UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS
Expert meeting on Statistical Data Editing
3-7 October 2022, (virtual)

Improving statistical data editing with Machine Learning: first use cases in Statistics Spain (INE)

S. Barragán and D. Salgado (Statistics Spain (INE))
sandra.barragan.andres@ine.es; david.salgado.fernandez@ine.es

I. INTRODUCTION

A. General view

1. During the last fifteen years the production of official statistics in national and international statistical offices is facing a double challenge, namely the modernization and industrialization of the production process and the incorporation of new digital data sources. The first challenge was clearly identified in the first decade of the present century [HLG-MOS, 2011] and gave rise to the advent of international production models such as the GSBPM [UNECE, 2019a], the GSIM [UNECE, 2019b], and the GAMSO [UNECE, 2019c], to name a few.

2. These standards are being gradually adopted by most statistical offices thus conforming a common working space in the international community [UNECE, 2021a]. The second challenge was already identified eight years ago [see e.g. DGINS, 2013] and is taking efforts in the form of complementary institutional initiatives [see e.g. Eurostat, 2014, 2021, UNECE, 2021b]. This challenge "will require amendments to the statistical business architecture, processes, production models, IT infrastructures, methodological and quality frameworks, and the corresponding governance structures [...]" [DGINS, 2018], i.e. basically to modify all aspects of the entire production framework.

3. In this context, the traditional production phase of statistical data editing appears as a natural candidate for the modernisation of the whole production process and the improvement of several key quality facets such as timeliness, cost-efficiency, and accuracy. The General Statistical Data Editing Model (GSDEM) [UNECE, 2021c] provides a versatile, evolvable, and comprehensive production framework to design, implement, execute, and monitor so-called statistical data strategies, i.e. a collection of sequential and concurrent business functions to detect and treat non-sampling errors. In parallel, the use of machine learning techniques with this same goal has received due attention [see Beck et al., 2018, Puts and Daas, 2021, and multiple references therein] and highly relevant international projects are already producing important empirical results [UNECE, 2022] (see especially theme 2 of work package 1 about edit and imputation).

4. Statistics Spain (INE) has been focusing during the last years on increasing the cost-efficiency of data editing in its production chain by improving the methodology of selective editing [Arbués et al., 2009, 2013, Salgado et al., 2018]. More specifically, with this formulation of selective editing, the computation of local (item) scores has been expressed as a statistical model of measurement errors,

whereas the selection and/or ranking of units arise as the solution of certain optimization problems. The construction of different models with statistical learning techniques fits very nicely in this setup, hence the exploration of use cases presented in this contribution.

B. Innovating in statistical data editing

5. Machine learning techniques offer a range of opportunities to improve multiple business functions along the whole statistical production process from data collection to data dissemination over editing, imputation, estimation, etc. Regarding statistical data editing, as stated above, the natural framework to deal with editing business functions is the Generic Statistical Data Editing Model [UNECE, 2021c]. It is in this context where we should approach the generic questions about what to improve and, then, how.

6. In our view, there exist at least three complementary facets to be taken into account, namely, the new statistical methods, the new data sources, and the development and deployment of innovation for production.

7. Regarding the new statistical methods, one can follow two immediate paths. On the one hand, a revision of existing business functions in the GSDEM can be undertaken by investigating the potential usage of machine learning techniques. For example, imputation methods with a higher-predictive capacity can be researched. On the other hand, new business functions within the review-selection-treatment classification might be devised. In both cases, it is mainly the cost-efficiency dimension of quality which will be improved.

8. Regarding the new data sources, we also identify two potential trends. On the one hand, a thorough revision of current business functions and their associated statistical methods is advised to be undertaken in the light of the new data sources. This should also be extended to the corresponding workflows for the implementation of editing and imputation strategies [see UNECE, 2021c]. For example, should web-scraped data be edited and imputed in the same spirit as traditional survey data?

9. On the other hand, however, we find it increasingly relevant the turn of attention towards input data quality in contrast to the traditional output data quality approach. There exist important novel factors affecting the data editing phase.

10. Firstly, new digital sources (administrative registers included) produce data with no statistical metadata in contrast to traditional survey data, where structural metadata are heavily taken into account during the design of the questionnaire, thus generating data with a high-quality standard. Current survey data editing business functions and workflows take advantage of this.

11. Secondly, current survey data editing methods are strongly oriented towards the use of design-based linear estimators (see, e.g., traditional item score functions in selective editing) for usual sample sizes (consider, e.g., interactive editing, not affordable for terabyte-sized datasets). This is a key ingredient supporting the output quality approach. Now, with the no-statistical-metadata data-driven paradigm brought by new data sources, data quality must be assessed from an input perspective.

12. Consider, for instance, financial transaction data coming from commercial banks. This promising data source can be potentially used and reused in many business and household statistics. It can also be linked to administrative data and to population frames and business registers. However, it does not seem meaningful to device an E&I strategy considering the precise use of these data into concrete final estimators for all cases. In contrast, in our opinion, an input data strategy should be considered

before using data further on in the statistical production chain. This suggests a new view of editing to be stressed in the GSDEM.

13. Thirdly, the continuous development and deployment in production with complete documentation, appropriate technological tools, and ready-for-use software should be a priority in the international realm to boost modernization of Official Statistics. An outstanding initiative in this direction can be found in [van der Loo and ten Bosch, 2022]. The interplay among statistical methodology, software development and deployment, and subject matter expertise should, in our opinion, be strengthened to achieve industrial standardisation.

14. In the following sections, we shall share some first experiences and reflections based on them regarding the efforts of improving data editing in the traditional output quality approach at Statistics Spain (INE).

II. USE CASES

A. Computation of local (item) scores

15. Local (item) scores are basically a measure of the degree of suspicion for a questionnaire item (variable) to contain a measurement error ($y_k \neq y_k^{\text{true}}$) [see e.g. de Waal et al., 2011]. Traditionally, it is computed by means of a choice of a local (item) score function $s(\cdot)$ on anticipated values $\hat{y}_k$ (usually low-quality predicted values) and raw variable values y_k^{raw} (collected values) such as, e.g., $s_k = s(\hat{y}_k, y_k^{\text{raw}}) = d_k \cdot |y_k^{\text{raw}} - \hat{y}_k|$, where d_k stands for the sampling design weight of unit k.

16. In the optimization approach proposed by Statistics Spain (INE) some time ago [Arbués et al., 2013], these scores arise as fitted values of a measurement error model:

$$s_k = \mathbb{E}\left[d_k |Y_k^{\text{raw}} - Y_k^{\text{true}}| \middle| \mathbf{x}_k\right], \tag{1}$$

where $\mathbf{x}_k$ stands for any available information used in the model. Equation (1) paves the way for a natural use of machine learning models [Hastie et al., 2009, Murphy, 2013].

17. We have applied these ideas on Short-Term Business Statistics, where historical time series are available. For continuous variables, using these historical time series of raw and validated values, we can define the absolute error $\epsilon_k = |y_r^{\text{raw}} - y_k^{\text{val}}|$ and construct a statistical model using the available information, mainly the raw values in the past $y_{kt-1}^{\text{raw}}, y_{kt-2}^{\text{raw}}, \ldots$ and their errors $\epsilon_{kt-1}, \epsilon_{kt-2}$, other variable values (e.g. number of employees), and especially data collection and data editing paradata from preceding periods (collection unit ID, data collection mode, day of the month of data collection, interview, etc).

18. However, this is no less than a predictive sophistication of the traditional approach. More interestingly, equation (1) can be further applied also to ***categorical*** variables. In this case, the variables $Y_k^{\text{raw/true}}$ are indeed indicator variables expressing class membership to estimate population class totals [Särndal et al., 1992], so that ϵ_k reduces to a binary variable denoting presence/absence of error. Furthermore, we can write

$$s_k = \mathbb{E}\left[d_k |Y_k^{\text{raw}} - Y_k^{\text{true}}| \middle| \mathbf{x}_k\right] = d_k \cdot \mathbb{P}\left[\epsilon_k = 1 \middle| \mathbf{x}_k\right], \tag{2}$$

and the computation of this probability can be undertaken by means of a classification problem in machine learning.

19. This approach has been applied to rank questionnaires for interactive editing in the last edition of the European Health Interview Survey in Spain [Barragán et al., 2020] reducing the number of questionnaires to revise manually. This use case focused on selecting questionnaires where the variable `occupation code` seems to have a measurement error. The selected statistical model was a random forest where the target variable was the error indicator (binary variable) in the occupation in preceding editions of this survey and the Spanish National Health Interview Survey (with nearly identical structure). The regressors $\mathbf{x}_k$, according to the subject matter expert, were those questionnaire items semantically connected to occupation such as educational attainment, economic activity of job position, wage, age, ... Data collection proceeded in batches, which were processed by subject matter experts using the sorting order provided by the local scores of this target variable. Every time a batch was completely edited (errors detected and corrected), the model was re-trained with the latest information to avoid drift. In the past, the whole sample (around 22000 questionnaires) was manually revised in no particular order. In this use case, on average the first half of the sorted sample already contained 75% of all measurement errors related to the `occupation code`.

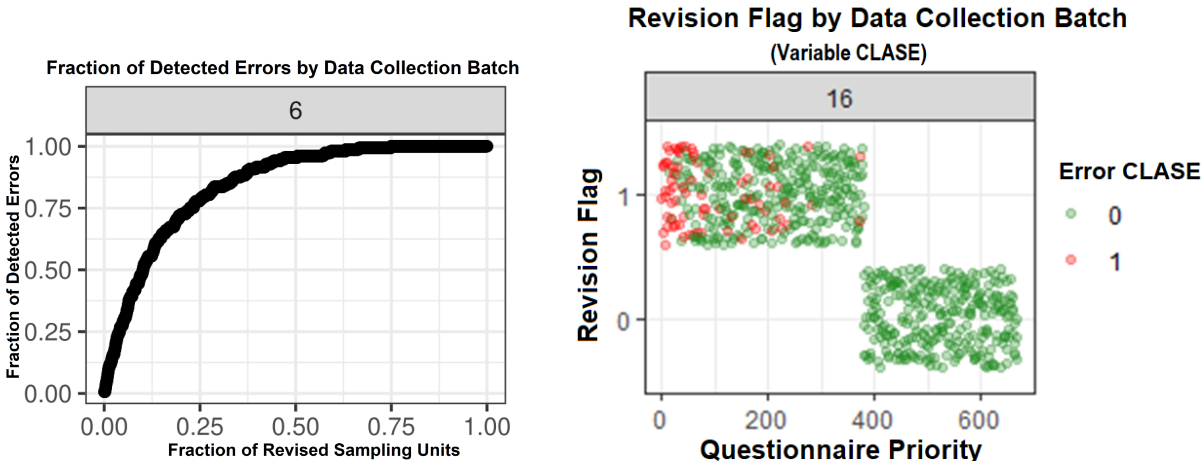

FIGURE 1. Detected errors according to sorting order and revised questionnaires (only one data collection batch).

B. Semicontinuous variables

20. A closer reflection on the approach introduced above drives immediately to the consideration of semicontinuous variables, i.e., of variables with values equal to 0 or in a continuous range. As a matter of fact, absolute values of measurement errors of continuous variables (ϵ_k above) are semicontinuous variables with values equal to 0 or in $\mathbb{R}^+$. A more adequate predictive model is needed to gain in efficiency.

21. Machine learning techniques provides a solution by combining a classification problem with a regression problem [Bohnensteffen, 2020]. The bottom line can be sketched as in figure 2. For units with missing values, a separate regression model is needed, since we miss one of the most important regressors, i.e. y_k^{raw}. For units with non-missing values we firstly build a classification model to decide whether ϵ_k is 0 or not. Finally, for those classified as having non-null measurement error, a new regression model is built now including as regressor the raw value of the variable under analysis (y_k^{raw}).

22. Under the same methodology, now the score values for each sampling unit k can be readily computed as

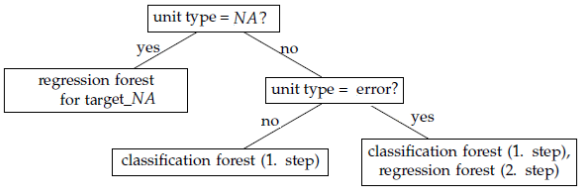

FIGURE 2. Two-stage approach to model semicontinuous variables including missing values.

$$s_k = d_k \cdot \mathbb{P}\left[\epsilon_k > 0 \middle| \mathbf{x}_k\right] \cdot \mathbb{E}\left[\epsilon_k \middle| \epsilon_k > 0, \mathbf{x}_k\right], \tag{3}$$

where the probability is computed according to the classification model and the conditional expectation represents the fitted values of the second regression model.

23. The statistical units are sorted out according to these local scores. As a figure of merit, as usual, we make use of the relative pseudobias in absolute value, i.e.

$$ARB\left(\widehat{Y}(n_{ed})\right) = \frac{|\widehat{Y}(n_{ed}) - \widehat{Y}^0|}{\widehat{Y}^0},$$

where $\widehat{Y}(n_{ed})$ stands for the estimator when n_{ed} units have been edited (errors detected and treated) and $\widehat{Y}^0$ denotes the estimator when all units have been revised ($n_{ed} = n$).

24. Using the historical time series of raw and edited values in the Spanish Service Sector Indicators Survey, we have trained and tested this proposal where the main results can be summarized in figure 3. When missing values are not considered, the model detects very rapidly the errors. However, the model for missing values needs further improvement, although already within the first half of the sorted-out sample all errors are detected.

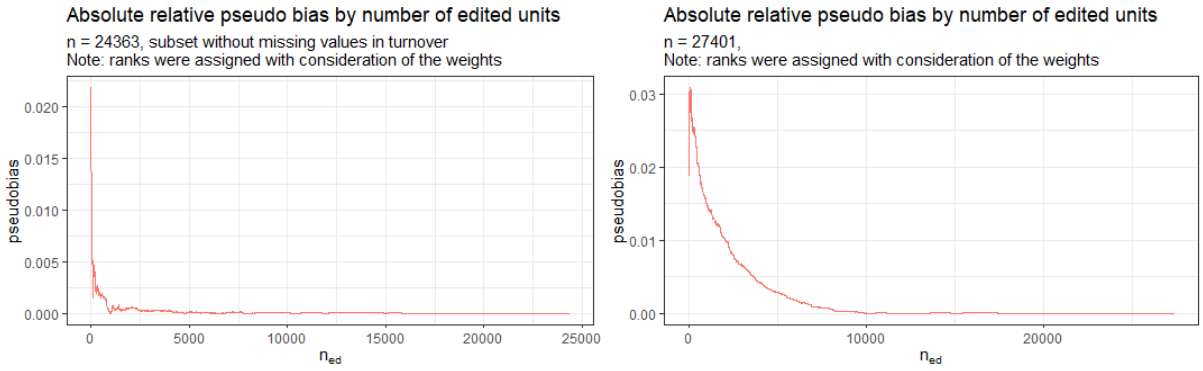

FIGURE 3. Absolute relative pseudobias for the turnover excluding units with missing values (left) and including units with missing values (right).

C. Imputation to complete information during collection: nowcasting

25. Imputation is a strategic production task in statistical data editing to deal with missing and erroneous values [see e.g. Dagdoug et al., 2021, for a recent study with machine learning algorithms]. Indeed, imputation can be approached in multiple ways [UNECE, 2022] and, in this sense, we have

carried out a mass imputation exercise with a gradient boosting algorithm over units with variable values not yet collected during the data collection phase to produce early estimates of the Spanish Industry Turnover Index, thus improving timeliness.

26. The model takes advantage of both past and ongoing collected information to produce more accurate predictions. The bottom line for producing early estimates is two-fold. On the one hand, use the prediction model to generate estimated values $\widehat{y}_k$ for those sampling units k $s-r(t)$ not yet collected in the sample $r(t)$ at time t. On the other hand, divide regressors into past-time and current-period. For the former, use information at unit level whereas for the latter, use aggregated information. The combination of both types produce accurate enough estimated values so that we can regenerate the total turnover in each population domain U_d under study[1]

$$\widehat{Y}_d = \sum_{k \in r_d(t)} y_k + \sum_{k \in s_d - r_d(t)} \widehat{y}_k.$$

27. Importantly enough, the use of statistical models allows us to provide an accuracy indicator (root mean square error) of every early estimate of every disseminated index. See figure 4 for the general index. The more data are collected, the more accurate the early estimates. Thus, a quantitative trade-off between timeliness and accuracy can be established.

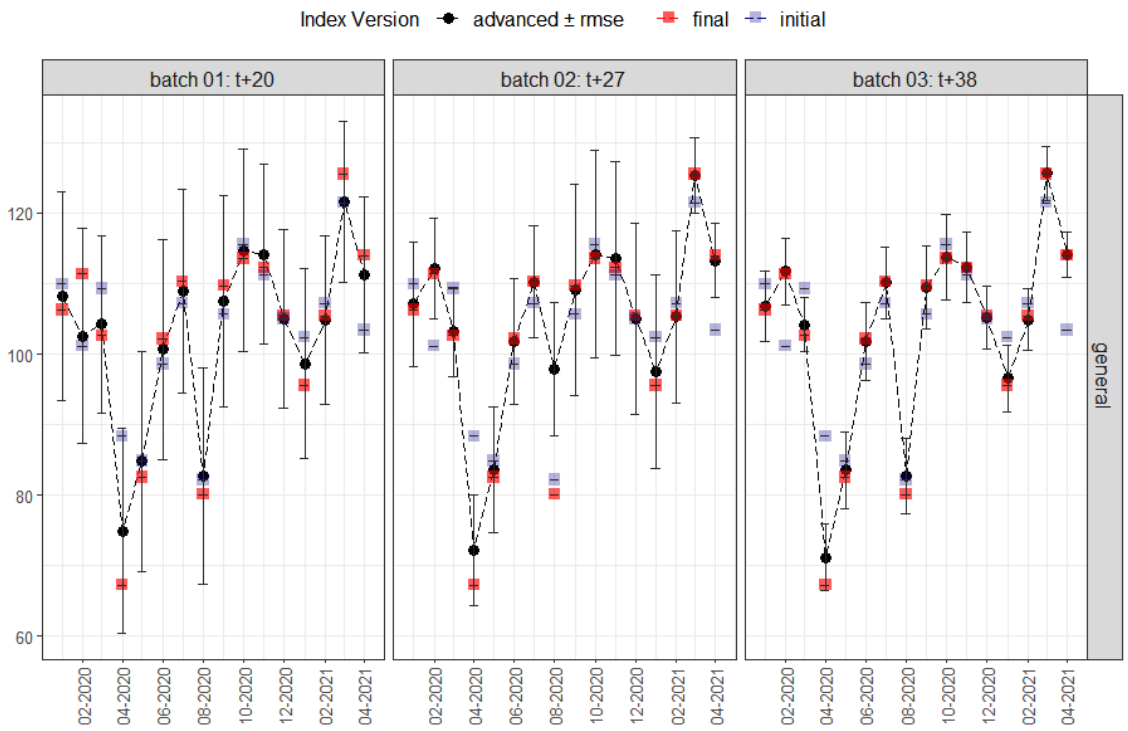

FIGURE 4. Early estimates of the Spanish Industrial Turnover Index (black) compared with the disseminated version (red) and an initial estimate without data from the reference time period (blue). Official release is at $t + 51$ on average.

[1]In our use case (the Spanish Industrial Turnover Index Survey), a cut-off sampling design is used so that the sampling weights are 1 by construction.

D. Imbalanced data

28. Another immediate observation when dealing with measurement errors is that they are deeply imbalanced [see e.g. López et al., 2013], i.e. no more than 10% of statistical units report erroneous values. This is valid both for business and household surveys as those introduced above. Machine learning algorithms must then be analysed accordingly and possibly adjusted.

29. We have explored three approaches, namely (i) undersampling, (ii) oversampling, and (iii) cost-sensitive learning. Undersampling amounts to taking a sample of those units with null measurement error to balance both classes. On the contrary, oversampling means sampling with replacement those units with non-null measurement error to reach a balance between both classes. Finally, cost-sensitive learning assigns a cost to each of the four possible classification results (true/false positive/negative).

30. In our analysis we have made the following choice for the costs in the latter approach:

		predicted	
		1	**0**
true	**1**	0	c
	0	1	0

31. This amounts to assigning a unitary cost to an overedited questionnaire (selected for editing but having no error) and to assigning an arbitrary cost $c > 0$ (to be fixed and analysed) to an erroneous value not selected for editing (typically $c > 1$). Under these choices, for the case with categorical cases in the European Health Interview Survey mentioned above, the local scores can be computed as

$$s_k = \begin{cases} d_k \cdot \mathbb{P}\left(\epsilon_k = 1|\mathbf{x}_k\right) \cdot c & \text{if } \mathbb{P}\left(\epsilon_k = 1|\mathbf{x}_k\right) \leq \frac{c}{1+c}, \\ d_k \cdot \mathbb{P}\left(\epsilon_k = 0|\mathbf{x}_k\right) & \text{if } \mathbb{P}\left(\epsilon_k = 1|\mathbf{x}_k\right) > \frac{c}{1+c}. \end{cases}$$

32. To select c, we compute the quartiles of $\frac{\mathbb{P}(\epsilon_k=1|\mathbf{x}_k)}{1-\mathbb{P}(\epsilon_k=1|\mathbf{x}_k)}$ to obtain three values $c_1 = 1.4$, $c_2 = 2.4$, and $c_3 = 5.7$. We divide the whole sample into train (80%) and test data sets (20%), sort out the units in decreasing order of s_k and compute the absolute relative pseudobias. We repeat this procedure 30 times and compute the average. Figure 5 represents the performance of the local scores in terms of

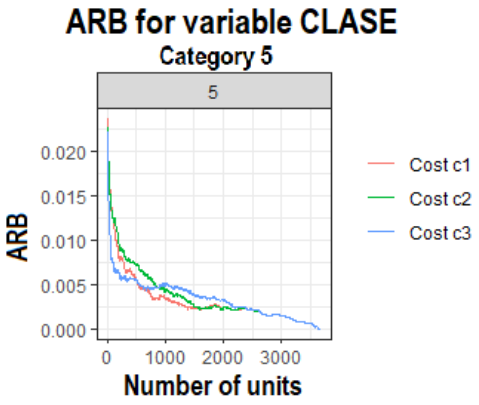

FIGURE 5. Absolute relative pseudobias in terms of the number edited units. Category 5 of the occupation variable `CLASE`.

the absolute relative pseudobias for a concrete category of the occupation variable `CLASE`. We observe how the larger value of c drives us to a steepest descent of the pseudobias.

33. In figure 6 we represent the average descent of the pseudobias comparing the four alternatives (no imbalanced data treatment, undersampling, oversampling, and cost-sensitive learning).

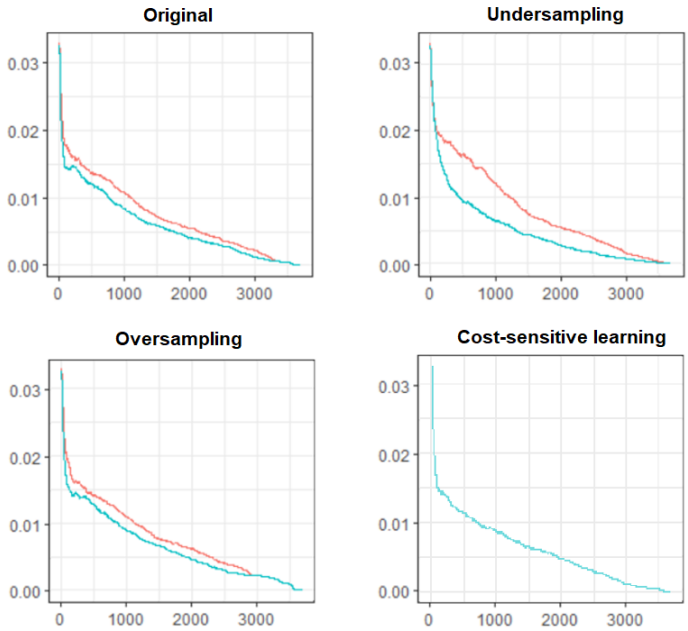

FIGURE 6. Averaged decrease of pseudobias per method. Red and blue lines are computed without/with the design weights d_k, respectively.

E. A metric to measure quality of unit prioritization

34. At this point it is evident that we need a quantitative measure to compare the quality of different unit prioritizations, independently of the underlying statistical method to compute the prioritization.

35. We can inspire ourselves by the area under the ROC curve (AUC) [see e.g. Fawcett, 2006] and propose to compute the area under the ARB cumulative curve after normalizing the horizontal and vertical axes by expressing the coordinates in terms of fractions of the number of edited units and of the ARB, respectively. In detail, we firstly define the horizontal coordinates x_k, $k = 0, 1, \ldots, n$ as $x_k = k/n$. Secondly, we define the vertical coordinates y_k as

$$y_k = \frac{\sum_{i=0}^{k} ARB(n_{ed} = i)}{\sum_{i=0}^{n} ARB(n_{ed} = i)},$$

where $ARB(n_{ed}) = \frac{|\widehat{Y}(n_{ed}) - \widehat{Y}^0|}{\widehat{Y}^0}$. Next, we define the (polygonal) curve with points $\{(x_0, y_0), \ldots, (x_n, y_n)\}$.

36. We propose to use the area under the cumulative curve as a quality indicator of the prioritization of units (see figure 7). Notice the following properties: (i) $0 \leq AUCC \leq 1$, (ii) the fastest the descent of ARB, the greatest the value of AUCC, (iii) it is independent of how the prioritization has been computed.

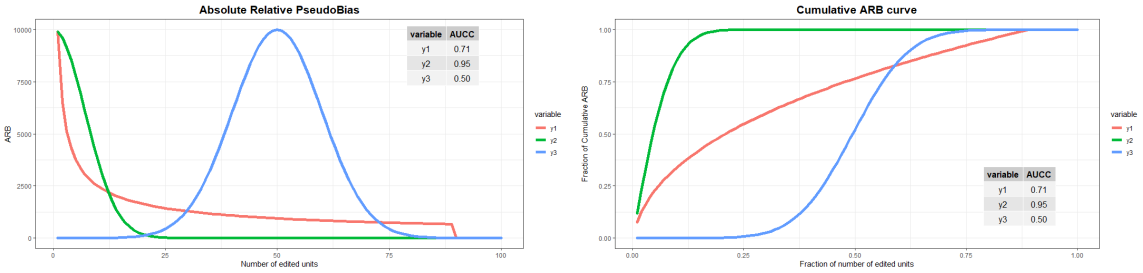

FIGURE 7. Absolute relative pseudobias and area under the cumulative curve (AUCC).

F. NLP on questionnaire comments and remarks

37. During data collection and microdata editing production tasks, many paradata are generated which prove to be extremely useful for later production phases such as macro editing, outlier detection and analysis, and estimation. Some of these paradata amounts to remarks and comments by respondents and/or editing clerks. Natural language processing stands up as a natural technique to go beyond plain human reading to make an optimal usage of this information. A very first preliminary attempt has been undertaken with the Spanish Industrial Turnover Index Survey, where around 12000 questionnaires are collected every month. Highly valuable information appears as clerical comments and remarks to the collection and editing tasks. Subject matter experts need to read these texts to incorporate this information into the production process.

38. In this first exploratory analysis [Mata, 2021], we proceeded as follows in steps: (i) we preprocess the collection of remarks assigned to each single questionnaire by lowercasing all the text, removing stopwords, and substituting personal data such as mobile phone numbers, person names and surnames, tax ID codes,... into generic labels such as *mobilephonenumber*, *name*, *taxID*, ...; (ii) we tokenize the generalised texts obtained above; (iii) we apply the hash trick to code all tokens to limit the size of the bag of words, and (iv) we build a random forest model with the target variable $\epsilon_k \in \{0, 1\}$ (questionnaire revised or not) and the coded tokens as regressors. The analysis has been performed for 1-grams and 2-grams. The ROC curves yield AUCs of 0.5923 and 0.5732 for 1-grams and 2-grams, respectively, thus suggesting that much research is needed in this direction (ensemble methods, more efficient preprocessing, deep learning techniques, etc.). Even with this naive approach, a weak signal is indeed detected with 1-grams (AUC$\approx$0.60).

III. SOME CONCLUSIONS AND REFLECTIONS

39. Beyond technical and methodological details, these first experiences with diverse degree of maturity have provided us with important insights to incorporate machine learning techniques into daily production tasks.

40. Firstly, we foresee an increasing need to have a close collaboration among subject matter experts, methodologists, and software developers. This arises not so much for the predictive part of these models, but especially for the information representation challenge to make an optimal use of them. For example, in the nowcasting exercise in section C and the score-value computation in section A, the selection of primary and derived variables to be used as regressors has been fundamental, since they incorporate subject-matter knowledge.

41. Traditional survey data editing comprises a non-negligible amount of interactive intervention (the so-called interactive editing modality), thus entailing the well-known resource consumption and timeliness restriction of this production phase. Machine learning is sometimes viewed as an opportunity to reduce this manual work. Our experience suggests that the expert intervention beyond statistical models will still be necessary, even with more sophisticated models and tools, especially to control model drift.

42. However, the incorporation of new data sources (administrative registers and digital sources) bringing extraordinarily large data sets will affect the interplay between editing and estimation. In traditional survey data, editing is clearly focused towards the design-based estimation (sample selection + linear estimator), concentrating expert resources on influential records [see UNECE, 2021c]. We foresee that with new sources and new estimation techniques, sample selection methods will be needed to be applied to rationalize resources for interactive editing: having data sets close to frame population size, now the problem of resources for interactive editing intensifies, especially in Short-Term Business Statistics.

43. As mentioned above, input data quality, intricatily related to editing and validation, is increasingly becoming more important with the incorporation of new data sources. Paradoxically, if the production process is designed to be modular [see e.g. UNECE, 2019a], the input of a production process step should be the output from the preceding step, so that input quality should derive from output quality. This means that knowledge of the preceding production tasks (data generation and data collection in the case of editing) will be necessary. This impinges clearly on the intricate issue of access to privately held data for new digital sources and to administrative register design and collection for admin data.

44. Finally, the renovation of professional skills bringing the gap between subject matter, statistical methodology, and software development becomes more urgent. Hiring new staff is only a marginal solution and training programmes for long-term production staff are necessary.

References

Arbués, I., González, M., and Revilla, P. (2009). Selective editing as a stochastic optimization problem. *Boletín de Estadística e Investigación Operativa*, 25(1) 32–41. Available at http://www.seio.es/BEIO/files/BEIOv25n1_EO_I.Arbues+M.Gonzalez+P.Revilla.pdf.

Arbués, I., Revilla, P., and Salgado, D. (2013). An Optimization Approach to Selective Editing. *Journal of Official Statistics*, 29(4) 489–510. doi: https://doi.org/10.2478/jos-2013-0037.

Barragán, S., González-García, M.R., Salgado, D., and Vázquez, T. (2020). Selective Editing of Categorical Variables with Random Forests: Results of the Implementation in the Statistical Production. 8th Conference on the Use of R in Official Statistics, 2-4 December, 2020, Vienna (Austria).

Barragán, S., Rosa-Pérez, E., and Salgado, D. (2022). Early provision of economic short-term indicators using Machine Learning. 10th European Conference on Quality in Official Statistics, 8-10 June, 2022, Vilnius (Lithuania).

Beck, M., Dumpert, F., and Feuerhake, J. (2018). Machine Learning in Official Statistics. *arXiv.1812.10422*. doi: 10.48550/ARXIV.1812.10422. Available at https://arxiv.org/abs/1812.10422.

Bohnensteffen, S. (2020). Selective Data Editing of Continuous Variables with Random Forests in Official Statistics. EMOS master thesis at Complutense University of Madrid Available at https://eprints.ucm.es/id/eprint/63245/.

Dagdoug, M., Goga, C., and Haziza, D. (2021). Imputation Procedures in Surveys Using Nonparametric and Machine Learning Methods: an Empirical Comparison. *Journal of Survey Statistics and Methodology*, 00 (September, 2021), 1-48. ISSN 2325-0984. doi: https://doi.org/10.1093/jssam/smab004. Available at https://academic.oup.com/jssam/advance-article-pdf/doi/10.1093/jssam/smab004/40215471/smab004.pdf.

de Waal, T., Pannekoek, J. and Scholtus, S. *Handbook of Statistical Data Editing and Imputation*. Wiley, Amsterdam, 2011.

DGINS (2013). The Scheveningen Memorandum. Technical report, European Union, 2013. Available at https://ec.europa.eu/eurostat/cros/news/scheveningen-memorandum-big-data-and-official-statistics-adopted-essc_en.

DGINS (2018). The Bucharest Memorandum. Technical report, European Union, 2018. Available at https://ec.europa.eu/eurostat/web/european-statistical-system/-/dgins2018-bucharest-memorandum-adopted.

Eurostat (2014). Vision 2020 Implementation Portfolio, 2014. Available at http://ec.europa.eu/eurostat/web/ess/about-us/ess-vision-2020/implementation-portfolio.

Eurostat (2021b). Big Data, 2021. Collaboration in Research and Methodology for Official Statistics. Available at https://ec.europa.eu/eurostat/cros/content/big-data_en.

Fawcett, T. (2006). An introduction to ROC analysis *Pattern Recognition Letters* 27(8), 861–874. doi: 10.1016/j.patrec.2005.10.010. Available at https://www.sciencedirect.com/science/article/pii/S016786550500303X.

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

HLG-MOS (2011). Strategic vision of the High-Level Group for strategic developments in business architecture in Statistics. *Conference of European Statisticians, Geneva, 14-16 June*, 2011.

López, V., Fernández, A., García, S., Palade, V., and Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences* 250, 113–141. doi: 10.1016/j.ins.2013.07.007.

Mata-Núñez, J.A. (2021). Uso de Procesamiento de Lenguaje Natural para la Generación de Variables a Partir de Paradatos de Recogida. Degree thesis (in Spanish). Faculty of Mathematics, Complutense University of Madrid. Unpublished.

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

Puts, M. and Daas, P.J.H. (2021). Machine Learning from the Perspective of Official Statistic. *The Survey Statistician* 84, 12–17. Available at http://isi-iass.org/home/wp-content/uploads/Survey_Statistician_2021_July_N84_02.pdf.

Salgado, D., Esteban, M.E., and Saldaña, S. (2018). SelEdit: a collection of R packages to implement some optimization-based selective editing techniques. *Romanian Statistical Review*, 4(Dec) 19–38. Available at https://www.revistadestatistica.ro/wp-content/uploads/2018/12/rrs4_2018_A03.pdf.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model assisted survey sampling*. Springer, New York, 1992.

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

UNECE (2019b). Generic statistical information model v1.2, 2019. Available at https://statswiki.unece.org/display/gsim.

UNECE (2019c). Generic Activity Model for Statistical Organizations v1.2, 2019. Available at https://statswiki.unece.org/display/GAMSO/GAMSO+v1.2.

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

UNECE (2021b). Big Data, 2021. Available at https://unece.org/statistics/ces/big-data.

UNECE (2021c). Generic Statistical Data Editing Model. Available at https://statswiki.unece.org/display/sde/GSDEM.

UNECE (2022). HLG-MOS Machine Learning Project. Available at https://statswiki.unece.org/display/ML/HLG-MOS+Machine+Learning+Project.

van der Loo, M. and ten Bosch, O. (2022). Awesome official statistics software. Available at https://github.com/SNStatComp/awesome-official-statistics-software.