Timeliness and Accuracy with Machine Learning Algorithms: Early Estimates of the Industrial Turnover Index

David Salgado[1], Sandra Barragán[1], and Elena Rosa-Pérez[2]

[1]S.G. for Methodology and Sampling Design, Statistics Spain (INE), Spain
[2]S.G. for Short-Term Statistics, Statistics Spain (INE), Spain

April 30, 2023

1 Introduction

The quest for quality improvement in the production of official statistics does not only embrace the traditional accuracy dimension but also, and with remarkable intensity in recent times, other quality dimensions such as timeliness and opportunity, cost-efficiency, and response burden. In this context, the incorporation of new data sources together with new statistical methods has been considered as a must to avoid losing relevance in the production of statistics.

More often than not, this key strategic direction has been used to completely disregard survey data due to their cost (especially regarding collection and editing), their burden (response is needed from each statistical unit), and their slowness (the difference between the release date and the time reference period goes usually beyond several weeks). The argument can be represented as in figure 1. The horizontal axis represents the timeline, where a reference period has been marked. The measurement unit in this axis is basically the number of days, which is related to timeliness. The vertical axis represents the mean squared error (MSE) as a measurement of the degree of accuracy. As time goes by, before the reference period, all we can do is to provide a prediction, usually of poor accuracy (high MSE), not based on any data from that period. As the reference period starts, at most we can begin incorporating some kind of data from the present time containing a signal of the phenomenon under analysis, thus hopefully improving the accuracy (reducing the MSE) but still a prediction. When the reference period is over, data from this period has been generated and the production machinery starts to work. Although execution phases are conceived of sequentially (collect, then edit, then estimate, then disseminate), in

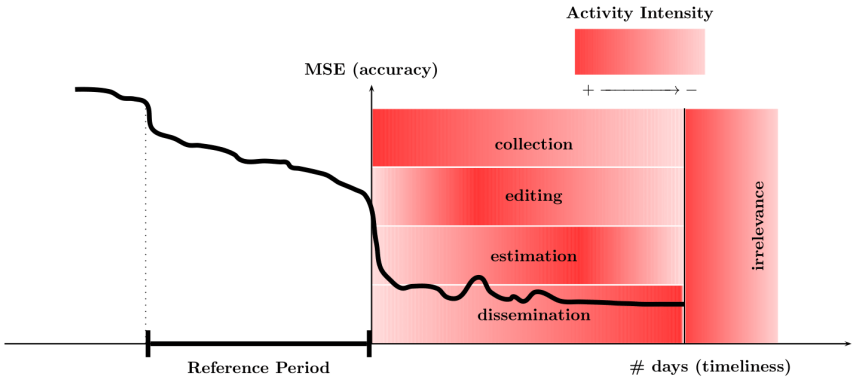

Figure 1: The timeliness-vs-accuracy argument

practice there is some overlapping: editing already starts when a fraction of the sample has been collected, preliminary estimates are produced even before the whole data collection and editing are concluded, and even some delayed data are collected after the dissemination of a first version of the statistics. After too many days from the reference period, we fall into irrelevance.

We have constructed a minimum viable product to show that timeliness can be noticeably increased even with survey data providing a controlled assesment of accuracy by resorting to statistical learning algorithms. The core idea is to train a prediction model for the target variable at the sampling unit level using microdata from the historic time series of the survey and microdata already collected and edited during the reference period on course. The model can be updated when new microdata are available (even daily) to have a complete microdata set at all times since the time reference period is over. Aggregates and indices can then be computed at all times being updated accordingly.

In this paper, we propose an early estimation of the Spanish Industrial Turnover Index (ITI), which is a monthly Short-Term Business Statistics produced by National Statistics Institutes (NSIs) with a cut-off sample of 12000 industrial establishments in the European Statistical System (ESS) under Regulations 2019/2152 and 2020/1197 (EP, 2019; EC, 2020). This early estimation of the ITI clearly improves the timeliness of this statistic using only the traditional survey data providing also an accuracy measure.

2 Preliminary remarks

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 (not stock data), it is conceptually impossible to **measure** any indicator relative to a reference period until this 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, 2019) 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 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 complete sample for each month m. 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.

3 The statistical learning model

Barragán et al. (2022) provide fully fledged methodological details about the inference paradigm, the editing and validation of target variable values, the total estimators and indices computation, the construction of regressors for the model, the treatment of missing values and outliers, the training, testing, and validation of the model, the hyperparameters and model selection, and the accuracy assessment in terms of bias, variance, and the mean squared error.

We single out the construction of regressors (feature engineering) because, in our opinion, it concentrates the key step in providing high-quality predicted values, namely the information representation step. For this short-term business statistics the monthly statistical variables are basically the turnover, the economic classification codes (NACE class) of the industrial establishment and its corporation (enterprise), and the municipality. From these variables, more variables can be elementarily derived such as NACE section, division and group codes, aggregated territorial regions, and cross-tabulations thereof. More interestingly we can also compute moving averages for all statistical units, quantiles across different domains, and moving averages of these quantiles. Basically, we construct two types of regressors: (i) at statistical unit level based on individual past data and (ii) at aggregated level for different domain sizes (geographical, NACE-code, cross-tabulated) using monthly cross-sectional data even from the reference time period on course. The construction of these regressors amounts to an information representation exercise which is guided by the subject matter expert knowledge applied during the standard production process to validate every microdata value.

The implementation of the minimum viable product has been carried out in a modular process following the principles of both GSBPM and GSIM in a single PC in R language (Barragán et al., 2021).

4 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 (at $m + 20$, $m + 27$, $m + 38$) as they are made available during data collection. These quantities are computed together with their respective conditional 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 1).

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

Table 1: Number of categories per index breakdown.

In figure 2 we represent an example comprising the three index versions (initial, batches, final) from January 2020 to April 2021 for the national index.

In figure 3 we represent the corresponding annual variation rates for these same time periods for the national index.

5 Reflections and conclusions

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

[1]Division-group is a specific code list of divisions and combinations of NACE Rev. 2 groups which are necessary to build the Main Industrial Groupings (MIGs).

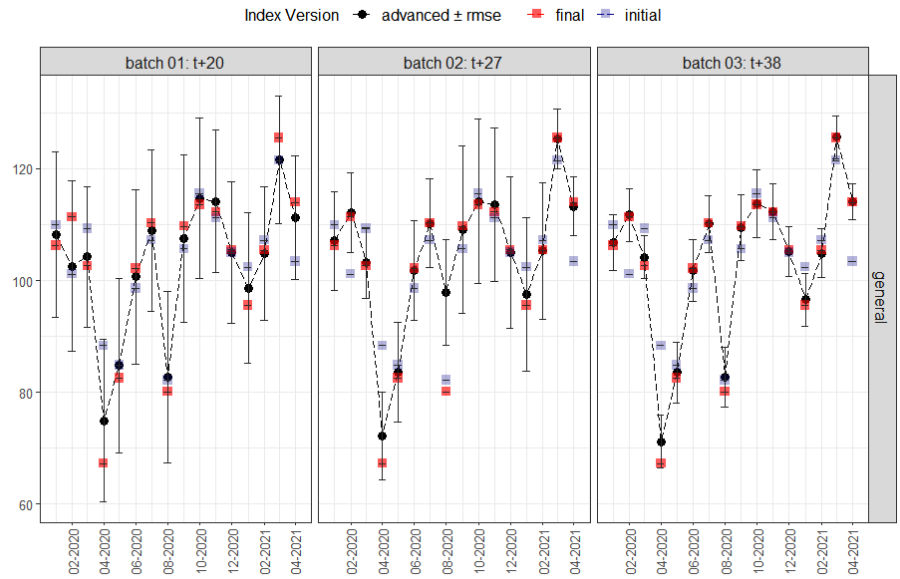

Figure 2: National index series from Jan2020 to Apr2021.

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 and edited 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.

This example proves that the goal to have a continuously updated synthetic microdata set is possible thus providing early estimates of aggregates and indices computed thereof. More important than model improvements mentioned above

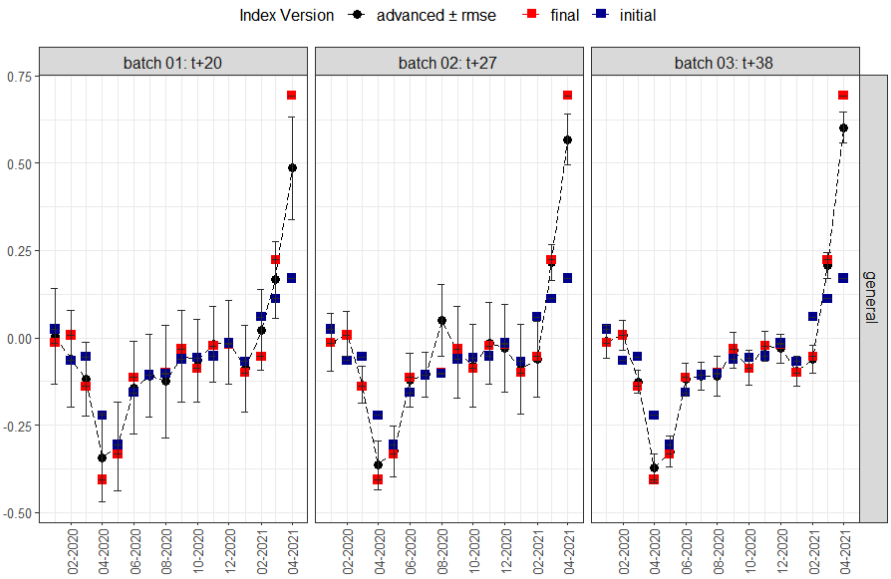

Figure 3: Annual variation rates series for the national index from Jan2020 to Apr2021.

we understand that strategic line of actions can be undertaken:

- If we can provide reliable predicted values for each statistical unit, we can reformulate the sample selection problem as the problem to select those units which allow us to maintain the model quality so that not every single unit must be required to provide a response every single time period. We can reduce response burden. The goal of the sample selection is not the final aggregate estimate but the quality of the prediction model.
- A continuously updated model can be readily used to deal with non-response.
- Furthermore, it can be used to early detect outliers, which are especially relevant in highly skewed distributions typical in business statistics.
- The model can be used to explore the possibility to provide values for those units in the frame population not originally selected in the originally trained model, thus possibly allowing us to provide higher levels of breakdowns.

We point out that, in our opinion, the information representation stage in the construction of the model is more critical than the selection of the statistical model (random forest, gradient boosting, etc.). The information representation

carried out here 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 official statistics is not only advisable but also necessary for a high-quality predictive model. The use of deep learning techniques to approach this stage is open to investigate its performance.

Moreover, 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 to avoid model drift and model deterioration. Further work and empirical evidence is needed in this line.

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 with a controlled assessement of accuracy.

References

Barragán, S., L. Barreñada, J. Calatrava, J. G. S. de Cueto, J. M. del Moral, E. Rosa-Pérez, and D. Salgado (2021). *AdvITI: Early Estimates of Spanish Industrial Turnover Index*.

Barragán, S., L. Barreñada, J. Calatrava, J. G. S. de Cueto, J. M. del Moral, E. Rosa-Pérez, and D. Salgado (2022). Early estimates of the industrial turnover index using statistical learning algorithms. Statistics Spain Working Paper 03/22.

EC (2020). Commission Implementing Regulation 2020/1197 laying down technical specifications and arrangements pursuant to Regulation (EU) 2019/2152 of the European Parliament and of the Council on European business statistics repealing 10 legal acts in the field of business statistic (General Implementing Act). Technical report. `https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32020R1197`.

EP (2019). Regulation (EU) 2019/2152 of the European Parliament and of the Council on European business statistics, repealing 10 legal acts in the field of business statistics (EBS-Regulation). `https://ec.europa.eu/eurostat/web/short-term-business-statistics/legislation`.

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.

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