Integration of administrative and survey data in a Short-Term Business Statistics with statistical learning algorithms

Sandra Barragán[1], David Salgado[1], Sergio Pardina[1], Esther Puerto[1]

[1]S.G. for Methodology and Sampling Design, Statistics Spain, Spain

Abstract

The use of administrative data (and digital data sources) is a must not only for the modernization of the production of official statistics but also for keeping relevance in the new data and AI international ecosystem. These new data sources must be integrated with survey data. However, as it is widely known, this incorporation of new data sources does not come without quality challenges. By and large, the direct substitution, use, or aggregation of administrative data cannot be undertaken since errors both in the representation and measurement lines arise even when formerly they were under control using only survey data.

Representation errors (especially regarding coverage) arise because of unit misclassification errors and other factors. Validity, measurement, and process errors easily occur because of the administrative (non-statistical) purposes of these data sources. Overall, the fact that the data generation mechanism lies outside the control of the statistical process revives both non-sampling errors (validity error, for example) and inferential challenges (non-ignorability, for instance).

We present a proposed end-to-end statistical production process integrating administrative data with survey data in a probability sample. Synthetic values produced from a tax source are computed using a statistical learning model so that validity and measurement errors can be a priori identified and kept under control. The statistical learning algorithm learns from past and present survey and administrative data producing high-quality values for non-influential units, which paves the way to reduce response burden. Influential units are still integrated using survey data.

We share a proof of concept on the monthly Services Sector Activity Indicators using VAT data. We discuss challenges regarding the quality of both the sampling design, the statistical model, and the training data.

Keywords: data integration, administrative data, end-to-end statistical production process

1. Introduction

The integration of new data sources with survey data is essential not only for the modernization of official statistics production, but also for maintaining relevance within the new international data and AI ecosystem. These novel data sources must be incorporated alongside survey data in a proper form. However, as it is widely recognized, this inclusion of new and external data sources such as administrative data does not come without quality challenges. By and large, the direct substitution, utilization, or aggregation of administrative data cannot be undertaken. This is due to the emergence of errors in both representation and measurement, even in areas previously well-controlled using only survey data.

The use of administrative data as the primary source must maintain the same objectives as in the case of survey data, i.e., the aim is to estimate a set of population aggregates in a finite population U, defined as $Y_{U_d} = f(\sum_{k\in U_d} y_k , \sum_{k\in U_k} \boldsymbol{x}_k)$ for a collection of population domains $U_d \subset U$ (publication cells) for various target variables y and auxiliary variables $\boldsymbol{x}$. Without loss of practical generality, we can focus on population totals in the form $Y_{U_d} = \sum_{k\in U_d} y_k$, as other more complex aggregates can be expressed as functions of these totals. We assume here to have the complete sample $s = \bigcup_d s_d$ where $s_d \subset U_d$.

Let $\hat{Y}_d = \sum_{k\in s_d} \omega_{ks}(x) y_k^{\circ}$ be a linear estimator with pseudo-sampling weights $\omega_{ks}(x)$ (or genuine sampling weights if a sampling design is used), and y_k° denotes a synthetic value for variable y, which can either be a transformation of the corresponding administrative variable or a predicted value for the survey variable based on all available information (administrative and survey). Accuracy measures must also be produced.

Within this framework, the cornerstone concepts of finite population (representation line) and target variable (measurement line) remain paramount. Consequently, the Total Survey Error (TSE) model, as outlined by Groves and Lyberg (2010), maintains its validity for quality assessment purposes, even when considered within the second phase of the two-phase life-cycle model proposed by Zhang (2012).

This work is focused on short-term business statistics, specifically those incorporating tax register data as the primary source alongside survey data obtained through a probabilistic sampling design. Notably, we will describe an ongoing pilot project involving the Service Sector Activity Indicators (SSAI) survey, which has begun utilizing Value Added Tax (VAT) data from the National Tax Agency to alleviate the response burden on respondents.

Let U^{adm} denote the collection of business units contained within the tax register. U represents the finite population of analysis, derived from the population frame U_F, which itself is constructed from the business register maintained by our office. Our first concern is centred on coverage error, particularly with regards to accurately identifying administrative units $k \in U^{adm}$ that can be used as statistical units $k \in U$. From the tax register, we will consider only those units that are also contained within the population frame, denoting this intersection as $U^{mdl} = U^{adm} \cap U_F$. The target statistical variable for these units will be synthesized using the raw administrative value y_{adm} in a dedicated statistical learning model.

A preliminary approach might involve directly substituting the administrative values y^{adm} for the target survey values y^{surv}. This strategy would offer increased cost-efficiency and timeliness, contingent upon the quality of the input administrative data.

1.1 Measuring the quality of the input

Within the domain of Official Statistics, numerous proposals have been put forth to assess quality throughout the various stages of the statistical production process. However, historically, greater emphasis has been placed upon evaluating the quality of final aggregates as opposed to the quality of the input data itself. This focus can be primarily attributed to the inherent control mechanisms employed during the generation of survey data. With the burgeoning incorporation of diverse data sources, there arises a growing necessity to evaluate their quality as well. Initiatives within the European Statistical System (ESS), such as the BLUE-ETS Project (Daas et al., 2011) and the ESSnet KOMUSO project (Ascari et al., 2020, and multiple references therein), have demonstrably addressed this need.

At Statistics Spain, we have recently started to engage in the development of a diverse set of indicators aimed at evaluating the quality of data sources across multiple dimensions. Furthermore, we are conducting retrospective analyses to juxtapose administrative data with survey data at the microdata level. This endeavour entails the creation of numerical indicators and graphical comparisons to facilitate a comprehensive and rigorous evaluation process. See the work “Measuring the quality of administrative sources: at macro level with novel indicators and micro level with distributions comparison” (Nieto et al. 2024) in this conference Q2024 for more details.

2. Methodology: proposal of integration of administrative and survey data with statistical algorithms

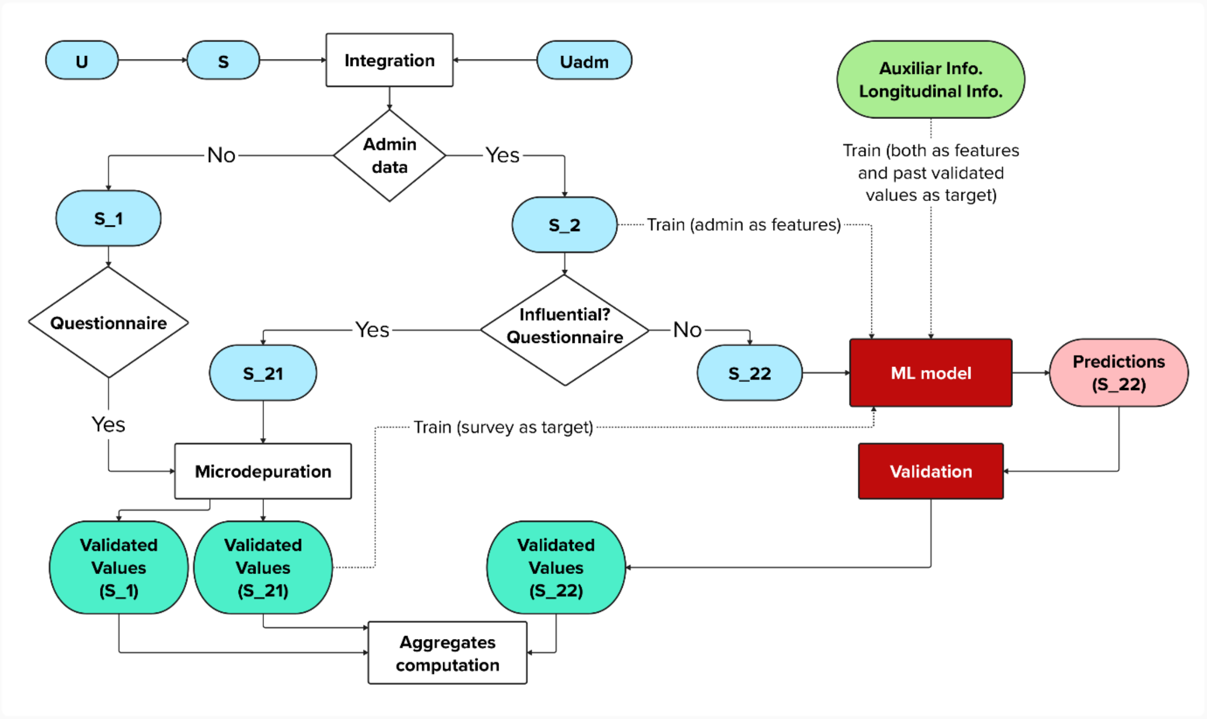
In this section the methodology developed to deal with the validity and measurement errors is exposed. It is well known (Ascari et al., 2020) that administrative data can severely differ from survey data since they are defined and collected for statistical purposes. In this sense, in the use case of the SSAI survey, the administrative total sales value declared for tax purposes, y^{adm}, may differ from the turnover survey value, y^{surv}, traditionally collected in questionnaires. These differences cautiously discourage the use of y^{adm} by mere substitution as the value of y^{surv}.

Our proposal aims at a combined use of statistical learning models and data validation techniques to control this difference (validity error, but also measurement error since we use validated microdata as training data). Longitudinal information is of special relevance as auxiliary information. Consider several datasets for reference time periods t_{-1}, t_{-2}, t_{-3}, and so on, where past periods t_{-i} will be used for model training. The proposal focuses on predicting and validating successively each dataset t_0, t_1, t_2, etc. with their past datasets.

Firstly, to initialise the recurrent modelling exercise in successive time periods, training sets for the reference period t_0 are identified with those units in the probabilistic samples and the tax register, $k \in s_{t_{-i}}^{mdl} = s_{t_{-i}} \cap U^{mdl}$. Their corresponding synthetic target variable values y_k° are the validated values entering the computation of the indices, i.e. $y_{kt_{-i}}^{\circ} = y_{kt_{-i}}^{stat}$. Then, a statistical model $y^{\circ} = f\left(y^{adm}; x\right) + \epsilon$ is adjusted using explicitly the value of the administrative variable as a feature (as part of the auxiliary information in the form of regressors). Once the model is constructed, it is used to predict the values of the variable $\hat{y}_{kt}^{\circ} = \hat{f}\left(y_{kt}^{adm}; x_k\right)$. Notice that this is the predicted value of the validated total turnover in terms of the raw administrative value of the total sales variable y^{adm} (and other features). The rest of features $\boldsymbol{x}$ are constructed following the ideas in the working paper by Barragan et al. (2022) about early estimates of the Spanish Industrial Turnover Index. These predicted values are candidates to enter the index computation. However, a data validation strategy is needed for the predictions obtained by the statistical learning model. This stage entails the design and application of error detection functions (edits) alongside their corresponding treatment methods. These treatments will likely need the implementation of a more specific imputation model. Ideally, for enhanced efficiency, this process should be automated to the greatest extent possible. The culmination of this phase will be a newly refined and validated set of synthetic target values, accompanied by the validated survey data values utilized for index computation.

This process described here is shown in the diagram of Figure 1 below.

Figure 1: Diagram of the end-to-end statistical production process with the integration of administrative data



The primary objective of employing administrative data as the principal source is to alleviate the burden on survey respondents. To achieve this goal, questionnaires should be eliminated entirely for those units possessing reliable administrative information. However, the selection of such units needs a meticulous approach, as insufficient information may obstruct the proper training of models designed to predict values for units not included within the survey sample. Several scenarios have been evaluated to differentiate between units reporting survey data and those reporting with their administrative records. The scenario presented here has demonstrably yielded the most favourable results thus far. Meticulously defined criteria have been established to identify units exhibiting erratic behaviour. This identification process guarantees the quality of their values through the traditional data collection and editing procedures. The units that are considered to be influential, which will remain under questionnaire data collection, satisfy at least one of the following criteria that are formulated in detail in Appendix 1.

- Criterion 1: Units with a high impact on the aggregate.
- Criterion 2: New units.
- Criterion 3: Units with high variability in the target variable.
- Criterion 4: High difference between the survey and administrative values.
- Criterion 5: High absolute differences between the survey value and the administrative record value.

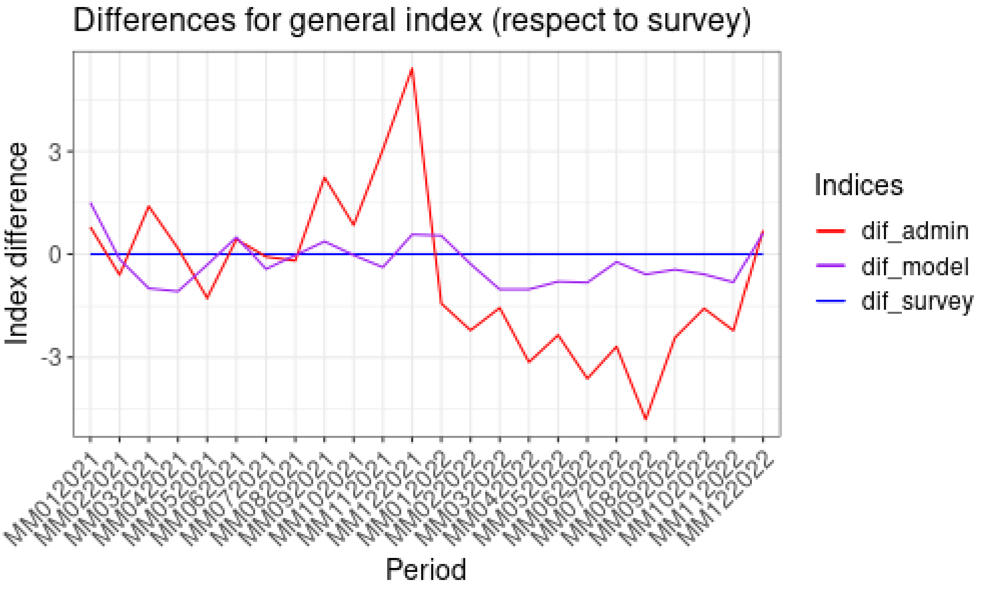
- Criterion 6: Zero values.

Units not selected according to this scoring system will be predicted by the statistical learning model. These criteria are conservative with respect to the reduction of response burden with the idea of keeping all challenging units in the survey. For example, for year 2021, there were 5281 units in the administrative dataset, and using these criteria, still 2320 would be needed to collect in the survey and 2961 could be dropped.

3. Results obtained in the use case with a Short-Term Business Statistics

As preliminary results, we show at the aggregate level the comparison between the direct substitution of the administrative value and the used of the process presented here by using statistical models. Figure 2 presents a comparison of the SSAI index obtained through direct substitution of administrative data for those units intersecting with the sample (red line) and the index obtained using model-predicted and validated values (purple line). The graph illustrates the differences between each of these indices and the index obtained exclusively from survey data. It can be observed how the differences are considerably smoothed out by using a model to account for validity and measurement errors in the statistical value based on the administrative value.

Figure 2: Comparison at aggregation level of the indices obtained with the survey, administrative and integration



4. Conclusions

In conclusion, several key points emerge from this analysis. Firstly, the observed discrepancies during the exploration of input quality strongly discourage the direct substitution of the administrative values into the statistical variable. This work proposes an end-to-end statistical production process that integrates administrative data with survey data within a probability sample. The process leverages a statistical learning model to compute synthetic values, utilizing administrative data as regressors alongside longitudinal information. Furthermore, to account for potential validity and measurement errors within administrative data, it is recommended to employ statistical learning models for response burden reduction, while concurrently implementing a selection process for units that upholds the model's quality.

This work is part of an ongoing project where a lot of issues can be solved in relation to the use of administrative data as primary source for official statistics. In fact, we can see some future work of special relevance such as: a) the development of indicators to evaluate the quality of data sources across multiple dimensions; b) the estimation of variances and mean squared errors when combining sampling designs and statistical model; c) giving solutions to the trade-off between accuracy (lack of measurement errors) and response burden reduction.

As a final highly relevant comment, we claim that new data validation methodology is needed when confronting survey and administrative values of a given target variable. The approach presented here rests on the assumption of considering validated survey values are closer to true values, since a fully-fledged statistical data editing and imputation strategy is implemented in the traditional production process to ensure data quality prior to the estimation stage. New editing and imputation strategies need to be investigated integrating administrative sources where business functions such recontacts and follow-ups for error treatment are not possible anymore.

References

Ascari, G., K. Blix, G. Brancato, T. Burg, A. McCourt, A. van Delden, D. Krapavickaite, N. Ploug, S. Scholtus, P. Stoltze, T. deWaal, and L.-C. Zhang (2020). Quality of Multisource Statistics – the KOMUSO Project. The Survey Statistician, 81,36–51.

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 SpainWorking Paper 03/22. Available at https://www.ine.es/GS_FILES/DocTrabajo/art_doctr032022.pdf.

Daas, P., S. Ossen, M. Tennekes, L.-C. Zhang, C. Hendriks, K. Foldal Haugen, F. Cerroni, G. Di Bella, T. Laitila, A. Wallgren, et al. (2011). Report on methods preferred for the quality indicators of administrative data sources.

Groves, R. and L. Lyberg (2010). Total survey error: past, present, and future. Public Opinion Quarterly 74, 849–879.

Nieto, A., Salgado, D., Barragán, S., Saldaña, S. and Rodriguez, A. (2024). Measuring the quality of administrative sources: at macro level with novel indicators and micro level with distributions comparison. Presented in the European Conference on Quality in Official Statistics.

Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. Statistica Neerlandica 66(1), 41–63.

Appendix I: Criteria for unit selection as influential units

In this appendix we describe in detail the criteria formulated to evaluate which units have to be considered influential. All the units that meet any of the following criteria enter into questionnaire data collection.

1. Criterion 1: Units with a high impact on the aggregate.

A first period-wise local score for unit k in period $t–i$ is defined as $s_{kt-i}^{(1)} = \frac{\omega_{kt-i} \cdot y_{kt-i}^{surv}}{\hat{Y}_{dt-i}}$, where $\hat{Y}_{dt-i}$ is the estimated population total for domain d in period $t–i$. The periods corresponding to the first nine months of the previous year to the reference year are used, and a first global score $S_k^{(1)}$ for each unit k is defined as the pth quantile $S_k^{(1)} = \ Q_p(s_{kt-1}^{(1)}, \dots, s_{kt-9}^{(1)})$. We have used the median (p = 0.5). A threshold is computed using a conservative elbow criterion (Tam, 2023), and, thus, units above the threshold are selected for survey data collection.

2. Criterion 2: New units.

All new units in the sample s of the previous year to the reference time period that have $\omega_{kt-i} = 1$ or annual turnover y_k^F in the population frame U_F for the preceding year greater than a chosen threshold t_F are selected. We have used $t_F = 10^7$. A second global score $S_k^{(2)}$ is thus defined as $S_k^{(2)} = I_k\big(\omega_{kt-i} = 1 \wedge y_k^F > t_F\big)$. Notice that $\omega_{kt-i} = \omega_{kt-j}$ for all $t–i$ and $t- j$ in the same year since sampling designs change only annually. Units with $S_k^{(2)} = 1$ are selected for survey data collection.

3. Criterion 3: Units with high variability in the target variable.

Another global score is defined as $S_k^{(3)} = std(\omega_{kt-1} \cdot y_{kt-1}^{surv}, \dots, \omega_{kt-9} \cdot y_{kt-9}^{surv})$. Again an elbow-based threshold is used to select those units for survey data collection.

4. **Criterion 4: High difference between the survey and administrative values.**

A time-wise score for unit k in period $t{-}i$ is defined as $s_{kt-i}^{(4)} = \frac{\omega_{kt-i} \cdot \left| y_{kt-i}^{adm} - y_{kt-i}^{surv} \right|}{\hat{Y}_{dt-i}}$. A new global score is defined as $S_k^{(4)} = std(s_{kt-1}^{(4)}, \dots, s_{kt-9}^{(4)})$. Again an elbow-based threshold is used to select those units for survey data collection.

5. **Criterion 5: High absolute differences between the survey value and the administrative record value.**

Using the same time-wise score for unit k in period $t{-}i$ as in criterion 4, a new global score is defined as the pth quantile so that $S_k^{(5)} = Q_p(s_{kt-1}^{(4)}, \dots, s_{kt-9}^{(4)})$.We have selected p = 0.5. Again an elbow-based threshold is used to select those units for survey-reporting.

6. **Criterion 6: Zero values.**

All units with any administrative record value equal to zero in the periods under consideration are selected. The global score is defined as $S_k^{(6)} = I_k\left(y_{kt-1}^{adm} = 0 \ \wedge \dots \wedge y_{kt-9}^{adm} = 0\right)$. Units with $S_k^{(6)}$ are selected for survey data collection.