United Nations

ECE/CES/2023/27

Economic and Social Council

Distr.: General
6 June 2023

English only

Economic Commission for Europe

Conference of European Statisticians

Seventy-first plenary session
Geneva, 22–23 June 2023
Item 9 of the provisional agenda
Timeliness, granularity and frequency of official statistics

Statistical learning in the Industrial Turnover Index: from a use case to strategic reflections

Prepared by Spain

Summary

Using statistical learning models the National Statistics Institute of Spain has reconstructed the microdata set for the Industrial Turnover Index updating the model as the data collection and data editing phases are underway, thus allowing to provide early estimates of the finally disseminated indices. This exercise covers 60 consecutive months of the survey. From this use case several reflections can be drawn regarding timeliness, accuracy (and their trade-off), response burden, granularity, cost-efficiency and organizational issues regarding the usage of these techniques at scale in a statistical office.

The document is presented to the Conference of European Statisticians’ session on “Timeliness, granularity and frequency of official statistics” for discussion.

I. Context

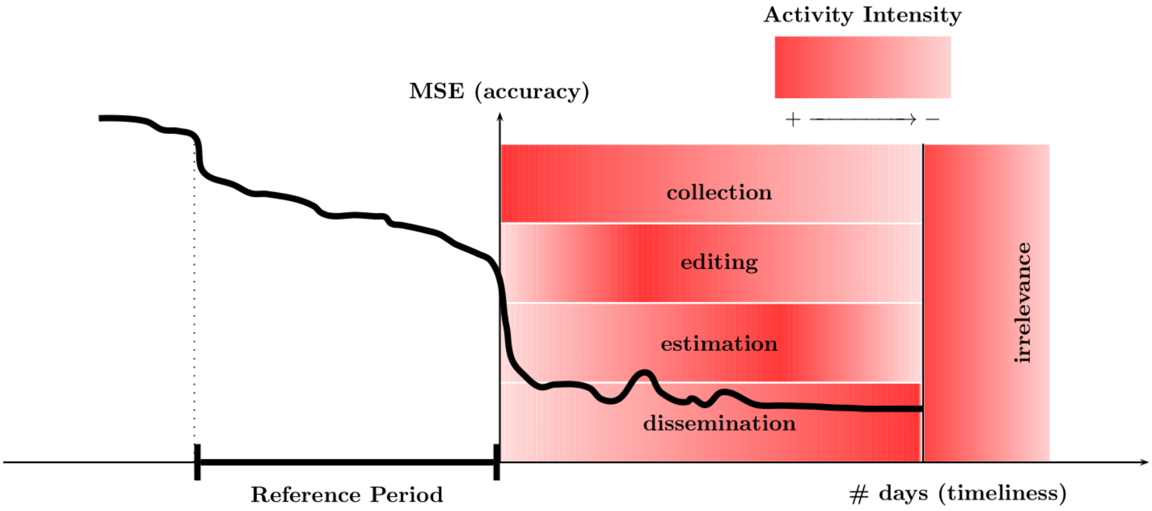
1. Accuracy has been the historic quality dimension in the production of official statistics but in the last decades more attention is demanded to timeliness and opportunity, cost-efficiency and response burden. Statistical learning techniques stand as a highly versatile tool to cover this demand by producing faster estimates under a controlled trade-off between timeliness and accuracy, but also as an opportunity to approach other quality dimensions by streamlining traditional business functions. These methods can readily be applied to survey data, as we show below, thus avoiding potential access and methodological issues from non-traditional data sources.

2. Usually, to face this demand for timeliness, the key strategic direction has been to focus on new data sources and new methods. Survey data are disregarded 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). We represent this argument 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 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.

3. We have conducted a pilot experience reconstructing the time series for the Spanish Industrial Turnover Index for 60 consecutive months for all publication cells. We have trained a prediction model for the target variable (total turnover) 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 is updated when new validated microdata are available (even daily) to have a complete microdata set at all times since the time reference period is over.

4. This experience, not implemented in production yet, allows us to draw some preliminary strategic reflections to approach the improvement of different quality dimensions and to streamline the production of multiple official statistics. We very shortly describe the pilot experience and provide our initial strategic reflections.

Figure 1
The timeliness-vs-accuracy argument



II. The pilot experience

5. We have conducted a simulated early estimation of the Spanish 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 Regulations 2019/2152 and 2020/1197 (EP, 2019; EC, 2020). In Spain, we use a monthly cut-off sample of 12000 industrial establishments.

6. 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.

7. 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 pro- duction process to validate every microdata value.

8. The pilot experience 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). The main results of this pilot study comprise the series of early estimates of the ITI broken down according to publication cells as well as their corresponding annual 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 (rmse). 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. 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 and their corresponding annual variation rates (figure 3).

Figure 2
National index series from Jan2020 to Apr2021

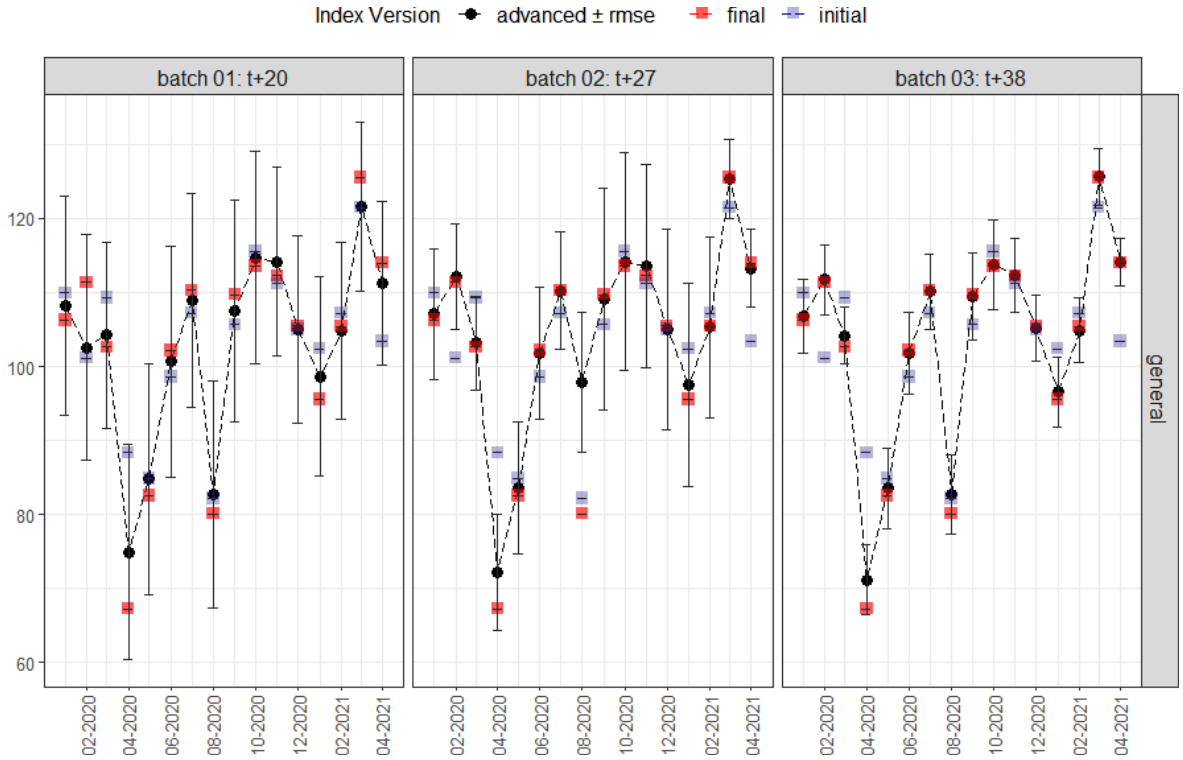
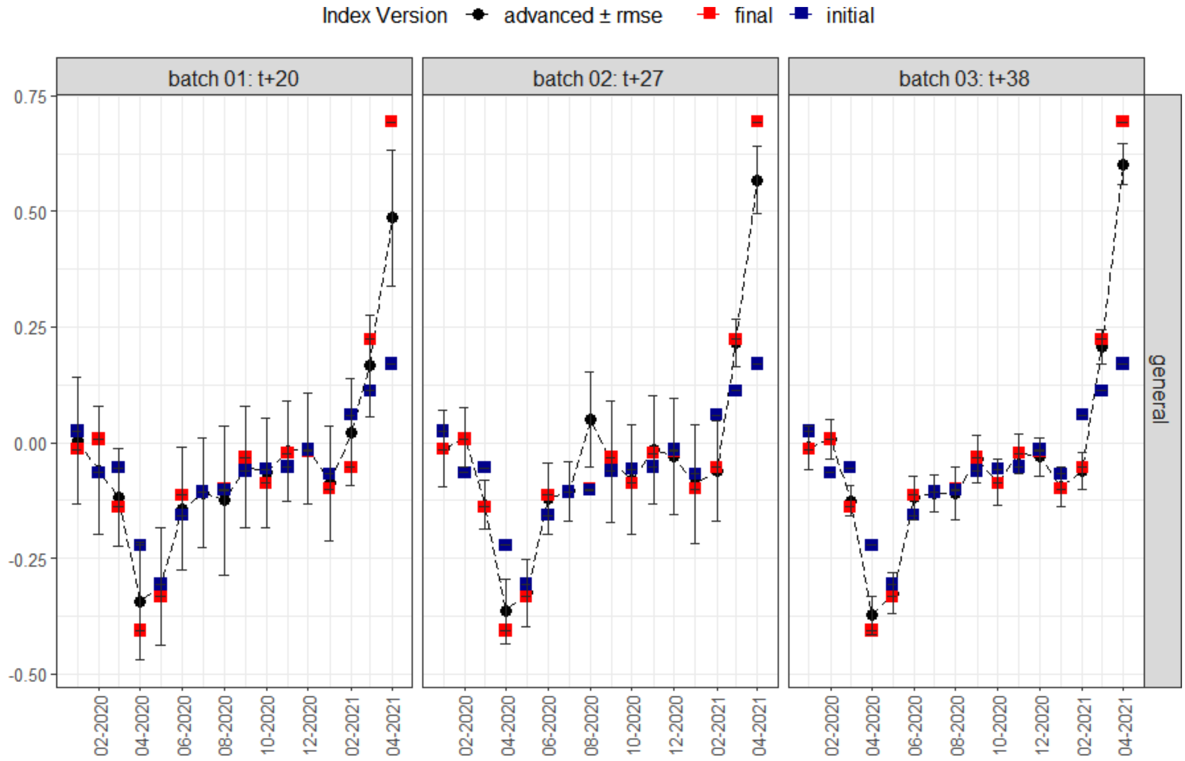


Figure 2
Annual variation rates series from Jan2020 to Apr2021



III. Strategic reflections

9. Although not having deployed the pilot experience to production, we can already draw some relevant reflections from the strategic point of view:

- This example strongly suggests that the goal to have a continuously updated synthetic microdata set is possible thus providing early estimates of aggregates and indices computed thereof. This could be considered for important short-term business statistics (especially for their usage in Quarterly National Accounts). Therefore, timeliness can be approached bottom-up working upon microdata by improving existing business functions with the predictive power of statistical learning models.

- The traditional production process needs some adjustments, for example, by making collected and vali- dated microdata as early available to model updating as possible, moving from a batch-based perspective to a continuous updating of the systems.

- In the use of statistical learning techniques, we find it more relevant the information representation step than the application and selection of a statistical model. This information representation step should be guided by subject matter knowledge, thus suggesting a strong collaboration between domain experts, methodologists, and computer scientists. The use of deep learning techniques to approach this stage is open to investigate its performance.

- In this line, the microdata validation process is crucial for a high-quality algorithm training. Again, editing and validation business functions need to be accommodated to incorporate subject matter knowledge as automatically and early as possible. This will reduce the chance of model drift and model deterioration.

- 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. The goal of the sample selection is not the final aggregate estimate but the quality of the prediction model. Therefore, we could reduce response burden.

- A continuously updated model can be readily used to deal with non-response. Needless to say, caution should still be necessary in presence of non-ignorability of the response mechanism and similar challenging conditions lurking the non-response.

- Furthermore, the model and predictions thereof can be used to early detect outliers, which are especially relevant in highly skewed distributions typical in business statistics. Subject matter knowledge (e.g. to distinguish between representative and non-representative outliers) is crucial. Thus, we can gain in cost- efficiency. Editing and imputation strategies (as e.g. in the Generic Statistical Data Editing Model) should be updated accordingly in production.

- The model could 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. The availability of auxiliary information (e.g. from administrative data) to be used as regressors becomes key. As a result, granularity could be increased.

- Survey data do not play a central role in this exercise and a similar approach could be considered for administrative data. However, access to these data needs to be restated. Usually, NSIs have access to these data after the reference period is over and after some pre-processing and revision tasks are concluded by their holders. Data access and use is approached in batches. Timeliness could be improved should NSIs have earlier and continuous access to admin data as soon as they are generated and/or collected.

10. These quality-oriented reflections bring potential consequences for architectural and organizational aspects of the production at scale:

- Data architecture is crucial to generalize this kind of approaches at scale. The amount of pre-processing tasks in our exercise to prepare data and compute regressors (feature engineering) would clearly benefit if a data architecture with fully-fledged metadata is put in place and shared among all surveys.

- In this same line, a repository of regressors or features would be highly advised with a continuously updating process in place incorporating subject matter knowledge.

- Skills related to statistical learning need to be generalized and not concentrated on specific units much in the same way as sample survey methodology is general knowledge among production staff in a statistical office.

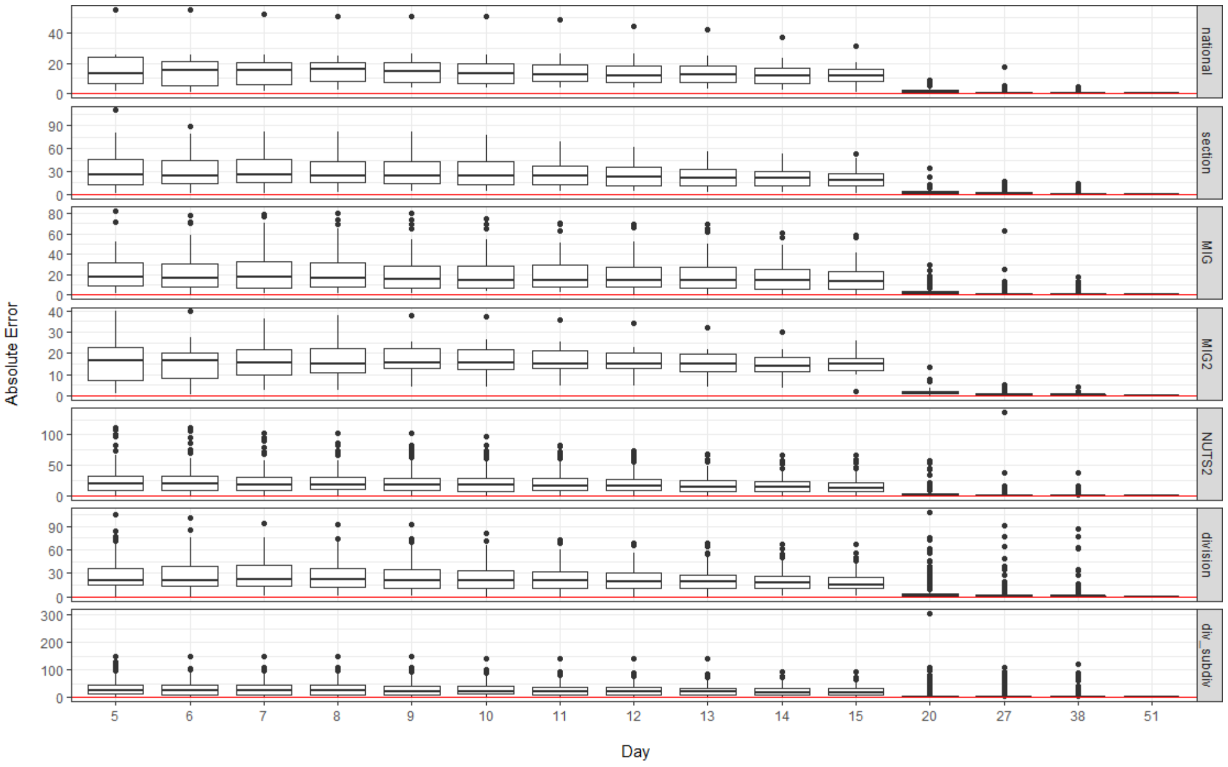
- Computational capacity is needed to continuously update the models with new collected and validated data. Technological platforms providing these new functionalities are necessary.

- A standard protocol to promote proofs of concept, minimal viable products, and experimental statistics to official statistics need to be also put in place hopefully following common international guidelines. For example, when should these new features be included (then made compulsory) in national and international legal regulations (e.g. can it be made legally compulsory to deliver a short-term business statistics in, say, 15 days?).

11. As a final observation, we tried to reproduce the exercise using daily collected data (see figure 4) to quantitatively assess the trade-off between accuracy and timeliness, which is indeed intimately related also to granularity (for a given sample size the more granular, the less accurate). This was too computationally demanding but, however, the organization of work in batches is visible so that accuracy is clearly improved when batches of validated data are made available to train the model.

Figure 4

Trade-off between accuracy ($|I(d) - I^{final}|$) and timeliness ($t + day$) for the 60 consecutive months grouped by publication cell size (notice different scales).



12. All in all, statistical learning techniques should become another versatile tool for producers of official statistics so that quality can be continuously improved in many dimensions. Nonetheless, architectural, methodological, and organizational challenges lie ahead to be tackled to transform official statistical production.

IV. 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. https://github.com/david-salgado/AdvITI.

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. Available at https://www.ine.es/GS_FILES/DocTrabajo/art_doctr032022.pdf.

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 statis- tics, repealing 10 legal acts in the field of business statistics (EBS-Regulation). https://ec.europa.eu/eurostat/web/ short-term-business-statistics/legislation.