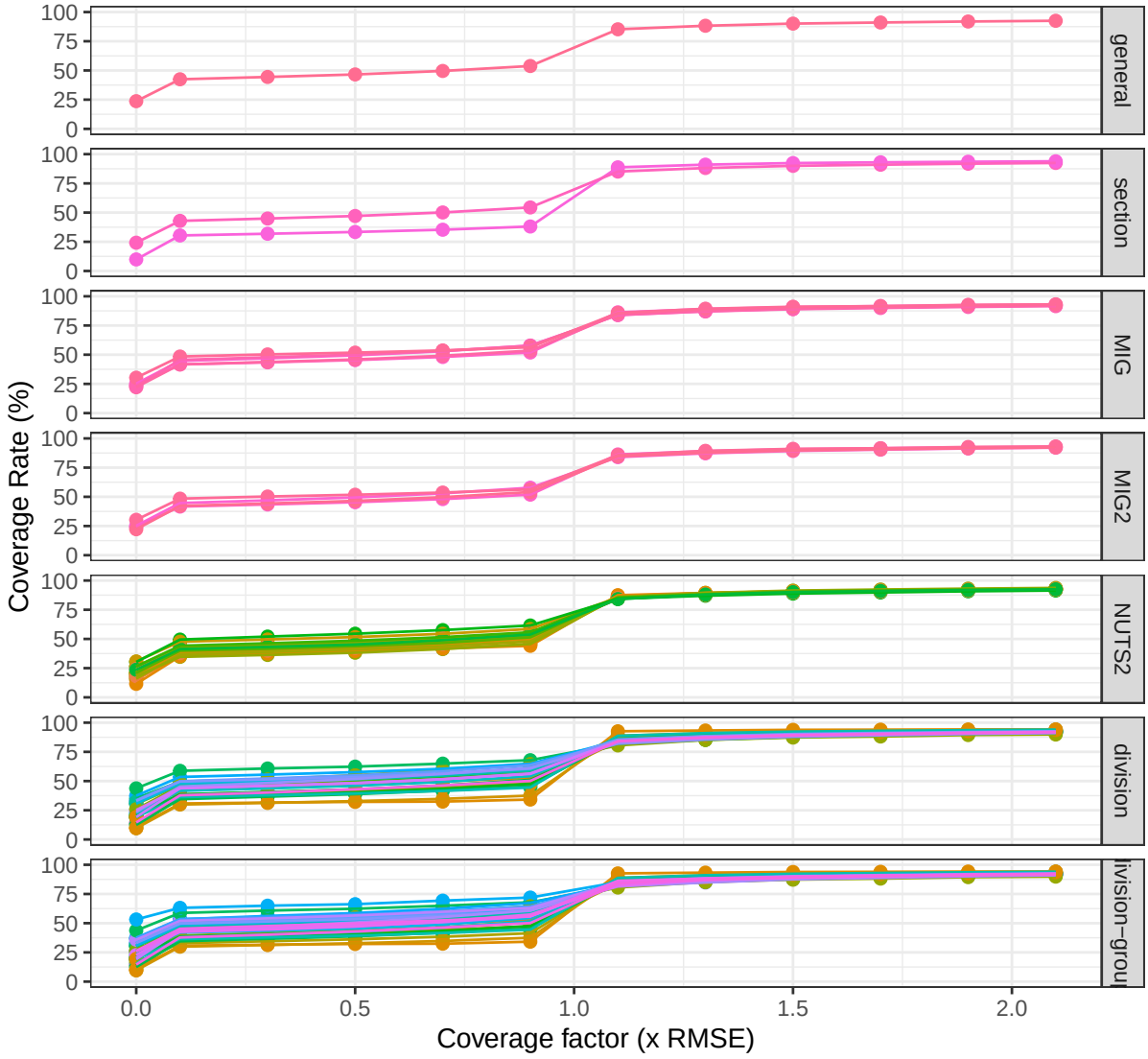

Figure 15: Coverage rate of rmse-based uncertainty intervals for the unit point predictions.

On the one hand, the fact that we only use data sources already collected by Statistics Spain (INE) itself provides autonomy and independence and makes it possible to compute both the predictions and the accuracy indicators with traditional survey data. No new data source needs to be accessed and integrated. On the other hand, despite the fact of using a fraction of the sample to reconstruct the entire microdata set, we show that it is possible to detect the turning points in the statistics' time series, such as the COVID crisis onset (reflected in several divisions during the lock-down and recovery period), the lack of microchips affecting the manufacture of motor vehicles, trailers and semi-trailers or the increase of prices affecting the manufacture of coke and refined petroleum products (both in 2021).

Finally, the data validation by subject matter experts proves to be important to provide reliably labelled variables (both target and regressors). The design and implementation of efficient data editing and imputation strategies, now also considering this use of statistical learning algorithms, gains relevance, especially regarding the detection and treatment of outliers produced by changes in the economic activity code and/or by units decreasing their activity below the threshold values.

8 Conclusions and future work

This work provides a first pilot experience in the construction of a prototyping end-to-end statistical process producing early estimates of a monthly short-term business statistics using survey data during their collection phase. Using a gradient boosting regression machine we make use of historical microdata and aggregated data from the same survey and aggregated data from the fraction of the sample already collected to predict the target variable of each single remaining statistical unit. Then, the standard process of computation is applied to the synthetic microdata set, together with a measure of uncertainty, to produce early estimates of the complete set of total turnover indices to be disseminated later on according to the official release calendar.

We have prioritised the design and test of a modular process susceptible of incremental improvements in different phases of the statistical learning model construction pipeline. In this sense, many aspects of the process described in preceding sections can be rightfully improved such as the hyperparameter search optimization, the inclusion of more regressors, the systematic study of alternative statistical learning algorithms (neural networks, random forests, etc.), the treatment of measurement errors with a specific complementary model, and some more. However, the current state of the process already produces fairly accurate early estimates together with uncertainty intervals assessing their reliability.

The present use of machine learning involves the human identification and construction of regressors, so that the participation of subject matter experts with a robust knowledge of the production process of the statistics is not only advisable but also necessary for a high-quality predictive model.

The role of statistical officers, in general, and of subject matter experts, in particular, needs some careful thoughts in this new context where statistical algorithms and process automation are considered. We may wonder whether manual tasks and the intervention of humans is not necessary any more. Our feeling is that the use of these new tools does not eliminate humans from the statistical process but rather on the contrary their role needs to be adjusted to the possibilities offered by the new context. Probably, the generic structure of editing and imputation strategies needs to be rethought and a new combination of business functions, or even new editing business functions, should be considered. Further work and empirical evidence is needed in this line.

The modular design of the process is a must for an efficient implementation and deployment where production tasks are reused in other surveys and/or statistical offices. This is completely aligned with international production standards such as the GSBPM. However, in our view, much work in the international community is still needed to reach an agreement on how to describe a statistical learning production pipeline in terms of GSBPM activities.

Nonetheless, the modular approach allows us to identify independent aspects which may be improved in the future. From the context of the present study with the Spanish Industrial Turnover Index Survey, we can recognise the following:

- We need to construct a complementary model to correct for the measurement error

(thus also of the bias of the early estimates). This can be generalised to predict any semicontinuous variable (Bohnensteffen, 2020).

- The underlying idea for these early estimates is not limited to cut-off sampling designs, and the approach can be readily extended to probabilistic sampling designs. More work is needed in this direction to analyse the computation of uncertainty intervals and accuracy measures (both conditional and unconditional variances).
- An extension of this study can be, in principle, applied to data collection procedures providing new daily data sets, thus producing daily updates of the early estimates. These new statistical methods strongly suggest that modifications need to be introduced in traditional statistical production processes to increase the quality of official statistics so that the best of statistical learning techniques can be put in place. This would imply to prepare the data collection procedures to store daily datasets for daily model training and use.
- An exploration of deep learning algorithms is needed to analyse whether manual identification and construction of regressors can be made automatic and even beaten. Complementarily, research on more regressors containing more information about the production process is also advised (e.g. use of historic register of edit activation per unit).
- A systematic study on model variable importance is needed to provide interpretability of the early estimates so that we can understand how regressors affect the predictions.

All in all, statistical learning techniques should become another versatile tool for producers of official statistics so that quality can be continuously improved. This pilot study shows a specific use improving the timeliness of existing survey-based short-term business statistics.

Appendix: Regressors

We standardise the names of the regressors to show their meaning and computation. We shall adopt the following conventions:

- Neither we shall distinguish in the notation between scalar and vector variables nor between categorical and continuous variables.
- We shall denote by N_A the cardinal of the set A.
- We shall denote by F_x^* the empirical cumulative distribution function computed from the realization x of a random variable X. This is computed with the `ecdf` function in R.
- The target variable (turnover) will be denoted by z_k for each unit k.
- Suffix `ent` will stand for *enterprise* (company or corporation possibly owning one or several industrial establishments in different locations). It should be reminded that statistical units are industrial establishments. In most cases, the enterprise has only one establishment and the statistical unit coincides with the enterprise, but in few cases a firm may have several premises.
- Suffixes `ed` and `val` will qualify computations using the edited and validated values of the turnover, respectively. Edited values refer to values along the execution of the data editing strategy and validated values refer to final validated values after the execution of the data editing strategy.
- Integer suffixes `i` qualify computations using values from the preceding ith time period.
- One-hot encoding is applied to categorical variables with less than 31 categories. This is implemented using the function `one_hot` from the `mltools` package (Gorman, 2018). Missing values are encoded as one of the categories. Coded values are $(1, 0, \ldots, 0)^t$, $(0, 1, \ldots, 0)^t$, $\ldots$, $(0, \ldots, 0, 1)^t$ corresponding to each category (no $(0, \ldots, 0)^t$ value is used).
- Mean encoding is applied to categorical variables with more than 30 categories. To avoid overfitting, the mean is computed on the moving average MA12_trnovr_val$_k$ of the validated value of the turnover over the past 12 months before the reference period for each category:

$$[\text{var}]_j\text{\_encoded} = \frac{1}{N_{\text{Cat}_j^{(12)}}} \sum_{k \in \text{Cat}_j^{(12)}} \text{MA12\_trnovr\_val}_k, \tag{12}$$

where $\text{Cat}_j^{(12)}$ stands for the units within category j of variable var in the past 12 months before the reference period.

- Quantiles are computed with the `quantile` function with the default algorithm (`type = 7`, i.e. definition 7 by Hyndman and Fan (1996)).

Geographical variables

`code_NUTS2_ent_ed`	
Definition	Edited value of the NUTS2 code for the territorial unit of the enterprise owning the industrial establishment
Stat Type	Categorical
Values	$01 - 17$
Example	05
Source	Internal-Primary
Formula	-
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Unit
Time Periods	$\{0\}$
Long/Cross	-
Cross-Domain Vars	-
Encoding	One-hot

`code_NUTS2_val_1`	
Definition	Validate value of the NUTS2 code for the territorial unit of the industrial establishment
Stat Type	Categorical
Values	$01 - 17$
Example	08
Source	Internal-Primary
Formula	-
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Unit
Time Periods	$\{-1\}$
Long/Cross	-
Cross-Domain Vars	-
Encoding	One-hot

`code_prov_ent_ed`	
Definition	Edited value of the code for the province of the enterprise owning the industrial establishment
Stat Type	Categorical
Values	$01 - 50$
Example	04
Source	Internal-Primary
Formula	-
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Unit
Time Periods	$\{0\}$
Long/Cross	-
Cross-Domain Vars	-
Encoding	Mean

`code_prov_ed`	
Definition	Edited value of the code for the province of the industrial establishment
Stat Type	Categorical
Values	$01 - 50$
Example	02
Source	Internal-Primary
Formula	-
Stat Progr Ref	Spanish IOE-30052

Unit/Aggr	Unit
Time Periods	{0}
Long/Cross	-
Cross-Domain Vars	-
Encoding	Mean

`code_munic_ent_ed`	
Definition	Edited value of the code for the municipality of the enterprise owning the industrial establishment obtained from the postal code
Stat Type	Categorical
Values	$000 - 999$
Example	009
Source	Internal-Primary
Formula	-
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Unit
Time Periods	{0}
Long/Cross	-
Cross-Domain Vars	-
Encoding	Mean

`code_postal_ent_ed`	
Definition	Edited value of the postal code of the enterprise owning the industrial establishment
Stat Type	Categorical
Values	$[01 - 50][000 - 999]$
Example	28050
Source	Internal-Primary
Formula	-
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Unit
Time Periods	{0}
Long/Cross	-
Cross-Domain Vars	-
Encoding	Mean

Time variables

year_ref	
Definition	Year of the reference time period for the corresponding target variable value z_k
Stat Type	Categorical
Values	$2015 - 2021$
Example	2018
Source	Internal-Primary
Formula	-
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Unit
Time Periods	$\{0\}$
Long/Cross	-
Cross-Domain Vars	-
Encoding	One-hot

month_ref	
Definition	Month of the reference time period for the corresponding target variable value z_k
Stat Type	Categorical
Values	$01 - 12$
Example	08
Source	Internal-Primary
Formula	-
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Unit
Time Periods	$\{0\}$
Long/Cross	-
Cross-Domain Vars	-
Encoding	One-hot

batch	
Definition	Batch of the data collection and transmission for the corresponding target variable value z_k
Stat Type	Numerical
Values	$1 - 3$
Example	2
Source	Internal-Primary
Formula	-
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Unit
Time Periods	$\{0\}$
Long/Cross	-
Cross-Domain Vars	-
Encoding	-

nmonths$i$_imputd_xprt	
Definition	Number of months within the last i months before the reference time period in which the target variable value z_k has been manually imputed by a subject matter expert
Stat Type	Numerical
Values	$0 - i$, $i = 3, 6, 12$ (three variables)
Example	2
Source	Internal-Derived
Formula	$\sum_{j=1}^{j=i} \delta(z_k^{(m-j)y,\text{imp}} = \text{expert})$, where δ is an indicator function and $z_k^{my,\text{imp}}$ is a paradata variable indicating the mode of imputation for the variable z of unit k in the reference time period with month m and year y
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Unit
Time Periods	$\{-1, -2, \ldots, -i\}$

Long/Cross	Long
Cross-Domain Vars	-
Encoding	-

`nmonths`i`_imputd_auto`	
Definition	Number of months within the last i months before the reference time period in which the target variable value z_k has been automatically imputed by a prefixed algorithm (monthly variation rate regression)
Stat Type	Numerical
Values	$0 - i$, $i = 3, 6, 12$ (three variables)
Example	2
Source	Internal-Derived
Formula	$\sum_{j=1}^{j=i} \delta(z_k^{(m-j)y,\text{imp}} = \text{auto})$, where δ is an indicator function and $z_k^{my,\text{imp}}$ is a paradata variable indicating the mode of imputation for the variable z of unit k in the reference time period with month m and year y
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Unit
Time Periods	$\{-1, -2, \ldots, -i\}$
Long/Cross	Long
Cross-Domain Vars	-
Encoding	-

`nmonths`i`_trnover0`	
Definition	Number of months within the last i months before the reference time period in which the target variable value z_k is 0 ($z_k = 0$)
Stat Type	Numerical
Values	$0 - i$, $i = 3, 6, 12$ (three variables)
Example	2
Source	Internal-Derived
Formula	$\sum_{j=1}^{j=i} \delta(z_k^{(m-j)y} = 0)$, where δ is an indicator function and $z_k^{my,\text{val}}$ is the validated variable value of unit k in the reference time period with month m and year y
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Unit
Time Periods	$\{-1, -2, \ldots, -i\}$
Long/Cross	Long
Cross-Domain Vars	-
Encoding	-

`nmonths13_notinsample`	
Definition	Number of months within the last 13 months before de reference time period in which the unit is not in the sample of the survey
Stat Type	Numerical
Values	$0 - 13$
Example	3
Source	Internal-Derived
Formula	$\sum_{j=1}^{j=13} \delta(k \notin U_c^{(m-j)y})$, where δ is an indicator function and U_c^{my} is the cut-off population frame in the reference time period with month m and year y
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Unit
Time Periods	$\{-1, -2, \ldots, -13\}$
Long/Cross	Long
Cross-Domain Vars	-
Encoding	-

Economic Activity Variables

code_NACE2class_frame_ed	
Definition	Edited value of the NACE Rev. 2 code for the class of the industrial establishment for the reference time period according to the population frame
Stat Type	Categorical
Values	[NACE Rev. 2 4-digit code]
Example	0510
Source	Internal-Primary
Formula	-
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Unit
Time Periods	{0}
Long/Cross	-
Cross-Domain Vars	-
Encoding	Mean

code_NACE2class_frame_ent_ed	
Definition	Edited value of the NACE Rev. 2 code for the class of the enterprise owning the industrial establishment for the reference time period according to the population frame
Stat Type	Categorical
Values	[NACE Rev. 2 4-digit code]
Example	0510
Source	Internal-Primary
Formula	-
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Unit
Time Periods	{0}
Long/Cross	-
Cross-Domain Vars	-
Encoding	Mean

code_NACE2class_ed	
Definition	Edited value of the NACE Rev. 2 code for the class of the industrial establishment for the reference time period according to the data collection process
Stat Type	Categorical
Values	[NACE Rev. 2 4-digit code]
Example	0510
Source	Internal-Primary
Formula	-
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Unit
Time Periods	{0}
Long/Cross	Long
Cross-Domain Vars	-
Encoding	Mean

code_NACE2class_val_1	
Definition	Validated value of the NACE Rev. 2 code for the class of the industrial establishment for the preceding reference time period
Stat Type	Categorical
Values	[NACE Rev. 2 4-digit code]
Example	0510
Source	Internal-Primary
Formula	-
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Unit
Time Periods	$\{-1\}$
Long/Cross	Long
Cross-Domain Vars	-

Encoding	Mean

`match_NACE2class_ed_val_1`	
Definition	Binary variable indicating whether `code_NACE2class_ed` is equal to `code_NACE2class_val_1` or not
Stat Type	Numerical
Values	0-1
Example	0
Source	Internal-Derived
Formula	$\delta(\text{code\_NACE2class\_ed}_k = \text{code\_NACE2class\_val\_1}_k)$, where δ denotes an indicator function
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Unit
Time Periods	$\{0, -1\}$
Long/Cross	-
Cross-Domain Vars	-
Encoding	-

`code_NACE2group_ed`	
Definition	Edited value of the NACE Rev. 2 code for the group of the industrial establishment for the reference time period
Stat Type	Categorical
Values	[NACE Rev. 2 3-digit code]
Example	051
Source	Internal-Primary
Formula	-
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Unit
Time Periods	$\{0\}$
Long/Cross	-
Cross-Domain Vars	-
Encoding	Mean

`code_NACE2group_ent_val_1`	
Definition	Validated value of the NACE Rev. 2 code for the group of the enterprise owning the industrial establishment for the preceding reference time period
Stat Type	Categorical
Values	[NACE Rev. 2 3-digit code]
Example	051
Source	Internal-Primary
Formula	-
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Unit
Time Periods	$\{-1\}$
Long/Cross	-
Cross-Domain Vars	-
Encoding	Mean

`match_NACE2group_est_ent_1`	
Definition	Binary variable indicating whether `code_NACE2group_ed` is equal to `code_NACE2group_val_1` or not
Stat Type	Categorical
Values	0, 1
Example	0
Source	Internal-Derived
Formula	$\delta(\text{code\_NACE2group\_ed}_k = \text{code\_NACE2group\_val\_1}_k)$, where δ denotes an indicator function
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Unit
Time Periods	$\{0, -1\}$
Long/Cross	-
Cross-Domain Vars	-

Encoding	-

`code_NACE2div_ed`	
Definition	Edited value of the NACE Rev. 2 code for the division of the industrial establishment for the reference time period; specific aggrupations of some divisions for ITI (for example division 10 is divided in 10A and 10B) is used.
Stat Type	Categorical
Values	[NACE Rev. 2 2-digit code (+ A|B)]
Example	15, 10*A*
Source	Internal-Primary
Formula	-
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Unit
Time Periods	{0}
Long/Cross	-
Cross-Domain Vars	-
Encoding	Mean

`code_NACE2div_ent_val_1`	
Definition	Validated value of the NACE Rev. 2 code for the division of the enterprise owning the industrial establishment for the preceding reference time period; specific aggrupations of some divisions for ITI (for example division 10 is divided in 10A and 10B) are used
Stat Type	Categorical
Values	[NACE Rev. 2 2-digit code (+ A|B)]
Example	14, 10*A*
Source	Internal-Primary
Formula	-
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Unit
Time Periods	$\{-1\}$
Long/Cross	-
Cross-Domain Vars	-
Encoding	Mean

`match_NACE2div_est_ent_1`	
Definition	Binary variable indicating whether `code_NACE2div_ed` is equal to `code_NACE2div_val_1` or not
Stat Type	Categorical
Values	0, 1
Example	0
Source	Internal-Derived
Formula	$\delta(\text{code\_NACE2div\_ed}_k = \text{code\_NACE2div\_val\_1}_k)$, where δ denotes an indicator function
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Unit
Time Periods	$\{0, -1\}$
Long/Cross	-
Cross-Domain Vars	-
Encoding	-

`code_NACE2sect_ed`	
Definition	Edited value of the NACE Rev. 2 code for the section of the industrial establishment for the reference time period
Stat Type	Categorical
Values	[NACE Rev. 2 character code]
Example	*C*
Source	Internal-Primary
Formula	-
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Unit
Time Periods	{0}

Long/Cross	-
Cross-Domain Vars	-
Encoding	One-hot

`code_NACE2sect_ent_val_1`	
Definition	Validated value of the NACE Rev. 2 code for the section of the enterprise owning the industrial establishment for the preceding reference time period
Stat Type	Categorical
Values	[NACE Rev. 2 1-digit code]
Example	*C*
Source	Internal-Primary
Formula	-
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Unit
Time Periods	$\{-1\}$
Long/Cross	-
Cross-Domain Vars	-
Encoding	One-hot

`match_NACE2sect_est_ent_1`	
Definition	Binary variable indicating whether `code_NACE2sect_ed` is equal to `code_NACE2sect_val_1` or not
Stat Type	Categorical
Values	0, 1
Example	0
Source	Internal-Derived
Formula	$\delta(\text{code\_NACE2sect\_ed}_k = \text{code\_NACE2sect\_val\_1}_k)$, where δ denotes an indicator function
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Unit
Time Periods	$\{0, -1\}$
Long/Cross	-
Cross-Domain Vars	-
Encoding	-

`code_NACE2MIG_ed`	
Definition	Edited value of the code for the Eurostat Main Industrial Grouping (MIGS) of the industrial establishment for the reference time period. If some 3-digit code is not in the standard MIGS classification, the 3-digit code is used directly.
Stat Type	Categorical
Values	[NACE Rev. 2 length-2 character code]
Example	*BC*
Source	Internal-Primary
Formula	-
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Unit
Time Periods	$\{0\}$
Long/Cross	-
Cross-Domain Vars	-
Encoding	One-hot

`code_NACE2MIG_ent_val_1`	
Definition	Validated value of the code for the Eurostat Main Industrial Grouping (MIGS) of the enterprise owning the industrial establishment for the preceding reference time period. If some 3-digit code is not in the standard MIGS classification, the 3-digit code is used directly.
Stat Type	Categorical
Values	[NACE Rev. 2 2-digit code]
Example	*BC*
Source	Internal-Primary
Formula	-
Stat Progr Ref	Spanish IOE-30052

Unit/Aggr	Unit
Time Periods	$\{-1\}$
Long/Cross	-
Cross-Domain Vars	-
Encoding	Mean or One-hot

Target-Related Variables

`trnovr_val_i`	
Definition	Validated total turnover value of the industrial establishment for the reference time period with month $m-i$ and year y
Stat Type	Numerical
Values	$\mathbb{R}$, $i = 1, \ldots, 12$ (12 variables)
Example	230000
Source	Internal-Primary
Formula	-
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Unit
Time Periods	$\{-i\}$
Long/Cross	-
Cross-Domain Vars	-
Encoding	-

`MA`i`_trnovr_val`	
Definition	Moving average of the validated total turnover values of the industrial establishment for the reference time periods within the last i months before the reference time period
Stat Type	Numerical
Values	$\mathbb{R}$, $i = 3, 6, 12$ (3 variables)
Example	230000
Source	Internal-Derived
Formula	$\frac{1}{i}\sum_{j=1}^{j=i} z_k^{(m-j)y,\text{val}}$
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Unit
Time Periods	$\{-1, \ldots, -i\}$
Long/Cross	Long
Cross-Domain Vars	-
Encoding	-

`q95_MA`i`trnovr_val_NACE2div`	
Definition	Quantiles 0.95 of the variable `MA`i`_trnovr_val` across the population domain defined by edited values of variable `code_NACE2div_ed` (NACE Rev. 2 division) of the industrial establishment from the reference time period
Stat Type	Numerical
Values	$\mathbb{R}$, $i = 3, 6, 12$ (3 variables)
Example	150000
Source	Internal-Derived
Formula	$Q_{0.95}^{\text{NACE2div}}(\text{MA}i(z_k^{my,\text{val}}))$, where $\text{MA}i(z_k^{my,\text{val}}) = \frac{1}{i}\sum_{j=1}^{i} z_k^{(m-j)y,\text{val}}$
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Aggr
Time Periods	$\{-1, \ldots, -i\}$
Long/Cross	Long + Cross
Cross-Domain Vars	code_NACE2div_ed
Encoding	-

`q95_MA`i`trnovr_val_NUTS2NACE2divEnt`

Definition	Quantiles 0.95 of the variable `MA`i`_trnovr_val` across the population domain defined by variables `code_NACE2div_ent_val_1` (NACE Rev. 2 division) of the enterprise owning the establishment and `code_NUTS2_val_1` (NUTS2 territorial unit) both from the preceding reference time period; validated values
Stat Type	Numerical
Values	$\mathbb{R}$, $i = 3, 6, 12$ (3 variables)
Example	75000
Source	Internal-Derived
Formula	$Q_{0.95}^{\text{NUTS2NACE2divEnt}}(\text{MA}i(z_k^{my,\text{val}}))$, where $\text{MA}i(z_k^{my,\text{val}}) = \frac{1}{i}\sum_{j=1}^{i} z_k^{(m-j)y,\text{val}}$
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Aggr
Time Periods	$\{-1, \ldots, -i\}$
Long/Cross	Long + Cross
Cross-Domain Vars	code_NACE2div_ent_val_1, code_NUTS2_val_1
Encoding	-

`q95_MA`i`trnovr_val_NUTS2NACE2div`

Definition	Quantiles 0.95 of the variable `MA`i`_trnovr_val` across the population domain defined by edited values of variable `code_NACE2div_ed` (NACE Rev. 2 division) of the establishment from the reference time period and validated values of variable `code_NUTS2_val_1` (NUTS2 territorial unit) of the industrial establishment from the preceding reference time period
Stat Type	Numerical
Values	$\mathbb{R}$, $i = 3, 6, 12$ (3 variables)
Example	75000
Source	Internal-Derived
Formula	$Q_{0.95}^{\text{NUTS2NACE2div}}(\text{MA}i(z_k^{my,\text{val}}))$, where $\text{MA}i(z_k^{my,\text{val}}) = \frac{1}{i}\sum_{j=1}^{i} z_k^{(m-j)y,\text{val}}$
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Aggr
Time Periods	$\{-1, \ldots, -i\}$
Long/Cross	Long + Cross
Cross-Domain Vars	code_NACE2div_ed, code_NUTS2_val_1
Encoding	-

`above_q95_MA`i`trnovr_val_NACE2div`

Definition	Binary variable indicating whether `MA`i`_trnovr_val` is greater or equal to `q95_MA`i`trnovr_val_NACE2div` or not
Stat Type	Numerical
Values	0-1, $i = 3, 6, 12$ (3 variables)
Example	0
Source	Internal-Derived
Formula	$\delta(\text{MA}i(z_k^{my,\text{val}}) \geq Q_{0.95}^{\text{NACE2div}}(\text{MA}i(z_k^{my,\text{val}}))$, where δ denotes an indicator function
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Unit
Time Periods	$\{-1, \ldots, -i\}$
Long/Cross	Long + Cross
Cross-Domain Vars	code_NACE2div_ed
Encoding	-

`above_q95_MA`i`trnovr_val_NUTS2NACE2divEnt`

Definition	Binary variable indicating whether `MA`i`_trnovr_val` is greater or equal to `q95_MA`i`trnovr_NUTS2NACE2divEnt_val` or not
Stat Type	Numerical
Values	0-1, $i = 3, 6, 12$ (3 variables)
Example	0
Source	Internal-Derived
Formula	$\delta(\text{MA}i(z_k^{my,\text{val}}) \geq Q_{0.95}^{\text{NUTS2NACE2divEnt}}(\text{MA}i(z_k^{my,\text{val}}))$,

	where δ denotes an indicator function
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Unit
Time Periods	$\{-1, \ldots, -i\}$
Long/Cross	Long + Cross
Cross-Domain Vars	code_NACE2div_ent_val_1, code_NUTS2_val_1
Encoding	-

`above_q95_MA`i`trnovr_val_NUTS2NACE2div`	
Definition	Binary variable indicating whether `MA`i`_trnovr_val` is greater or equal to `q95_MA`i`trnovr_NUTS2NACE2div_val` or not
Stat Type	Numerical
Values	0-1, $i = 3, 6, 12$ (3 variables)
Example	0
Source	Internal-Derived
Formula	$\delta(\text{MA}i(z_k^{my,\text{val}}) \geq Q_{0.95}^{\text{NUTS2NACE2div}}(\text{MA}i(z_k^{my,\text{val}}))$, where δ denotes an indicator function
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Unit
Time Periods	$\{-1, \ldots, -i\}$
Long/Cross	Long + Cross
Cross-Domain Vars	code_NACE2div_ed, code_NUTS2_val_1
Encoding	-

`prob_trnovr_val_`i`_MA12_NACE2div`	
Definition	Validated value of the empirical cumulative distribution function generated by variable MA12_NACE2div (12-month moving average of turnover by NACE Rev. 2 division of the industrial establishment) at the validated turnover $z_k^{(m-i)y,\text{val}}$ from i time periods before the reference time period of the industrial turnover
Stat Type	Numerical
Values	$[0, 1]$, $i = 1, 3, 6, 12$ (4 variables)
Example	0.8
Source	Internal-Derived
Formula	$F^*_{\text{MA12\_NACE2div}}(z_k^{(m-i)y,\text{val}})$
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Aggr
Time Periods	$\{-1, \ldots, -i\}$
Long/Cross	Long + Cross
Cross-Domain Vars	code_NACE2div_ed
Encoding	-

`prob_trnovr_val_`i`_MA12_NUTS2NACE2divEnt`	
Definition	Validated value of the empirical cumulative distribution function generated by variable MA12_NUTS2NACE2divEnt (12-month moving average of turnover by NUTS2 of the industrial establishment and NACE Rev. 2 division of the enterprise owning the enterprise) at the validated turnover $z_k^{(m-i)y,\text{val}}$ from i time periods before the reference time period of the industrial establishment
Stat Type	Numerical
Values	$[0, 1]$, $i = 1, 3, 6, 12$ (4 variables)
Example	0.8
Source	Internal-Derived
Formula	$F^*_{\text{MA12\_NUTS2NACE2divEnt}}(z_k^{(m-i)y,\text{val}})$
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Aggr
Time Periods	$\{-1, \ldots, -i\}$
Long/Cross	Long + Cross
Cross-Domain Vars	code_NACE2div_ent_val_1, code_NUTS2_val_1
Encoding	-

`prob_trnovr_val_`i`_MA12_NUTS2NACE2div`

Definition	Validated value of the empirical cumulative distribution function generated by variable MA12_NUTS2NACE2div (12-month moving average of turnover by NUTS2 and NACE Rev. 2 division of the industrial establishment) at the validated turnover $z_k^{(m-i)y,\text{val}}$ from i time periods before the reference time period of the industrial turnover; validated values
Stat Type	Numerical
Values	$[0, 1]$, $i = 1, 3, 6, 12$ (4 variables)
Example	0.8
Source	Internal-Derived
Formula	$F^*_{\text{MA12\_NUTS2NACE2div}}(z_k^{(m-i)y,\text{val}})$
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Aggr
Time Periods	$\{-1, \ldots, -i\}$
Long/Cross	Long + Cross
Cross-Domain Vars	code_NACE2div_ed, code_NUTS2_val_1
Encoding	-

`cv_sd`i`trnovr_val_MA`i`trnovr_val`

Definition	Coefficient of variation obtained with the standard deviation of the validated total turnover values $z_k^{(m-j)y,\text{val}})$ of the industrial establishment for the reference time periods within the last i months before the reference time period $(j = 1, \ldots, i)$ and the moving average of the same values
Stat Type	Numerical
Values	$\mathbb{R}$, $i = 3, 6, 12$ (3 variables)
Example	0.9
Formula	$\frac{\sqrt{\frac{1}{i-1}\sum_{j=1}^{j=i}(z_k^{(m-j)y,\text{val}} - \text{MA}i(z_k^{my,\text{val}}))^2}}{\text{MA}i(z_k^{my,\text{val}})}$, where $\text{MA}i(z_k^{my,\text{val}}) = \frac{1}{i}\sum_{j=1}^{j=i} z_k^{(m-j)y,\text{valb}}$
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Unit
Time Periods	$\{-1, \ldots, -i\}$
Long/Cross	Long
Cross-Domain Vars	-
Encoding	-

`min_trnovr_val_`i

Definition	Minimum value of the validated total turnover values $z_k^{(m-j)y,\text{val}})$ of the industrial establishment for the reference time periods within the last i months before the reference time period $(j = 1, \ldots, i)$
Stat Type	Numerical
Values	$\mathbb{R}$, $i = 3, 6, 12$ (3 variables)
Example	150000
Source	Internal-Derived
Formula	$min(z_k^{(m-1)y,\text{val}}, \ldots, z_k^{(m-i)y,\text{val}})$
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Unit
Time Periods	$\{-1, \ldots, -i\}$
Long/Cross	Long
Cross-Domain Vars	-
Encoding	-

`max_trnovr_val_`i

Definition	Maximum value of the validated total turnover values $z_k^{(m-j)y,\text{val}}$ of the industrial establishment for the reference time periods within the last i months before the reference time period $(j = 1, \ldots, i)$
Stat Type	Numerical
Values	$\mathbb{R}$, $i = 3, 6, 12$ (3 variables)
Example	150000
Source	Internal-Derived

Formula	$max(z_k^{(m-1)y,\text{val}}, \ldots, z_k^{(m-i)y,\text{val}})$
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Unit
Time Periods	$\{-1, \ldots, -i\}$
Long/Cross	Long
Cross-Domain Vars	-
Encoding	-

`mean_trnovr_ed_NACE2group`	
Definition	Mean value of the edited total turnover values $z_k^{my,\text{ed}}$ of the responding industrial establishments at time t for the reference time period across the domain defined by variable `code_NACE2group_ed` (NACE Rev. 2 group)
Stat Type	Numerical
Values	$\mathbb{R}$
Example	150000
Source	Internal-Derived
Formula	$\frac{1}{N_{r_d(t)}} \sum_{k \in r_d(t)} z_k^{my,\text{ed}}$, where $d \in$ NACE2group
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Aggr
Time Periods	$\{0\}$
Long/Cross	Cross
Cross-Domain Vars	`code_NACE2group_ed`
Encoding	-

`mean_trnovr_ed_NUTS2NACE2group`	
Definition	Mean value of the edited total turnover values $z_k^{my,\text{ed}}$ of the responding industrial establishments at time t for the reference time period across the domain defined by variables `code_NACE2group_ed` (NACE Rev. 2 group) and `code_NUTS2_val_1` (NUTS2 territorial unit)
Stat Type	Numerical
Values	$\mathbb{R}$
Example	150000
Source	Internal-Derived
Formula	$\frac{1}{N_{r_d(t)}} \sum_{k \in r_d(t)} z_k^{my,\text{ed}}$, where $d \in$ (NUTS2, NACE2group)
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Aggr
Time Periods	$\{0\}$
Long/Cross	Cross
Cross-Domain Vars	`code_NACE2group_ed`, `code_NUTS2_val_1`
Encoding	-

`mean_trnovr_ed_NACE2div`	
Definition	Mean value of the edited total turnover values $z_k^{my,\text{ed}}$ of the responding industrial establishments at time t for the reference time period across the domain defined by variable `code_NACE2div_ed` (NACE Rev. 2 division)
Stat Type	Numerical
Values	$\mathbb{R}$
Example	150000
Source	Internal-Derived
Formula	$\frac{1}{N_{r_d(t)}} \sum_{k \in r_d(t)} z_k^{my,\text{ed}}$, where $d \in$ NACE2div
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Aggr
Time Periods	$\{0\}$
Long/Cross	Cross
Cross-Domain Vars	`code_NACE2div_ed`
Encoding	-

`mean_trnovr_ed_NUTS2NACE2div`

Definition	Mean value of the edited total turnover values $z_k^{my,\text{ed}}$ of the responding industrial establishments at time t for the reference time period across the domain defined by variables `code_NACE2div_ed` (NACE Rev. 2 division) and `code_NUTS2_val_1` (NUTS2 territorial unit)
Stat Type	Numerical
Values	$\mathbb{R}$
Example	150000
Source	Internal-Derived
Formula	$\frac{1}{N_{r_d(t)}} \sum_{k \in r_d(t)} z_k^{my,\text{ed}}$, where $d \in (\text{NUTS2}, \text{NACE2div})$
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Aggr
Time Periods	$\{0\}$
Long/Cross	Cross
Cross-Domain Vars	`code_NACE2div_ed`, `code_NUTS2_val_1`
Encoding	-

`mean_trnovr_ed_NACE2section`	
Definition	Mean value of the edited total turnover values $z_k^{my,\text{ed}}$ of the responding industrial establishments at time t for the reference time period across the domain defined by variable `code_NACE2section_ed` (NACE Rev. 2 section)
Stat Type	Numerical
Values	$\mathbb{R}$
Example	150000
Source	Internal-Derived
Formula	$\frac{1}{N_{r_d(t)}} \sum_{k \in r_d(t)} z_k^{my,\text{ed}}$, where $d \in \text{NACE2section}$
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Aggr
Time Periods	$\{0\}$
Long/Cross	Cross
Cross-Domain Vars	`code_NACE2section_ed`
Encoding	-

`mean_trnovr_ed_NUTS2NACE2section`	
Definition	Mean value of the edited total turnover values $z_k^{my,\text{ed}}$ of the responding industrial establishments at time t for the reference time period across the domain defined by variables `code_NACE2section_ed` (NACE Rev. 2 section) and `code_NUTS2_val_1` (NUTS2 territorial unit)
Stat Type	Numerical
Values	$\mathbb{R}$
Example	150000
Source	Internal-Derived
Formula	$\frac{1}{N_{r_d(t)}} \sum_{k \in r_d(t)} z_k^{my,\text{ed}}$, where $d \in (\text{NUTS2}, \text{NACE2section})$
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Aggr
Time Periods	$\{0\}$
Long/Cross	Cross
Cross-Domain Vars	`code_NACE2section_ed`, `code_NUTS2_val_1`
Encoding	-

`mean_trnovr_ed_NACE2class`	
Definition	Mean value of the edited total turnover values $z_k^{my,\text{ed}}$ of the responding industrial establishments at time t for the reference time period across the domain defined by variable `code_NACE2class_val_1` (NACE Rev. 2 class)
Stat Type	Numerical
Values	$\mathbb{R}$
Example	150000
Source	Internal-Derived
Formula	$\frac{1}{N_{r_d(t)}} \sum_{k \in r_d(t)} z_k^{my,\text{ed}}$, where $d \in \text{NACE2class}$
Stat Progr Ref	Spanish IOE-30052

Unit/Aggr	Aggr
Time Periods	$\{-1\}$
Long/Cross	Long + Cross
Cross-Domain Vars	`code_NACE2class_val_1`
Encoding	-

`mean_trnovr_ed_NUTS2NACE2class`	
Definition	Mean value of the edited total turnover values $z_k^{my,\text{ed}}$ of the responding industrial establishments at time t for the reference time period across the domain defined by variables `code_NACE2class_val_1` (NACE Rev. 2 class) and `code_NUTS2_val_1` (NUTS2 territorial unit)
Stat Type	Numerical
Values	$\mathbb{R}$
Example	150000
Source	Internal-Derived
Formula	$\frac{1}{N_{r_d(t)}} \sum_{k \in r_d(t)} z_k^{my,\text{ed}}$, where $d \in (\text{NUTS2}, \text{NACE2class})$
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Aggr
Time Periods	$\{-1\}$
Long/Cross	Long + Cross
Cross-Domain Vars	`code_NACE2class_val_1`, `code_NUTS2_val_1`
Encoding	-

`sd_trnovr_ed_NACE2group`	
Definition	Standard deviation of the edited total turnover values $z_k^{my,\text{ed}}$ of the responding industrial establishments at time t for the reference time period across the domain defined by variables `code_NACE2group_ed` (NACE Rev. 2 group)
Stat Type	Numerical
Values	$\mathbb{R}$
Example	150000
Source	Internal-Derived
Formula	$\sqrt{\frac{1}{N_{r_d(t)}-1} \sum_{k \in r_d(t)} \left(z_k^{my,\text{ed}} - \bar{z}_k^{my,\text{ed}}\right)^2}$, where $d \in$ NACE2group and $\bar{z}_k^{my,\text{ed}} = \frac{1}{N_{r_d(t)}} \sum_{k \in r_d(t)} z_k^{my,\text{ed}}$
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Aggr
Time Periods	$\{0\}$
Long/Cross	Cross
Cross-Domain Vars	`code_NACE2group_ed`
Encoding	-

`sd_trnovr_ed_NUTS2NACE2group`	
Definition	Standard deviation of the edited total turnover values $z_k^{my,\text{ed}}$ of the responding industrial establishments at time t for the reference time period across the domain defined by variables `code_NACE2group_ed` (NACE Rev. 2 group) and `code_NUTS2_val_1` (NUTS2 territorial unit)
Stat Type	Numerical
Values	$\mathbb{R}$
Example	150000
Source	Internal-Derived
Formula	$\sqrt{\frac{1}{N_{r_d(t)}-1} \sum_{k \in r_d(t)} \left(z_k^{my,\text{ed}} - \bar{z}_k^{my,\text{ed}}\right)^2}$, where $d \in (\text{NUTS2}, \text{NACE2group})$ and $\bar{z}_k^{my,\text{ed}} = \frac{1}{N_{r_d(t)}} \sum_{k \in r_d(t)} z_k^{my,\text{ed}}$
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Aggr
Time Periods	$\{0\}$
Long/Cross	Cross
Cross-Domain Vars	`code_NACE2group_ed`, `code_NUTS2_val_1`
Encoding	-

`sd_trnovr_ed_NACE2div`

Definition	Standard deviation of the edited total turnover values $z_k^{my,\text{ed}}$ of the responding industrial establishments at time t for the reference time period across the domain defined by variable `code_NACE2div_ed` (NACE Rev. 2 division)
Stat Type	Numerical
Values	$\mathbb{R}$
Example	150000
Source	Internal-Derived
Formula	$\sqrt{\frac{1}{N_{r_d(t)}-1}\sum_{k\in r_d(t)}\left(z_k^{my,\text{ed}}-\bar{z}_k^{my,\text{ed}}\right)^2}$, where $d \in$ NACE2div and $\bar{z}_k^{my,\text{ed}} = \frac{1}{N_{r_d(t)}}\sum_{k\in r_d(t)} z_k^{my,\text{ed}}$
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Aggr
Time Periods	{0}
Long/Cross	Cross
Cross-Domain Vars	`code_NACE2div_ed`
Encoding	-

`sd_trnovr_ed_NUTS2NACE2div`

Definition	Standard deviation of the edited total turnover values $z_k^{my,\text{ed}}$ of the responding industrial establishments at time t for the reference time period across the domain defined by variables `code_NACE2div_ed` (NACE Rev. 2 division) and `code_NUTS2_val_1` (NUTS2 territorial unit)
Stat Type	Numerical
Values	$\mathbb{R}$
Example	150000
Source	Internal-Derived
Formula	$\sqrt{\frac{1}{N_{r_d(t)}-1}\sum_{k\in r_d(t)}\left(z_k^{my,\text{ed}}-\bar{z}_k^{my,\text{ed}}\right)^2}$, where $d \in$ (NUTS2, NACE2div) and $\bar{z}_k^{my,\text{ed}} = \frac{1}{N_{r_d(t)}}\sum_{k\in r_d(t)} z_k^{my,\text{ed}}$
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Aggr
Time Periods	{0}
Long/Cross	Cross
Cross-Domain Vars	`code_NACE2div_ed`, `code_NUTS2_val_1`
Encoding	-

`sd_trnovr_ed_NACE2section`

Definition	Standard deviation of the edited total turnover values $z_k^{my,\text{ed}}$ of the responding industrial establishments at time t for the reference time period across the domain defined by variable `code_NACE2section_ed` (NACE Rev. 2 section)
Stat Type	Numerical
Values	$\mathbb{R}$
Example	150000
Source	Internal-Derived
Formula	$\sqrt{\frac{1}{N_{r_d(t)}-1}\sum_{k\in r_d(t)}\left(z_k^{my,\text{ed}}-\bar{z}_k^{my,\text{ed}}\right)^2}$, where $d \in$ NACE2section and $\bar{z}_k^{my,\text{ed}} = \frac{1}{N_{r_d(t)}}\sum_{k\in r_d(t)} z_k^{my,\text{ed}}$
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Aggr
Time Periods	{0}
Long/Cross	Cross
Cross-Domain Vars	`code_NACE2section_ed`
Encoding	-

`sd_trnovr_ed_NUTS2NACE2section`

Definition	Standard deviation of the edited total turnover values $z_k^{my,\text{ed}}$ of the responding industrial establishments at time t for the reference time period across the domain defined by variables `code_NACE2section_ed` (NACE Rev. 2 section) and `code_NUTS2_val_1` (NUTS2 territorial unit)
Stat Type	Numerical

Values	$\mathbb{R}$
Example	150000
Source	Internal-Derived
Formula	$\sqrt{\frac{1}{N_{r_d(t)}-1}\sum_{k\in r_d(t)}\left(z_k^{my,\text{ed}}-\bar{z}_k^{my,\text{ed}}\right)^2}$, where $d \in$ (NUTS2, NACE2section) and $\bar{z}_k^{my,\text{ed}}=\frac{1}{N_{r_d(t)}}\sum_{k\in r_d(t)} z_k^{my,\text{ed}}$
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Aggr
Time Periods	$\{0\}$
Long/Cross	Cross
Cross-Domain Vars	`code_NACE2section_ed`, `code_NUTS2_val_1`
Encoding	-

`sd_trnovr_ed_NACE2class`

Definition	Standard deviation of the edited total turnover values $z_k^{my,\text{ed}}$ of the responding industrial establishments at time t for the reference time period across the domain defined by variable `code_NACE2class_ed` (NACE Rev. 2 class)
Stat Type	Numerical
Values	$\mathbb{R}$
Example	150000
Source	Internal-Derived
Formula	$\sqrt{\frac{1}{N_{r_d(t)}-1}\sum_{k\in r_d(t)}\left(z_k^{my,\text{ed}}-\bar{z}_{r_d(t)}^{my,\text{ed}}\right)^2}$, where $d \in$ NACE2class and $\bar{z}_k^{my,\text{ed}}=\frac{1}{N_{r_d(t)}}\sum_{k\in r_d(t)} z_k^{my,\text{ed}}$
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Aggr
Time Periods	$\{-1\}$
Long/Cross	Long + Cross
Cross-Domain Vars	`code_NACE2class_val_1`
Encoding	-

`sd_trnovr_ed_NUTS2NACE2class`

Definition	Standard deviation of the edited total turnover values $z_k^{my,\text{ed}}$ of the responding industrial establishments at time t for the reference time period across the domain defined by variables `code_NACE2class_ed` (NACE Rev. 2 class) and `code_NUTS2_val_1` (NUTS2 territorial unit)
Stat Type	Numerical
Values	$\mathbb{R}$
Example	150000
Source	Internal-Derived
Formula	$\sqrt{\frac{1}{N_{r_d(t)}-1}\sum_{k\in r_d(t)}\left(z_k^{my,\text{ed}}-\bar{z}_{r_d(t)}^{my,\text{ed}}\right)^2}$, where $d \in$ (NUTS2, NACE2class) and $\bar{z}_k^{my,\text{ed}}=\frac{1}{N_{r_d(t)}}\sum_{k\in r_d(t)} z_k^{my,\text{ed}}$
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Aggr
Time Periods	$\{-1\}$
Long/Cross	Long + Cross
Cross-Domain Vars	`code_NACE2class_val_1`, `code_NUTS2_val_1`

`rate_trnovr_ed0_vali`

Definition	Relative variation rate between the edited total turnover value from the reference time period and the validated total turnover value from the ith preceding reference time period
Stat Type	Numerical
Values	$\mathbb{R}$, $i = 1, 3, 6, 12$ (4 variables)
Example	0.91
Source	Internal-Derived
Formula	$\frac{z_k^{my}-z_k^{(m-i)y}}{z_k^{(m-i)y}}$
Stat Progr Ref	Spanish IOE-30052

Unit/Aggr	Unit
Time Periods	$\{0, -i\}$
Long/Cross	Long
Cross-Domain Vars	-
Encoding	-

`rate_trnovr_val1_vali`	
Definition	Relative variation rate between the validated total turnover value from the preceding reference time period and the validated total turnover value from the ith preceding reference time period
Stat Type	Numerical
Values	$\mathbb{R}$, $i = 2, 4, 7, 13$ (4 variables)
Example	0.11
Source	Internal-Derived
Formula	$\frac{z_k^{(m-1)y} - z_k^{(m-i)y}}{z_k^{(m-i)y}}$
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Unit
Time Periods	$\{1, -i\}$
Long/Cross	Long
Cross-Domain Vars	-
Encoding	-

`rate_meanTrnovr_ed0_vali_NUTS2NACE2div`	
Definition	Relative variation rate between the mean value of edited total turnover values from the reference time period and the mean value of validated total turnover values from the ith preceding reference time period of the industrial establishments across domains defined by variables `code_NACE2div_ed` (NACE Rev. 2 division) and `code_NUTS2_val_1` (NUTS2 territorial unit)
Stat Type	Numerical
Values	$\mathbb{R}$, $i = 1, 3, 6, 12$ (3 variables)
Example	0.11
Source	Internal-Derived
Formula	$\frac{\frac{1}{r_d(t)} \sum_{k \in d} z_k^{my,\text{ed}} - \frac{1}{N_d} \sum_{k \in d} z_k^{(m-i)y,\text{val}}}{\frac{1}{N_d} \sum_{k \in d} z_k^{(m-i)y,\text{val}}}$ where $d \in (\text{NUTS2}, \text{NACE2div})$
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Aggr
Time Periods	$\{0, -i\}$
Long/Cross	Long + Cross
Cross-Domain Vars	`code_NACE2div_ed`, `code_NUTS2_val_1`
Encoding	-

`rate_meanTrnovr_val1_vali_NUTS2NACE2div`	
Definition	Relative variation rate between the mean value of validated total turnover values from the preceding reference time period and the mean value of validated total turnover values from the ith preceding reference time period of the industrial establishments across domains defined by variables `code_NACE2div_ed` (NACE Rev. 2 division) and `code_NUTS2_val_1` (NUTS2 territorial unit)
Stat Type	Numerical
Values	$\mathbb{R}$, $i = 2, 4, 7, 13$ (3 variables)
Example	0.11
Source	Internal-Derived
Formula	$\frac{\frac{1}{N_d} \sum_{k \in d} z_k^{(m-1)y,\text{val}} - \frac{1}{N_d} \sum_{k \in d} z_k^{(m-i)y,\text{val}}}{\frac{1}{N_d} \sum_{k \in d} z_k^{(m-i)y,\text{val}}}$ where $d \in (\text{NUTS2}, \text{NACE2div})$
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Aggr
Time Periods	$\{1, -i\}$
Long/Cross	Long + Cross
Cross-Domain Vars	`code_NACE2div_ed`, `code_NUTS2_val_1`
Encoding	-

rate_meanTrnovr_ed0_vali_NACE2div	
Definition	Relative variation rate between the mean value of edited total turnover values from the reference time period and the mean value of validated total turnover values from the ith preceding reference time period of the industrial establishments across domains defined by variable `code_NACE2div_ed` (NACE Rev. 2 division)
Stat Type	Numerical
Values	$\mathbb{R}$, $i = 1, 3, 6, 12$ (3 variables)
Example	0.11
Source	Internal-Derived
Formula	$\frac{\frac{1}{r_d(t)} \sum_{k \in d} z_k^{my} - \frac{1}{N_d} \sum_{k \in d} z_k^{(m-i)y}}{\frac{1}{N_d} \sum_{k \in d} z_k^{(m-i)y}}$ where $d \in$ NACE2div
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Aggr
Time Periods	$\{0, -i\}$
Long/Cross	Long + Cross
Cross-Domain Vars	`code_NACE2div_ed`
Encoding	-

rate_meanTrnovr_val1_vali_NACE2div	
Definition	Relative variation rate between the mean value of validated total turnover values from the preceding reference time period and the mean value of validated total turnover values from the ith preceding reference time period of the industrial establishments across domains defined by variable `code_NACE2div_ed` (NACE Rev. 2 division)
Stat Type	Numerical
Values	$\mathbb{R}$, $i = 2, 4, 7, 13$ (3 variables)
Example	0.11
Source	Internal-Derived
Formula	$\frac{\frac{1}{N_d} \sum_{k \in d} z_k^{(m-1)y,\text{val}} - \frac{1}{N_d} \sum_{k \in d} z_k^{(m-i)y,\text{val}}}{\frac{1}{N_d} \sum_{k \in d} z_k^{(m-i)y,\text{val}}}$ where $d \in$ NACE2div
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Aggr
Time Periods	$\{1, -i\}$
Long/Cross	Long + Cross
Cross-Domain Vars	`code_NACE2div_ed`
Encoding	-

rate_meanTrnovr_ed0_vali_NACE2group	
Definition	Relative variation rate between the mean value of edited total turnover values from the reference time period and the mean value of validated total turnover values from the ith preceding reference time period of the industrial establishments across domains defined by variable `code_NACE2group_ed` (NACE Rev. 2 group)
Stat Type	Numerical
Values	$\mathbb{R}$, $i = 1, 3, 6, 12$ (3 variables)
Example	0.11
Source	Internal-Derived
Formula	$\frac{\frac{1}{r_d(t)} \sum_{k \in d} z_k^{my,\text{ed}} - \frac{1}{N_d} \sum_{k \in d} z_k^{(m-i)y,\text{val}}}{\frac{1}{N_d} \sum_{k \in d} z_k^{(m-i)y,\text{val}}}$ where $d \in$ NACE2group
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Aggr
Time Periods	$\{0, -i\}$
Long/Cross	Long + Cross
Cross-Domain Vars	`code_NACE2group_ed`
Encoding	-

rate_meanTrnovr_val1_vali_NACE2group

Definition	Relative variation rate between the mean value of validated total turnover values from the preceding reference time period and the mean value of validated total turnover values from the ith preceding reference time period of the industrial establishments across domains defined by variable `code_NACE2group_ed` (NACE Rev. 2 group)
Stat Type	Numerical
Values	$\mathbb{R}$, $i = 2, 4, 7, 13$ (3 variables)
Example	0.11
Source	Internal-Derived
Formula	$\frac{\frac{1}{N_d}\sum_{k\in d} z_k^{(m-1)y,\text{val}} - \frac{1}{N_d}\sum_{k\in d} z_k^{(m-i)y,\text{val}}}{\frac{1}{N_d}\sum_{k\in d} z_k^{(m-i)y,\text{val}}}$ where $d \in$ NACE2group
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Aggr
Time Periods	$\{1, -i\}$
Long/Cross	Long + Cross
Cross-Domain Vars	`code_NACE2group_ed`
Encoding	-

`rate_meanTrnovr_ed0_vali_NUTS2NACE2group`

Definition	Relative variation rate between the mean value of edited total turnover values from the reference time period and the mean value of validated total turnover values from the ith preceding reference time period of the industrial establishments across domains defined by variables `code_NACE2group_ed` (NACE Rev. 2 group) and `code_NUTS2_val_1` (NUTS2 territorial unit)
Stat Type	Numerical
Values	$\mathbb{R}$, $i = 1, 3, 6, 12$ (4 variables)
Example	0.11
Source	Internal-Derived
Formula	$\frac{\frac{1}{r_d(t)}\sum_{k\in d} z_k^{my,\text{ed}} - \frac{1}{N_d}\sum_{k\in d} z_k^{(m-i)y,\text{val}}}{\frac{1}{N_d}\sum_{k\in d} z_k^{(m-i)y,\text{val}}}$ where $d \in$ (NUTS2, NACE2group)
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Aggr
Time Periods	$\{0, -i\}$
Long/Cross	Long + Cross
Cross-Domain Vars	`code_NACE2group_ed`, code_NUTS2_val_1
Encoding	-

`rate_meanTrnovr_val1_vali_NUTS2NACE2group`

Definition	Relative variation rate between the mean value of validated total turnover values from the preceding reference time period and the mean value of validated total turnover values from the ith preceding reference time period of the industrial establishments across domains defined by variables `code_NACE2group_ed` (NACE Rev. 2 group) and `code_NUTS2_val_1` (NUTS2 territorial unit)
Stat Type	Numerical
Values	$\mathbb{R}$, $i = 2, 4, 7, 13$ (4 variables)
Example	0.11
Source	Internal-Derived
Formula	$\frac{\frac{1}{N_d}\sum_{k\in d} z_k^{(m-1)y,\text{val}} - \frac{1}{N_d}\sum_{k\in d} z_k^{(m-i)y,\text{val}}}{\frac{1}{N_d}\sum_{k\in d} z_k^{(m-i)y,\text{val}}}$ where $d \in$ (NUTS2, NACE2group)
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Aggr
Time Periods	$\{-1, -i\}$
Long/Cross	Long + Cross
Cross-Domain Vars	`code_NACE2group_ed`, code_NUTS2_val_1
Encoding	-

`rate_meanTrnovr_ed0_vali_NUTS2NACE2divEnt`

Definition	Relative variation rate between the mean value of edited total turnover values from the reference time period and the mean value of validated total turnover values from the ith preceding reference time period of the enterprises owning the industrial establishments across domains defined by variables `code_NACE2divEnt_ed` (NACE Rev. 2 division) and `code_NUTS2_val_1` (NUTS2 territorial unit)
Stat Type	Numerical
Values	$\mathbb{R}$, $i = 1, 3, 6, 12$ (4 variables)
Example	0.11
Source	Internal-Derived
Formula	$\frac{\frac{1}{r_d(t)} \sum_{k \in d} z_k^{my,\text{ed}} - \frac{1}{N_d} \sum_{k \in d} z_k^{(m-i)y,\text{val}}}{\frac{1}{N_d} \sum_{k \in d} z_k^{(m-i)y,\text{val}}}$ where $d \in (\text{NUTS2}, \text{NACE2divEnt})$
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Aggr
Time Periods	$\{0, -i\}$
Long/Cross	Long + Cross
Cross-Domain Vars `code_NACE2divEnt_ed`, code_NUTS2_val_1	
Encoding	-

`rate_meanTrnovr_val1_vali_NUTS2NACE2divEnt`

Definition	Relative variation rate between the mean value of validated total turnover values from the preceding reference time period and the mean value of validated total turnover values from the ith preceding reference time period of the enterprises owning the industrial establishments across domains defined by variables `code_NACE2divEnt_ed` (NACE Rev. 2 division) and `code_NUTS2_val_1` (NUTS2 territorial unit)
Stat Type	Numerical
Values	$\mathbb{R}$, $i = 2, 4, 7, 13$ (4 variables)
Example	0.11
Source	Internal-Derived
Formula	$\frac{\frac{1}{N_d} \sum_{k \in d} z_k^{(m-1)y,\text{val}} - \frac{1}{N_d} \sum_{k \in d} z_k^{(m-i)y,\text{val}}}{\frac{1}{N_d} \sum_{k \in d} z_k^{(m-i)y,\text{val}}}$ where $d \in (\text{NUTS2}, \text{NACE2divEnt})$
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Aggr
Time Periods	$\{-1, -i\}$
Long/Cross	Long + Cross
Cross-Domain Vars `code_NACE2divEnt_ed`, code_NUTS2_val_1	
Encoding	-

`rate_meanTrnovr_ed0_vali_NUTS2NACE2class`

Definition	Relative variation rate between the mean value of edited total turnover values from the reference time period and the mean value of validated total turnover values from the ith preceding reference time period of the industrial establishments across domains defined by variables `code_NACE2class_ed` (NACE Rev. 2 class) and `code_NUTS2_val_1` (NUTS2 territorial unit)
Stat Type	Numerical
Values	$\mathbb{R}$, $i = 1, 3, 6, 12$ (4 variables)
Example	0.11
Source	Internal-Derived
Formula	$\frac{\frac{1}{r_d(t)} \sum_{k \in d} z_k^{my,\text{ed}} - \frac{1}{N_d} \sum_{k \in d} z_k^{(m-i)y,\text{val}}}{\frac{1}{N_d} \sum_{k \in d} z_k^{(m-i)y,\text{val}}}$ where $d \in (\text{NUTS2}, \text{NACE2class})$
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Aggr
Time Periods	$\{0, -i\}$
Long/Cross	Long + Cross

Cross-Domain Vars	`code_NACE2class_val_1`, code_NUTS2_val_1
Encoding	-

`rate_meanTrnovr_val1_vali_NUTS2NACE2class`	
Definition	Relative variation rate between the mean value of validated total turnover values from the preceding reference time period and the mean value of validated total turnover values from the ith preceding reference time period of the industrial establishments across domains defined by variables `code_NACE2class_val_1` (NACE Rev. 2 class) and `code_NUTS2_val_1` (NUTS2 territorial unit)
Stat Type	Numerical
Values	$\mathbb{R}$, $i = 2, 4, 7, 13$ (4 variables)
Example	0.11
Source	Internal-Derived
Formula	$\frac{\frac{1}{N_d}\sum_{k\in d} z_k^{(m-1)y,\text{val}} - \frac{1}{N_d}\sum_{k\in d} z_k^{(m-i)y,\text{val}}}{\frac{1}{N_d}\sum_{k\in d} z_k^{(m-i)y,\text{val}}}$ where $d \in$ (NUTS2, NACE2class)
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Aggr
Time Periods	$\{-1, -i\}$
Long/Cross	Long + Cross
Cross-Domain Vars	`code_NACE2class_val_1`, code_NUTS2_val_1
Encoding	-

`rate_meanTrnovr_ed0_vali_NACE2class`	
Definition	Relative variation rate between the mean value of edited total turnover values from the reference time period and the mean value of validated total turnover values from the ith preceding reference time period of the industrial establishments across domains defined by variable `code_NACE2class_val_1` (NACE Rev. 2 class)
Stat Type	Numerical
Values	$\mathbb{R}$, $i = 1, 3, 6, 12$ (4 variables)
Example	0.11
Source	Internal-Derived
Formula	$\frac{\frac{1}{r_d(t)}\sum_{k\in d} z_k^{my,\text{ed}} - \frac{1}{N_d}\sum_{k\in d} z_k^{(m-i)y,\text{val}}}{\frac{1}{N_d}\sum_{k\in d} z_k^{(m-i)y,\text{val}}}$ where $d \in$ NACE2class
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Aggr
Time Periods	$\{0, -i\}$
Long/Cross	Long + Cross
Cross-Domain Vars	`code_NACE2class_val_1`
Encoding	-

`rate_meanTrnovr_val1_vali_NACE2class`	
Definition	Relative variation rate between the mean value of validated total turnover values from the preceding reference time period and the mean value of validated total turnover values from the ith preceding reference time period of the industrial establishments across domains defined by variable `code_NACE2class_val_1` (NACE Rev. 2 class)
Stat Type	Numerical
Values	$\mathbb{R}$, $i = 2, 4, 7, 13$ (4 variables)
Example	0.11
Source	Internal-Derived
Formula	$\frac{\frac{1}{N_d}\sum_{k\in d} z_k^{(m-1)y,\text{val}} - \frac{1}{N_d}\sum_{k\in d} z_k^{(m-i)y,\text{val}}}{\frac{1}{N_d}\sum_{k\in d} z_k^{(m-i)y,\text{val}}}$ where $d \in$ NACE2class
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Aggr
Time Periods	$\{-1, -i\}$
Long/Cross	Long + Cross
Cross-Domain Vars	`code_NACE2class_val_1`
Encoding	-

External Survey Variables

IPI_ed_0_NUTS2NACE2class	
Definition	Industrial Production Index computed with edited values from the reference time period across domains defined by variables `code_NACE2class_val_1` (NACE Rev. 2 class) and `code_NUTS2_val_1` (NUTS2 territorial unit)
Stat Type	Numerical
Values	$\mathbb{R}^+$
Example	104.45
Source	External-Primary
Formula	-
Stat Progr Ref	Spanish IOE-30050
Unit/Aggr	Aggr
Time Periods	{0}
Long/Cross	Cross
Cross-Domain Vars	`code_NACE2class_val_1`, `code_NUTS2_val_1`
Encoding	-

IPI_ed_0_NACE2class	
Definition	Industrial Production Index computed with edited values from the reference time period across domains defined by variable `code_NACE2class_val_1` (NACE Rev. 2 class)
Stat Type	Numerical
Values	$\mathbb{R}^+$
Example	104.45
Source	External-Primary
Formula	-
Stat Progr Ref	Spanish IOE-30050
Unit/Aggr	Aggr
Time Periods	{0}
Long/Cross	Cross
Cross-Domain Vars	`code_NACE2class_val_1`
Encoding	-

IPI_ed_0_NUTS2NACE2group	
Definition	Industrial Production Index computed with edited values from the reference time period across domains defined by variables `code_NACE2group_val_1` (NACE Rev. 2 group) and `code_NUTS2_val_1` (NUTS2 territorial unit)
Stat Type	Numerical
Values	$\mathbb{R}^+$
Example	104.45
Source	External-Primary
Formula	-
Stat Progr Ref	Spanish IOE-30050
Unit/Aggr	Aggr
Time Periods	{0}
Long/Cross	Cross
Cross-Domain Vars	`code_NACE2group_val_1`, `code_NUTS2_val_1`
Encoding	-

IPI_ed_0_NACE2group	
Definition	Industrial Production Index computed with edited values from the reference time period across domains defined by variable `code_NACE2group_val_1` (NACE Rev. 2 group)
Stat Type	Numerical
Values	$\mathbb{R}^+$
Example	104.45
Source	External-Primary
Formula	-
Stat Progr Ref	Spanish IOE-30050
Unit/Aggr	Aggr

Time Periods	{0}
Long/Cross	Cross
Cross-Domain Vars	`code_NACE2group_val_1`
Encoding	-

`rate_IPI_ed0_val1_NUTS2NACE2class`	
Definition	Relative variation rate between the Industrial Production Index computed with edited values from the reference time period and the Industrial Production Index computed with validated values from the preceding time period across domains defined by variables `code_NACE2class_val_1` (NACE Rev. 2 class) and `code_NUTS2_val_1` (NUTS2 territorial unit)
Stat Type	Numerical
Values	$\mathbb{R}$
Example	0.85
Source	External-Derived
Formula	$\frac{\text{IPI}_d^{my,\text{ed}} - \text{IPI}_d^{(m-1)y,\text{val}}}{\text{IPI}_d^{(m-1)y,\text{val}}}$ where $d \in (\text{NUTS2}, \text{NACE2class})$
Stat Progr Ref	Spanish IOE-30050
Unit/Aggr	Aggr
Time Periods	$\{0, -1\}$
Long/Cross	Long + Cross
Cross-Domain Vars	`code_NACE2class_val_1`, `code_NUTS2_val_1`
Encoding	-

`rate_IPI_ed0_val1_NACE2class`	
Definition	Relative variation rate between the Industrial Production Index computed with edited values from the reference time period and the Industrial Production Index computed with validated values from the preceding time period across domains defined by variable `code_NACE2class_val_1` (NACE Rev. 2 class)
Stat Type	Numerical
Values	$\mathbb{R}$
Example	0.85
Source	External-Derived
Formula	$\frac{\text{IPI}_d^{my,\text{ed}} - \text{IPI}_d^{(m-1)y,\text{val}}}{\text{IPI}_d^{(m-1)y,\text{val}}}$ where $d \in \text{NACE2class}$
Stat Progr Ref	Spanish IOE-30050
Unit/Aggr	Aggr
Time Periods	$\{0, -1\}$
Long/Cross	Long + Cross
Cross-Domain Vars	`code_NACE2class_val_1`
Encoding	-

`rate_IPI_ed0_val12_NUTS2NACE2class`	
Definition	Relative variation rate between the Industrial Production Index computed with edited values from the reference time period and the Industrial Production Index computed with validated values from the yearly preceding time period across domains defined by variables `code_NACE2class_val_1` (NACE Rev. 2 class) and `code_NUTS2_val_1` (NUTS2 territorial unit)
Stat Type	Numerical
Values	$\mathbb{R}$
Example	0.85
Source	External-Derived
Formula	$\frac{\text{IPI}_d^{my,\text{ed}} - \text{IPI}_d^{m(y-1),\text{val}}}{\text{IPI}_d^{m(y-1),\text{val}}}$ where $d \in (\text{NUTS2}, \text{NACE2class})$
Stat Progr Ref	Spanish IOE-30050
Unit/Aggr	Aggr
Time Periods	$\{0, -12\}$
Long/Cross	Long + Cross
Cross-Domain Vars	`code_NACE2class_val_1`, `code_NUTS2_val_1`
Encoding	-

`rate_IPI_ed0_val12_NACE2class`

Definition	Relative variation rate between the Industrial Production Index computed with edited values from the reference time period and the Industrial Production Index computed with validated values from the yearly preceding time period across domains defined by variable `code_NACE2class_val_1` (NACE Rev. 2 class)
Stat Type	Numerical
Values	$\mathbb{R}$
Example	0.85
Source	External-Derived
Formula	$\frac{\text{IPI}_d^{my,\text{ed}} - \text{IPI}_d^{m(y-1),\text{val}}}{\text{IPI}_d^{m(y-1),\text{val}}}$ where $d \in$ NACE2class
Stat Progr Ref	Spanish IOE-30050
Unit/Aggr	Aggr
Time Periods	$\{0, -12\}$
Long/Cross	Long + Cross
Cross-Domain Vars	`code_NACE2class_val_1`
Encoding	-

`rate_yearToDateMeanIPI_ed0_val12_NUTS2NACE2class`

Definition	Relative variation rate between the year-to-date mean of the Industrial Production Index computed with edited values from the reference time period year and the year-to-date mean of the Industrial Production Index computed with validated values from the preceding time period year across domains defined by variables `code_NACE2class_val_1` (NACE Rev. 2 class) and `code_NUTS2_val_1` (NUTS2 territorial unit)
Stat Type	Numerical
Values	$\mathbb{R}$
Example	0.97
Source	External-Derived
Formula	$\frac{y2dMean\left(\text{IPI}_d^{my,\text{ed}}\right) - y2dMean\left(\text{IPI}_d^{m(y-1),\text{val}}\right)}{y2dMean\left(\text{IPI}_d^{m(y-1),\text{val}}\right)}$ where $y2dMean\left(\text{IPI}_d^{my}\right) = \frac{1}{m}\sum_{j=1}^{m} \text{IPI}_d^{jy}$ and $d \in (\text{NUTS2}, \text{NACE2class})$
Stat Progr Ref	Spanish IOE-30050
Unit/Aggr	Aggr
Time Periods	$\{0, \ldots, -i(\text{Jan})\} \cup \{-12, \ldots, -i - 12(\text{Jan})\}$
Long/Cross	Long + Cross
Cross-Domain Vars	`code_NACE2class_val_1`, `code_NUTS2_val_1`
Encoding	-

`rate_yearToDateMeanIPI_ed0_val12_NACE2class`

Definition	Relative variation rate between the year-to-date mean of the Industrial Production Index computed with edited values from the reference time period year and the year-to-date mean of the Industrial Production Index computed with validated values from the preceding time period year across domains defined by variable `code_NACE2class_val_1` (NACE Rev. 2 class)
Stat Type	Numerical
Values	$\mathbb{R}$
Example	0.97
Definition	
Source	External-Derived
Formula	$\frac{y2dMean\left(\text{IPI}_d^{my,\text{ed}}\right) - y2dMean\left(\text{IPI}_d^{m(y-1),\text{val}}\right)}{y2dMean\left(\text{IPI}_d^{m(y-1),\text{val}}\right)}$ where $y2dMean\left(\text{IPI}_d^{my}\right) = \frac{1}{m}\sum_{j=1}^{m} \text{IPI}_d^{jy}$ and $d \in$ NACE2class
Stat Progr Ref	Spanish IOE-30050
Unit/Aggr	Aggr
Time Periods	$\{0, \ldots, -i(\text{Jan})\} \cup \{-12, \ldots, -i - 12(\text{Jan})\}$
Long/Cross	Long + Cross
Cross-Domain Vars	`code_NACE2class_val_1`

Encoding	-

`rate_MA12IPI_ed0_val12_NUTS2NACE2class`	
Definition	Relative variation rate between the 12-month moving average of the Industrial Production Index computed with edited values from the reference time period backwards and the 12-month moving average of the Industrial Production Index computed with validated values from the preceding time period year backwards across domains defined by variables `code_NACE2class_val_1` (NACE Rev. 2 class) and `code_NUTS2_val_1` (NUTS2 territorial unit)
Stat Type	Numerical
Values	$\mathbb{R}$
Example	0.97
Source	External-Derived
Formula	$\frac{\text{MA12}\left(IPI_d^{my,\text{ed}}\right)-\text{MA12}\left(\text{IPI}_d^{m(y-1),\text{val}}\right)}{\text{MA12}\left(\text{IPI}_d^{m(y-1),\text{val}}\right)}$ where $\text{MA12}\left(\text{IPI}_d^{my}\right)=\frac{1}{12}\sum_{j=0}^{11}\text{IPI}_d^{(m-j)y}$ and $d \in (\text{NUTS2}, \text{NACE2class})$
Stat Progr Ref	Spanish IOE-30050
Unit/Aggr	Aggr
Time Periods	$\{0,\ldots,-11\}\cup\{-12,\ldots,-23\}$
Long/Cross	Long + Cross
Cross-Domain Vars	`code_NACE2class_val_1`, `code_NUTS2_val_1`
Encoding	-

`rate_MA12IPI_ed0_val12_NACE2class`	
Definition	Relative variation rate between the 12-month moving average of the Industrial Production Index computed with edited values from the reference time period backwards and the 12-month moving average of the Industrial Production Index computed with validated values from the preceding time period year backwards across domains defined by variable `code_NACE2class_val_1` (NACE Rev. 2 class)
Stat Type	Numerical
Values	$\mathbb{R}$
Example	0.97
Source	External-Derived
Formula	$\frac{\text{MA12}\left(IPI_d^{my,\text{ed}}\right)-\text{MA12}\left(\text{IPI}_d^{m(y-1),\text{val}}\right)}{\text{MA12}\left(\text{IPI}_d^{m(y-1),\text{val}}\right)}$ where $\text{MA12}\left(\text{IPI}_d^{my}\right)=\frac{1}{12}\sum_{j=0}^{11}\text{IPI}_d^{(m-j)y}$ and $d \in \text{NACE2class}$
Stat Progr Ref	Spanish IOE-30050
Unit/Aggr	Aggr
Time Periods	$\{0,\ldots,-11\}\cup\{-12,\ldots,-23\}$
Long/Cross	Long + Cross
Cross-Domain Vars	`code_NACE2class_val_1`
Encoding	-

`rate_IPI_ed0_val1_NUTS2NACE2group`	
Definition	Relative variation rate between the Industrial Production Index computed with edited values from the reference time period and the Industrial Production Index computed with validated values from the preceding time period across domains defined by variables `code_NACE2group_ed` (NACE Rev. 2 group) and `code_NUTS2_val_1` (NUTS2 territorial unit)
Stat Type	Numerical
Values	$\mathbb{R}$
Example	0.85
Source	External-Derived
Formula	$\frac{\text{IPI}_d^{my,\text{ed}}-\text{IPI}_d^{(m-1)y,\text{val}}}{\text{IPI}_d^{(m-1)y,\text{val}}}$ where $d \in (\text{NUTS2}, \text{NACE2group})$
Stat Progr Ref	Spanish IOE-30050
Unit/Aggr	Aggr
Time Periods	$\{0,-1\}$

Long/Cross	Long + Cross
Cross-Domain Vars	`code_NACE2group_ed`, `code_NUTS2_val_1`
Encoding	-

`rate_IPI_ed0_val1_NACE2group`	
Definition	Relative variation rate between the Industrial Production Index computed with edited values from the reference time period and the Industrial Production Index computed with validated values from the preceding time period across domains defined by variable `code_NACE2group_ed` (NACE Rev. 2 group)
Stat Type	Numerical
Values	$\mathbb{R}$
Example	0.85
Source	External-Derived
Formula	$\frac{\text{IPI}_d^{my,\text{ed}} - \text{IPI}_d^{(m-1)y,\text{val}}}{\text{IPI}_d^{(m-1)y,\text{val}}}$ where $d \in$ NACE2group
Stat Progr Ref	Spanish IOE-30050
Unit/Aggr	Aggr
Time Periods	$\{0, -1\}$
Long/Cross	Long + Cross
Cross-Domain Vars	`code_NACE2group_ed`
Encoding	-

`rate_IPI_ed0_val12_NUTS2NACE2group`	
Definition	Relative variation rate between the Industrial Production Index computed with edited values from the reference time period and the Industrial Production Index computed with validated values from the preceding time period year across domains defined by variable `code_NACE2group_ed` (NACE Rev. 2 group) and `code_NUTS2_val_1` (NUTS2 territorial unit)
Stat Type	Numerical
Values	$\mathbb{R}$
Example	0.85
Source	External-Derived
Formula	$\frac{\text{IPI}_d^{my,\text{ed}} - \text{IPI}_d^{m(y-1),\text{val}}}{\text{IPI}_d^{m(y-1),\text{val}}}$ where $d \in$ (NUTS2, NACE2group)
Stat Progr Ref	Spanish IOE-30050
Unit/Aggr	Aggr
Time Periods	$\{0, -12\}$
Long/Cross	Long + Cross
Cross-Domain Vars	`code_NACE2group_ed`, `code_NUTS2_val_1`
Encoding	-

`rate_IPI_ed0_val12_NACE2group`	
Definition	Relative variation rate between the Industrial Production Index computed with edited values from the reference time period and the Industrial Production Index computed with validated values from the preceding time period year across domains defined by variable `code_NACE2group_ed` (NACE Rev. 2 group)
Stat Type	Numerical
Values	$\mathbb{R}$
Example	0.85
Definition	
Source	External-Derived
Formula	$\frac{\text{IPI}_d^{my,\text{ed}} - \text{IPI}_d^{m(y-1),\text{val}}}{\text{IPI}_d^{m(y-1),\text{val}}}$ where $d \in$ NACE2group
Stat Progr Ref	Spanish IOE-30050
Unit/Aggr	Aggr
Time Periods	$\{0, -12\}$
Long/Cross	Long + Cross
Cross-Domain Vars	`code_NACE2group_ed`
Encoding	-

`rate_yearToDateMeanIPI_ed0_val12_NUTS2NACE2group`

Definition	Relative variation rate between the year-to-date mean of the Industrial Production Index computed with edited values from the reference time period year and the year-to-date mean of the Industrial Production Index computed with validated values from the preceding time period year across domains defined by variables `code_NACE2group_ed` (NACE Rev. 2 class) and `code_NUTS2_val_1` (NUTS2 territorial unit)
Stat Type	Numerical
Values	$\mathbb{R}$
Example	0.97
Source	External-Derived
Formula	$\frac{y2dMean\left(\mathrm{IPI}_d^{my,\mathrm{ed}}\right)-y2dMean\left(\mathrm{IPI}_d^{m(y-1),\mathrm{val}}\right)}{y2dMean\left(\mathrm{IPI}_d^{m(y-1),\mathrm{val}}\right)}$ where $y2dMean\left(\mathrm{IPI}_d^{my}\right)=\frac{1}{m}\sum_{j=1}^{m}\mathrm{IPI}_d^{jy}$ and $d\in(\mathrm{NUTS2},\mathrm{NACE2group})$
Stat Progr Ref	Spanish IOE-30050
Unit/Aggr	Aggr
Time Periods	$\{0,\ldots,-i(\mathrm{Jan})\}\cup\{-12,\ldots,-i-12(\mathrm{Jan})\}$
Long/Cross	Long + Cross
Cross-Domain Vars	`code_NACE2group_ed`, `code_NUTS2_val_1`
Encoding	-

`rate_yearToDateMeanIPI_ed0_val12_NACE2group`	
Definition	Relative variation rate between the year-to-date mean of the Industrial Production Index computed with edited values from the reference time period year and the year-to-date mean of the Industrial Production Index computed with validated values from the preceding time period year across domains defined by variable `code_NACE2group_ed` (NACE Rev. 2 class)
Stat Type	Numerical
Values	$\mathbb{R}$
Example	0.97
Source	External-Derived
Formula	$\frac{y2dMean\left(\mathrm{IPI}_d^{my,\mathrm{ed}}\right)-y2dMean\left(\mathrm{IPI}_d^{m(y-1),\mathrm{val}}\right)}{y2dMean\left(\mathrm{IPI}_d^{m(y-1),\mathrm{val}}\right)}$ where $y2dMean\left(\mathrm{IPI}_d^{my}\right)=\frac{1}{m}\sum_{j=1}^{m}\mathrm{IPI}_d^{jy}$ and $d\in\mathrm{NACE2group}$
Stat Progr Ref	Spanish IOE-30050
Unit/Aggr	Aggr
Time Periods	$\{0,\ldots,-i(\mathrm{Jan})\}\cup\{-12,\ldots,-i-12(\mathrm{Jan})\}$
Long/Cross	Long + Cross
Cross-Domain Vars	`code_NACE2group_ed`
Encoding	-

`rate_MA12IPI_ed0_val12_NUTS2NACE2group`	
Definition	Relative variation rate between the moving average of the Industrial Production Index computed with edited values from the reference time period year and the year-to-date mean of the Industrial Production Index computed with validated values from the preceding time period year across domains defined by variables `code_NACE2group_ed` (NACE Rev. 2 class) and `code_NUTS2_val_1` (NUTS2 territorial units)
Stat Type	Numerical
Values	$\mathbb{R}$
Example	0.97
Source	External-Derived
Formula	$\frac{\mathrm{MA12}\left(IPI_d^{my,\mathrm{ed}}\right)-\mathrm{MA12}\left(\mathrm{IPI}_d^{m(y-1),\mathrm{val}}\right)}{\mathrm{MA12}\left(\mathrm{IPI}_d^{m(y-1),\mathrm{val}}\right)}$ where $\mathrm{MA12}\left(\mathrm{IPI}_d^{my}\right)=\frac{1}{12}\sum_{j=0}^{11}\mathrm{IPI}_d^{(m-j)y}$ and $d\in(\mathrm{NUTS2},\mathrm{NACE2group})$
Stat Progr Ref	Spanish IOE-30050
Unit/Aggr	Aggr
Time Periods	$\{0,\ldots,-11\}\cup\{-12,\ldots,-23\}$
Long/Cross	Long + Cross

Cross-Domain Vars	code_NACE2group_ed, code_NUTS2_val_1
Encoding	-

rate_MA12IPI_ed0_val12_NACE2group	
Definition	Relative variation rate between the moving average of the Industrial Production Index computed with edited values from the reference time period year and the year-to-date mean of the Industrial Production Index computed with validated values from the preceding time period year across domains defined by variable code_NACE2group_ed (NACE Rev. 2 class)
Stat Type	Numerical
Values	$\mathbb{R}$
Example	0.97
Source	External-Derived
Formula	$\frac{\mathrm{MA12}\left(IPI_d^{my,\mathrm{ed}}\right)-\mathrm{MA12}\left(\mathrm{IPI}_d^{m(y-1),\mathrm{val}}\right)}{\mathrm{MA12}\left(\mathrm{IPI}_d^{m(y-1),\mathrm{val}}\right)}$ where $\mathrm{MA12}\left(\mathrm{IPI}_d^{my}\right)=\frac{1}{12}\sum_{j=0}^{11}\mathrm{IPI}_d^{(m-j)y}$ and $d \in$ NACE2group
Stat Progr Ref	Spanish IOE-30050
Unit/Aggr	Aggr
Time Periods	$\{0,\ldots,-11\}\cup\{-12,\ldots,-23\}$
Long/Cross	Long + Cross
Cross-Domain Vars	code_NACE2group_ed
Encoding	-

IPRI_val_0_NUTS2NACE2class	
Definition	Industrial Price Index computed from validated values of the reference time period across domains defined by variables code_NACE2class_val_1 (NACE Rev. 2 class) and code_NUTS2_val_1 (NUTS2 territorial unit)
Stat Type	Numerical
Values	$\mathbb{R}^+$
Example	101.03
Source	External-Primary
Formula	-
Stat Progr Ref	Spanish IOE-30051
Unit/Aggr	Aggr
Time Periods	$\{0\}$
Long/Cross	Cross
Cross-Domain Vars	code_NACE2class_val_1, code_NUTS2_val_1
Encoding	-

IPRI_val_0_NACE2class	
Definition	Industrial Price Index computed from validated values of the reference time period across domains defined by variable code_NACE2class_val_1 (NACE Rev. 2 class)
Stat Type	Numerical
Values	$\mathbb{R}^+$
Example	101.03
Source	External-Primary
Formula	-
Stat Progr Ref	Spanish IOE-30051
Unit/Aggr	Aggr
Time Periods	$\{0\}$
Long/Cross	Cross
Cross-Domain Vars	code_NACE2class_val_1
Encoding	-

IPRI_val_0_NUTS2NACE2group	
Definition	Industrial Price Index computed from validated values of the reference time period across domains defined by variables code_NACE2group_val_1 (NACE Rev. 2 class) and code_NUTS2_val_1 (NUTS2 territorial unit)
Stat Type	Numerical
Values	$\mathbb{R}^+$

Example	101.03
Source	External-Primary
Formula	-
Stat Progr Ref	Spanish IOE-30051
Unit/Aggr	Aggr
Time Periods	{0}
Long/Cross	Cross
Cross-Domain Vars	`code_NACE2group_val_1`, `code_NUTS2_val_1`
Encoding	-

`IPRI_val_0_NACE2group`	
Definition	Industrial Price Index computed from validated values of the reference time period across domains defined by variable `code_NACE2group_val_1` (NACE Rev. 2 class)
Stat Type	Numerical
Values	$\mathbb{R}^+$
Example	101.03
Definition	
Source	External-Primary
Formula	-
Stat Progr Ref	Spanish IOE-30051
Unit/Aggr	Aggr
Time Periods	{0}
Long/Cross	Cross
Cross-Domain Vars	`code_NACE2group_val_1`
Encoding	-

References

Bentéjac, C., A. Csörgő, and G. M.-M. noz (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review 54*, 1937–1967.

Bohnensteffen, S. (2020). Selective data editing of continuous variables with random forests in official statistics. Awarded Best Master Thesis in the NTTS2020 Master Thesis EMOS Competition. https://ec.europa.eu/eurostat/cros/content/emos_en.

Bok, B., D. Caratelli, D. Giannone, A. Sbordone, and A. Tambalotti (2017). Macroeconomic nowcasting and forecasting with Big Data. Technical report, Staff Report, No. 830, Federal Reserve Bank of New York. https://www.econstor.eu/handle/10419/189871.

Broe, S. D., O. ten Bosch, P. Daas, G. Buiten, B. Laevens, and B. Kroese (2021). The need for timely official statistics. The COVID-19 pandemic as a driver for innovation. *Statistical Journal of the IAOS 37*, 1221â€“1227.

Chambers, R. (1986). Outlier Robust Finite Population Estimation. *Journal of the American Statistical Association 81*(396), 1063–1069.

Chang, W., J. Cheng, J. Allaire, C. Sievert, B. Schloerke, Y. Xie, J. Allen, J. McPherson, A. Dipert, and B. Borges (2021). *shiny: Web Application Framework for R*. R package version 1.7.1.

D. Lazer, D., R. Kennedy, G. King, and A. Vespignani (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science 343*, 1203–1205.

Dagdoug, M., C., Goga, and D. Haziza (2021, 09). Imputation Procedures in Surveys Using Nonparametric and Machine Learning Methods: an Empirical Comparison. *Journal of Survey Statistics and Methodology*.

ESS (2014). ESS Handbook for Quality Reports. https://ec.europa.eu/eurostat/documents/3859598/6651706/KS-GQ-15-003-EN-N.pdf/18dd4bf0-8de6-4f3f-9adb-fab92db1a568.

Esteban, E., M. Novás, S. S. na, D. Salgado, and L. Sanguiao (2018). Data organisation and process design based on functional modularity for a standard production process.

Eurostat (2010). European business statistics regulation. commission implementing regulation 2020/1197. Technical report.

Eurostat (2017). Handbook of Rapid Estimates. https://ec.europa.eu/eurostat/documents/3859598/8555708/KS-GQ-17-008-EN-N.pdf/7f40c70d-0a44-4459-b5b3-72894e13ca6d?t=1513758176000.

Eurostat (2019). Regulation (eu) 2019/2152 of the european parliament and of the council on european business statistics. Technical report.

Eurostat (2021). Ess reference architecture reference framework. https://ec.europa.eu/eurostat/cros/content/ess-enterprise-architecture-reference-framework_en.

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics 29*, 1189–1232.

Giannone, D., L. Reichlin, M. Bańbura, and M. Modugno (2013). Now-casting and the real-time data flow. In G. Elliott and A. Timmermann (Eds.), *Handbook of Economic Nowcasting*, Chapter 4, pp. 195–237. Amsterdam: Elsevier.

Ginsberg, L., M. Mohebbi, R. Patel, L. Brammer, M. Smolinski, and L. Brilliant (2009). Detecting influenza epidemics using search engine query data. *Nature 457*, 1012–1014.

Gorman, B. (2018). *mltools: Machine Learning Tools*. R package version 0.3.5.

Hand, D. (2018). Statistical challenges of administrative and transaction data. *Journal of the Royal Statistical Society A 181*, 555–605.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning* (Second ed.). New York: Springer.

Hyndman, R. and Y. Fan (1996). Sample Quantiles in Statistical Packages. *The American Statistician 50*, 361–365.

INE (2018). *Industrial Turnover Indices & Industrial New Orders Received Indices. Base 2015*. https://www.ine.es/en/metodologia/t05/t0530053_2015_en.pdf.

Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems 30*, 3149–3157. NIPS 2017.

Kitchin, R. (2015). Big Data and Official Statistics: Opportunities, challenges and risks. *Statistical Journal of the IAOS 31*, 471–481.

Little, R. and D. Rubin (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken: Wiley.

Maronna, R., R. Martin, V. Yohai, and M. Salibián-Barrera (2009). *Robust Statistics: Theory and Methods (with R)*. Wiley.

Mazzi, G. and R. Cannata (2017). Rapid Estimates: Different Products for Different Purposes. In Eurostat (Ed.), *Handbook of Rapid Estimates*, Chapter 2, pp. 28–51. Luxembourg: Eurostat.

Microsoft Corporation (2022). Lightgbm.

Mohri, M., A. Rostamizadeh, and A. Talwalkar (2018). *Foundations of Machine Learning* (2nd ed.). MIT Press.

Murphy, K. (2013). *Machine learning: a probabilistic perspective*. MIT Press.

Salgado, D. and B. Oancea (2020). On new data sources for the production of official statistics. arXiv:2003.06797v1.

Sanguiao, L. and L.-C. Zhang (2021). Design-Unbiased Statistical Learning in Survey Sampling. *Sankhya A: The Indian Journal of Statistics 83*, 714–744.

Shi, Y., G. Ke, D. Soukhavong, J. Lamb, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, and N. Titov (2021). *lightgbm: Light Gradient Boosting Machine*. R package version 3.3.1.

UNECE (2013). Generic statistical information model v1.2. https://statswiki.unece.org/display/gsim.

UNECE (2019a). Generic activity model for statistical organizations v1.2. https://statswiki.unece.org/display/GAMSO/GAMSO+v1.2.

UNECE (2019b). Generic Statistical Business Process Model v5.1. https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1.

UNECE (2019c). Generic statistical data editing model v2.0. https://statswiki.unece.org/display/sde/GSDEM.

UNECE (2021). High-level group for the modernisation of statistical production and services. https://unece.org/statistics/networks-of-experts/high-level-group-modernisation-statistical-production-and-services.

UNECE (2022). Linking GSBPM and GSIM v1.0. https://statswiki.unece.org/display/GSBPM/Information+flow+within+GSBPM+using+GSIM.

Wand, Y. and R. Wang (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM 39*, 86–95.

Watt, J., R. Borhani, and A. K. Katsaggelos (2020). *Machine Learning Refined: Foundations, Algorithms, and Applications* (2nd ed.). Cambridge University Press.