UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS
Expert Meeting on Statistical Data Editing
7-9 October 2024, Vienna

Moving towards the standardized process of automatic statistical data editing using machine learning techniques

Ieva Burakauskaitė (State Data Agency (Statistics Lithuania), Lithuania)
ieva.burakauskaite@stat.gov.lt

I. Introduction

1. Statistical data editing and imputation (hereinafter referred to as E&I) is a significant but time-consuming process during the production of official statistics at National Statistical Institutes. In order to increase the efficiency of E&I, the Generic Statistical Data Editing Model (GSDEM) offers some valuable insights on various steps during the latter process such as the detection of the most influential errors using selective editing and the error treatment with either interactive or preferably automatic editing.

2. As stated in the description of the GSDEM, E&I "is prone to be influenced by innovative procedures like machine learning" (UNECE, June 2019), hence, E&I process is often considered for improvement when it comes to the modernization of official statistics. Properly implemented machine learning algorithms enables the wider usage of automatic editing and reduce the extent of interactive editing while only focusing on the most influential units in the population. Such a shift in the E&I methodology might shorten data editing time and increase process efficiency.

3. The working document is organized as follows. Section II gives the motivation behind the standardization and automatization of the E&I process at State Data Agency (Statistics Lithuania) (hereinafter referred to as SDA), naming a few desirable outcomes of such a modernization. Section III introduces the current E&I process at SDA which is the subject of the latter modernization. Section IV overviews the envisioned standardized automatic E&I process, and outlines the three process development phases from Design and Implementation to Production. Sections V and VI provide a short overview of two case studies carried out by the academic community – outlier detection and outlier correction, respectively. These studies were initiated by SDA as the social partner for the final master's degree theses. Finally, some closing remarks are given in Section VII.

II. Motivation

4. The E&I process at SDA has already been a subject of modernization. At first, it was centralized for a number of statistical surveys, as Methodology and Data Science Group was entrusted with the outlier detection task. A set of outlier detection methods were implemented, that is, various deterministic rules, Hidiroglou-Berthelot method, selective editing (see Section V(30) for a more detailed description). Then, the method for selective editing was refined even further considering the findings of Burakauskaitė and Nekrašaitė-Liegė (2022).

5. However, the outlier treatment (correction) task still needs to be refined and automated. Until now, Statistics Divisions are fully responsible of the latter part of the E&I process. Although Methodology and Data Science Group have initiated the necessary case studies to find the best working methods for the outlier correction, e.g., Uogelė (2023), and is still continuing the related research, a better solution for the integration of the latter methods into the production process of statistical information is needed.

6. The migration of statistical surveys into the uniform platform (Palantir Foundry) has motivated SDA to utilize the latter platform as the solution for the integration of the E&I process. It also prompted to re-evaluate the current E&I process, as such a platform offers an opportunity not only for the integration but also for the standardization of the process.

7. The efficiency of E&I might be improved as the uniform platform enables an easier and safer data file exchange, a convenient user interface for Statistics Divisions, e.g., for those employees with very minimal or no programming knowledge, as well as the reduced data editing time and re-contact with respondents, as only the influential outliers might be flagged during the selective editing.

8. Hence, a few desirable outcomes are expected from the modernization of the E&I process through its integration into such a uniform platform, that is,

(a) Efficient resource allocation. Employees would be able to focus on data analysis more instead of spending the majority of time on manual review / follow-up of the collected observations.
(b) Faster production of official statistics. Less time would be spent on data editing, hence, on the preparation of statistical information in general.
(c) Lower response burden. Re-contact counts would be minimized.
(d) Increase in quality of statistics. The decrease in manual editing would lessen the chances of human error.

III. Current statistical data E&I process

9. Figure 1 below depicts the E&I process which is currently implemented at SDA. The following points 10–14 of Section III describe each process step in detail using GSDEM terms, as the latter framework has already been used for reference during the previous centralization of the E&I process at SDA, mentioned in Section II(4).

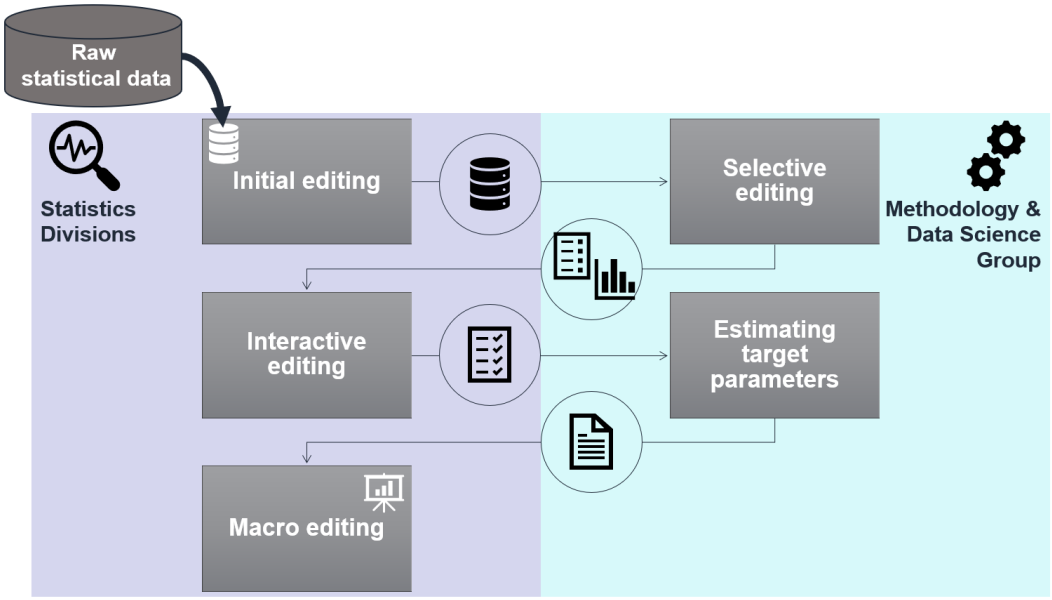

Figure 1: Current statistical data E&I process at SDA.

10. "Initial editing" is carried out by Statistics Divisions and is comprised of

(a) Domain editing, that is, checking structural components that define the population and variables, e.g., verifying classification variables such as NACE codes.
(b) Editing systematic errors, which includes addressing both obvious and systematic errors. The latter ones are harder to detect, however, treating these errors at the early process step ensures more reliable statistical data.

11. "Selective editing", or outlier detection in general, is carried out either by Statistics Divisions using mostly deterministic edit rules, or by Methodology and Data Science Group using

(a) Deterministic rules prepared by the responsible Statistics Division based on the expert knowledge and experience from the previous periods of the corresponding statistical survey.
(b) Mathematical methods developed and tested by Methodology and Data Science Group. The idea behind such mathematical methods for selective editing is to identify the most influential outliers that could greatly impact the target estimate. Hence, it focuses on selecting an optimal subset of units, reducing the extent of a costly interactive editing (preventing overediting) while maintaining the quality of the estimate of interest.

12. "Interactive editing", or outlier analysis and correction, is carried out by Statistics Divisions. Statistical data are checked for errors using expert knowledge and experience. If needed, adjustments are made applying deterministic rules, basic mathematical methods or re-contact.

13. "Estimating target parameters" is carried out either by
(a) Methodology and Data Science Group for sample-based surveys, or
(b) Statistics Divisions for census-based surveys.

14. "Macro editing" (also referred to as output editing or selection at the macro level in the description of the GSDEM) carried out by Statistics Divisions. It includes the identification of units in a statistical data set that may contain influential outliers by analyzing population aggregates or estimates. Comparison might be performed
(a) within statistical data set itself,
(b) with external sources, or
(c) with historical information.

15. After the E&I process steps portrayed in Figure 1, one additional action might take place to ensure the quality of the output (statistical information). If the final aggregates or estimates do not pass the macro edit rules set by Statistics Divisions, it is usually returned to the "Initial editing" process step and the procedures of Figure 1 are repeated.

IV. First steps towards the improved statistical data E&I process

16. Our aim is to further the integration of the E&I process according to the Generic Statistical Business Process Model (GSBPM) described in UNECE (January 2019), and continue the standardization started by adopting the GSDEM as the reference framework, while developing the E&I process in the uniform platform.

17. In order to standardize the E&I process according to GSDEM, we follow the characterization of the E&I process as the execution of three tasks: review, selection and / or treatment. Hence, three classes of functions are considered (Pannekoek and Zhang, 2012), that is,
(a) Review: Functions that examine the data to identify potential errors. It usually employs a set of quality indicators or measures (edit rules) that indicate specific outliers in the statistical data.
(b) Selection: Functions that select units or fields within units that may need further treatment, i.e., to be adjusted or imputed. Such functions use the results of "Review" and, based on some chosen selection criteria (thresholds) and statistical data as inputs, produce indicators identifying units or fields within units to be passed over to the next class of functions for treatment.
(c) Treatment: Functions that change selected data values to improve the data quality. It means changing or imputing the selected data values in order to treat the outliers detected earlier. The edited statistical data set may then become an input for another round of "Review".

A. E&I process standardization and automatization phases

18. Figure 2 below portrays the standardized automatic E&I process (also referred to as the improved E&I process) which is envisioned at SDA. The following points 19–21 of Section IV briefly describe each improved E&I process implementation phase to give a better understanding on its basic objective.

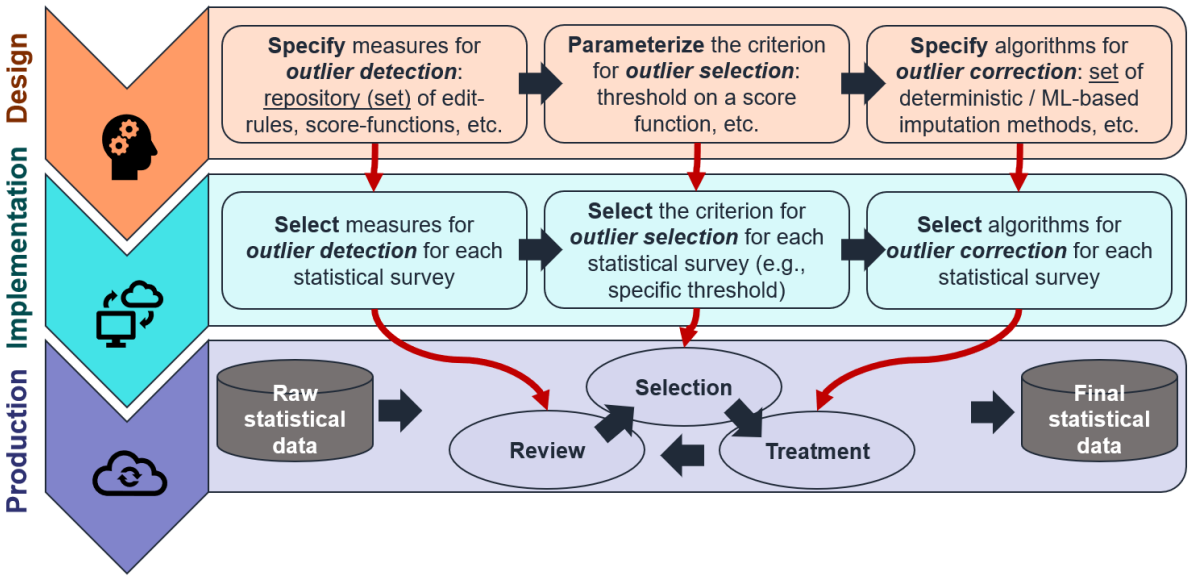

Figure 2: The goal: Standardized automatic statistical data E&I process.

19. "Design". During the first phase, a technical report describing statistical data validation, editing and imputation strategies is prepared by Statistics Divisions and Methodology and Data Science Group for each statistical survey. These technical reports are then combined to identify a unique set of E&I methods that could be applied to statistical data at SDA.

20. "Implementation". The repository of data validation, editing and imputation methods is developed, based on the strategies envisioned by Statistics Divisions and Methodology and Data Science Group during the "Design" phase. It might include both "traditional" and machine learning based (ML-based) algorithms. The parameterization of each method offers an opportunity for the customization of the strategies. However, as the choice of parameters can have a large impact on the quality of E&I, default options are set by Methodology and Data Science Group during the "Design" phase. The latter repository is integrated into the uniform platform, where a user interface for performing E&I is developed by Information Technology (IT) and State Data Governance Information System Divisions and tested together with Statistics Divisions and Methodology and Data Science Group.

21. "Production". When the goal of working automated data editing tool is reached, the E&I related tasks are performed using the developed parameterized method repository through the user interface of a uniform platform. These E&I tasks, that is, Review, Selection and Treatment type functions, lead to the edited statistical data set – transformed (before macro editing) and final (after passing the macro edit rules) statistical data.

22. To assess and ensure E&I quality, standardized reports both on statistical data as well as E&I process quality might be generated using the output metadata, i.e., paradata. It contains such measurements concerning the quality of the input, transformed and output data, as imputation rates, number of edit rule failures, etc.

23. In order to reach the goal of the improved E&I process, collaboration between various divisions at SDA (i.e., IT Division, State Data Governance Information System Division, Methodology and Data Science Group, and Statistics Divisions) is essential:

(a) IT infrastructure and software solutions are needed to develop the working statistical data editing tool and enable the automatization of the E&I process.
(b) Development of deterministic, mathematical methods and machine learning techniques are needed to perform outlier detection and correction tasks. To this end, collaboration with the academic community is important, for instance, SDA acts as a social partner for the final master's degree theses, suggesting topics related to selective editing, missing value imputation, etc. Such a collaboration leads to various case studies, see Sections V and VI for a few examples.
(c) Expert knowledge and experience are needed to assess the quality of E&I machine learning techniques compared to the "traditional" methods.

B. E&I flow model: Example for business statistics surveys

24. Figure A1 in Appendix A contains the envisioned improved E&I flow model for business statistics surveys based on the examples of generic statistical data editing flow models given in the description of the GSDEM. The given flow model may be partitioned into three parts, see points 25–27 below.

25. Similarly to the current E&I process at SDA given in Figure 1, the first process step is "Initial editing" which includes Domain editing and Editing systematic errors as it was described in Section III(10). The idea behind this process step is to first adjust for errors that can be reliably treated with a small cost.

26. Then, after the "Selective editing" process step, which selects the most influential outliers for further treatment, E&I flow model splits into two branches based on whether the survey is of structural business statistics or not, that is,

(a) For structural business statistics, only a set of the most influential units containing outliers are treated during the "Interactive editing", as described in Section III(12), through the uniform platform for E&I. If further adjustment is needed, other part of outliers is then treated during the "Automatic editing" process step.
(b) For short-term business statistics and business censuses, due to time constraints and to a large number of units and variables, respectively, "Automatic editing" is performed first. However, after the latter process step, there is a possibility to revisit the most influential outliers if needed. This "Interactive editing" process step may be performed on those units that comprise suspicious aggregates or estimates, in order to ensure better accuracy of statistical information.

27. Finally, "Macro editing" is performed similarly as it was described in Section III(14), that is, E&I impact on statistical information is measured. If E&I procedures need to be refined, it is suggested to return to the "Selective editing" process step.

V. Case study I: Outlier detection

28. We summarize the results of an outlier detections study by Burakauskaitė and Nekrašaitė-Liegė (2022) using the data of the quarterly statistical survey on Service Enterprises of SDA. The suggested improvement of the selective editing method is already implemented in the production process of statistical information at SDA.

29. Here the target variable is enterprise turnover of the accounting period (quarter) t, say, y_t. We have a set of unit-level auxiliary variables, associated with the target variable, – enterprise turnover of the previous period (denoted as y_{t-1}) and of the same period of the previous year (denoted as y_{t-4}), and enterprise turnover from Value Added Tax (VAT) declarations of the accounting period (denoted as y_t^*).

30. As it was briefly mentioned in Section II(4), a few algorithms are used for outlier detection, that is,

(a) Deterministic approach, which includes edit rules based on a comparison with overall trend of other observations belonging to the same subset of the population. The acceptance intervals for each subset are constructed according to the respective interquartile range of observations, that is, an observation of some subset (group) g is considered to be an outlier if it does not belong to the interval $\left[Q_1^{(g)} - 3IQR^{(g)}, Q_3^{(g)} + 3IQR^{(g)}\right]$, where $Q_1^{(g)}$ and $Q_3^{(g)}$ are the first and third quartiles of the vector of target variable observations, respectively, and $IQR^{(g)}$ denotes the interquartile range.
(b) Hidiroglou-Berthelot method, which is based on the idea of acceptance boundary that varies according to the size of a unit. Here ratios from (i) y_{t-1}, (ii) y_{t-4} to y_t are compared to the corresponding overall trend of other observations belonging to the same subset of the population (Belcher, 2003).
(c) Selective editing method, which is based on the idea of selecting a set of outliers that have the biggest impact on the target estimate. For the latter method, enterprise turnover from VAT declarations of the accounting period (y_t^*) is used as the auxiliary variable.

A. Case study I overview and findings

31. During the case study, predictions for the target variable were obtained using the contamination model (Di Zio and Guarnera, 2013). The impact of the potential error on the target estimate was evaluated using the

score function with a standard structure – the difference between the observed value of the target variable and its prediction multiplied by the sampling weight and a suspicion component. A purpose of a score function is to rank the detected outliers based on their influence on the estimate of interest.

32. Burakauskaitė and Nekrašaitė-Liegė (2022) evaluated an impact of the suspicion component on the effectiveness of selective editing by using

(a) A discrete suspicion component (Di Zio and Guarnera, 2013), which is an indicator variable, denoting whether the corresponding observation is considered suspicious (possibly erroneous) by some edit rule, or not.
(b) A continuous suspicion component (Norberg et al., 2010), which depends on the deviation from a chosen acceptance interval, for instance, with a lower bound equal to the first quartile and an upper bound equal to the third quartile of the vector of target variable predictions. The bigger the deviation from such an acceptance interval, the higher the value of the continuous suspicion component.

33. For the calculations, R software environment for statistical computing was used. Selective editing was performed using two main functions from the package *SeleMix* by Guarnera and Buglielli (2013), that is,

(a) *ml.est* for fitting the contamination model, i.e., estimating model parameters, predicting the "true" values of the target variable, and
(b) *sel.edit* for the identification of the most influential outliers, i.e., ranking observations based on the values of the score function.

34. Findings of the study suggest that the inclusion of the continuous suspicion component into the score function expression increases the efficiency of the selective editing method. This can be observed from the comparison between Figures 3 and 4, where Figure 3 depicts the relative absolute bias (RAB) dependency on the number of edited influential outliers using a discrete suspicion component, while Figure 4 portrays the same dependency using a continuous instead of a discrete suspicion component. With a discrete suspicion component, a total of 134 influential outliers are identified, however, the RAB calculation shows that only 92 of them have to be edited in order to achieve the desired level of accuracy (i.e., 0.011). On the contrary, the use of the continuous suspicion component lets to take into consideration distances between observations that do not fall into the chosen acceptance interval and the corresponding bounds of the interval. With this additional impact on the selective editing method, the calculation of RAB shows that almost every identified influential outlier (92 out of 93) has to be edited in order to achieve the desired level of accuracy.

35. Here RAB is calculated as

$$RAB = \frac{\left|\hat{T}_{\tilde{y}} - \hat{T}_y\right|}{\hat{T}_y},$$

where $\hat{T}_{\tilde{y}} = \sum_{i \in S} w_i \tilde{y}_i$ and $\hat{T}_y = \sum_{i \in S} w_i y_i$ are the estimators of the population sum of the target variable ($y_{t,i}$) and the target variable after the treatment of influential outliers ($\tilde{y}_{t,i}$), respectively, with the subscript t omitted in formulas for notation simplicity, and w_i denotes sampling weights for each unit i in the probability sample S. Here the influential outliers are adjusted using the contamination model predictions.

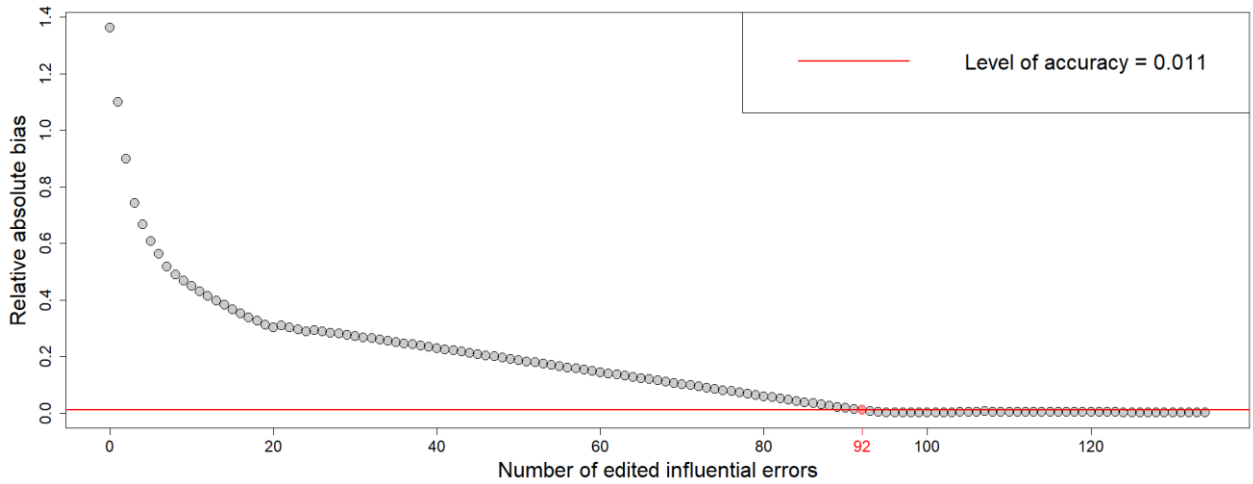

Figure 3: RAB dependency on the number of edited influential errors (outliers) using a discrete suspicion component (Burakauskaitė and Nekrašaitė-Liegė, 2022; Figure 1).

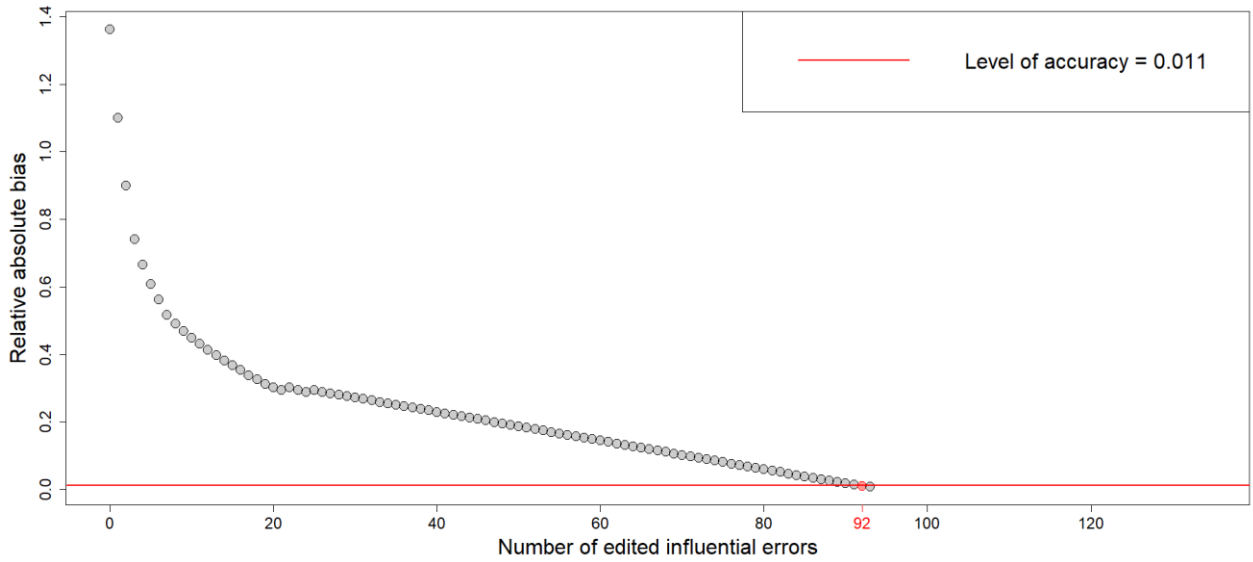

Figure 4: RAB dependency on the number of edited influential errors (outliers) using a continuous suspicion component (Burakauskaitė and Nekrašaitė-Liegė, 2022; Figure 2).

VI. Case study II: Outlier correction

36. We summarize the results of a missing value imputation study by Uogelė (2023) using the data of the monthly statistical survey on Trade and Catering Enterprises of SDA. The suggested imputation methods are planned to be applied for the automatic outlier treatment tasks using the statistical data editing tool in the uniform platform.

37. Here the target variable is enterprise turnover of the accounting period t (month), say, y_t. A set of unit-level auxiliary variables is available, which includes

(a) enterprise turnover of the previous period – y_{t-1},
(b) enterprise turnover from VAT declarations of the previous period – y^*_{t-1},
(c) enterprise turnover from VAT declarations of the accounting period – y^*_t,
(d) categorical variable of four-digit numerical code (classes) of economic activity group, and
(e) categorical variable for 7 enterprise size groups based on the number of employees.

A. Case study II overview and findings

38. The objective of the study was to compare different machine learning algorithms for the missing data imputation of the target variable. To this end, Uogelė (2023) considered different scenarios of response mechanism in the statistical data set, that is, various degree of missingness was generated with missing completely at random (MCAR) and missing at random (MAR) assumptions.

39. A number of imputation methods were used to impute the missing values, including the "traditional" methods usually employed at SDA (such as the *k*-nearest neighbors imputation, which was considered as the reference method, and a simple mean imputation in subsets of the population based on the enterprise size), as well as machine learning algorithms such as

(a) Bayesian linear regression,
(b) Stochastic regression,
(c) Predictive mean matching,
(d) Non-parametric missing value imputation using random forest,
(e) Bootstrapping and expectation-maximization algorithm.

40. For the calculations, four R software packages were used:
   (a) *VIM* (Kowarik and Templ, 2016) for the *k*-nearest neighbors imputation according to the Gower's distance. R function – *kNN*.
   (b) *mice* (van Buuren and Groothuis-Oudshoorn, 2011) for the imputation of multivariate data by chained equations. R function – *mice* with specified methods "norm" (point 39(a)), "norm.nob" (point 39(b)), "pmm" (point 39(c)).
   (c) *missForest* (Stekhoven and Bühlmann, 2012) for the imputation based on the random forest algorithm. R function – *missForest* (point 39(d)). It is also worth noting that an alternative *missRanger* package by Mayer (2019) offers an option of using predictive mean matching, while *missForest* uses mean.
   (d) *Amelia* (Honaker et al., 2011) for the imputation based on a bootstrap expectation-maximization algorithm, producing multiple output data sets. R function – *amelia* (point 39(e)).

41. Table 1 below provides a comparison of the considered missing value imputation methods under MCAR assumption for the response mechanism. Here the results are compared according to two accuracy measures, that is, the normalized root mean squared error (NRMSE) and the mean absolute error (MAE). It is observed that NRMSE values are much higher when the Mean imputation is performed compared to using MissForest and MissRanger. That implies the advantage of the random forest algorithm over the "traditional" approach. The same tendency is observed for MAE values, as well as for different degrees of missingness.

	NRMSE				MAE ($\times 10^3$)			
Missing	5%	10%	20%	30%	5%	10%	20%	30%
Mean	0.83	0.77	0.75	0.70	212.08	160.98	145.31	159.84
MissForest	0.09	0.13	0.12	0.16	19.95	17.42	16.50	24.76
MissRanger	0.15	0.18	0.13	0.16	28.35	21.42	17.09	23.13

Table 1: NRMSE and MAE of monthly enterprise turnover imputation under MCAR response mechanism (Uogelė, 2023; Table 6).

42. Table 2 below provides a comparison of the considered missing value imputation methods under MAR assumption for the response mechanism, according to NRMSE and MAE. In this scenario, we identify four methods with a desirable accuracy based on both NRMSE and MAE, that is, MICE-norm and MICE-norm.nob, as well as MissForest and MissRanger. However, the first two seem to perform slightly better than the latter two for the majority of missingness scenarios. Other methods (MICE-pmm and kNN) performed worse according to high NRMSE and MAE values.

	NRMSE				MAE ($\times 10^3$)			
Missing	5%	10%	20%	30%	5%	10%	20%	30%
MICE-pmm	0.54	0.41	0.51	0.47	88.70	90.10	84.79	97.32
MICE-norm	0.11	0.07	0.24	0.10	25.11	22.98	30.50	24.43
MICE-norm.nob	0.12	0.07	0.14	0.09	24.67	21.28	25.40	24.90
kNN	0.56	0.43	0.43	0.47	120.55	118.75	92.18	112.65
MissForest	0.18	0.14	0.17	0.26	34.89	33.66	34.39	42.85
MissRanger	0.27	0.23	0.21	0.24	42.08	45.75	39.66	43.69

Table 2: NRMSE and MAE of monthly enterprise turnover imputation under MAR response mechanism (Uogelė, 2023; Table 21).

VII. Final remarks

43. The development of a working data editing tool for the standardized automated E&I process at SDA is a complex task with several challenges. While some progress has been made in integrating innovative methods, such as machine learning techniques, into the E&I workflow, the realization of an automated process might still be subject to practical limitations. These include constraints related to IT infrastructure, software solutions, and the knowledge capacity at SDA. The implementation of the process is thus likely to undergo adjustments to accommodate these limitations.

44. Future plans for the improvement of the E&I process might focus on continued research into innovative statistical data editing techniques. In particular, further collaboration with academic community is seen as vital for the exploration and application of advanced methods for outlier detection and correction tasks. SDA will also focus on capacity building, ensuring that employees can effectively use these advanced tools and methods, which will be critical for the long-term success of the modernization effort.

45. While the standardized automated E&I process shows great promise in increasing efficiency and improving the quality of statistical information, there remains a long way toward a fully functional system. The data editing tool in the uniform platform, as envisioned, will require continuous refinement and collaboration among various divisions of SDA. The successful deployment of such a system will depend on overcoming current limitations and addressing emerging challenges during the Production phase.

References

Belcher, R. (June 2003). *Application of the Hidiroglou-Berthelot Method of Outlier Detection for Periodic Business Surveys*. Paper presented at the SSC Annual Meeting, Halifax, Nova Scotia, Canada.

Burakauskaitė, I. and Nekrašaitė-Liegė, V. (2022). Selective Editing Using Contamination Model. *Romanian Statistical Review*, 1:55–65.

Di Zio, M. and Guarnera, U. (2013). A Contamination Model for Selective Editing. *Journal of Official Statistics*, 29(4):539–555.

Guarnera, U. and Buglielli, M. T. (2013). SeleMix: An R Package for Selective Editing.

Honaker, J., King, G. and Blackwell, M. (2011). Amelia II: A Program for Missing Data. *Journal of Statistical Software*, 45:1–47.

Kowarik, A. and Templ, M. (2016). Imputation with the R Package VIM. *Journal of Statistical Software*, 74:1–16.

Mayer, M. (June 2019). missRanger: An R-package for Fast Imputation of Missing Values, Version 2.6.0, updated August 2024.

Norberg, A., Adolfsson, C., Arvidson, G., Gidlund, P. and Nordberg, L. (2010). *A General Methodology for Selective Data Editing*. Stockholm: Statistics Sweden.

Pannekoek, J. and Zhang, L.-C. (September 2012). *On the General Flow of Editing*. Paper presented at the UNECE Work Session on Statistical Data Editing, working paper 26, Oslo, Norway.

Stekhoven, D. J. and Bühlmann, P. (2012). MissForest – Non-parametric Missing Value Imputation for Mixed-type Data. *Bioinformatics*, 28(1):112–118.

UNECE (January 2019). Generic Statistical Business Process Model (GSBPM), Version 5.1, updated July 2024.

UNECE (June 2019). Generic Statistical Data Editing Model (GSDEM), Version 2.0, updated July 2024.

Uogelė, J. (2023). *Missing Data Imputation Methods for Monthly Statistical Survey on Trade and Catering Enterprises* [Unpublished master's thesis]. Vilnius University.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45:1–67.

Appendix A

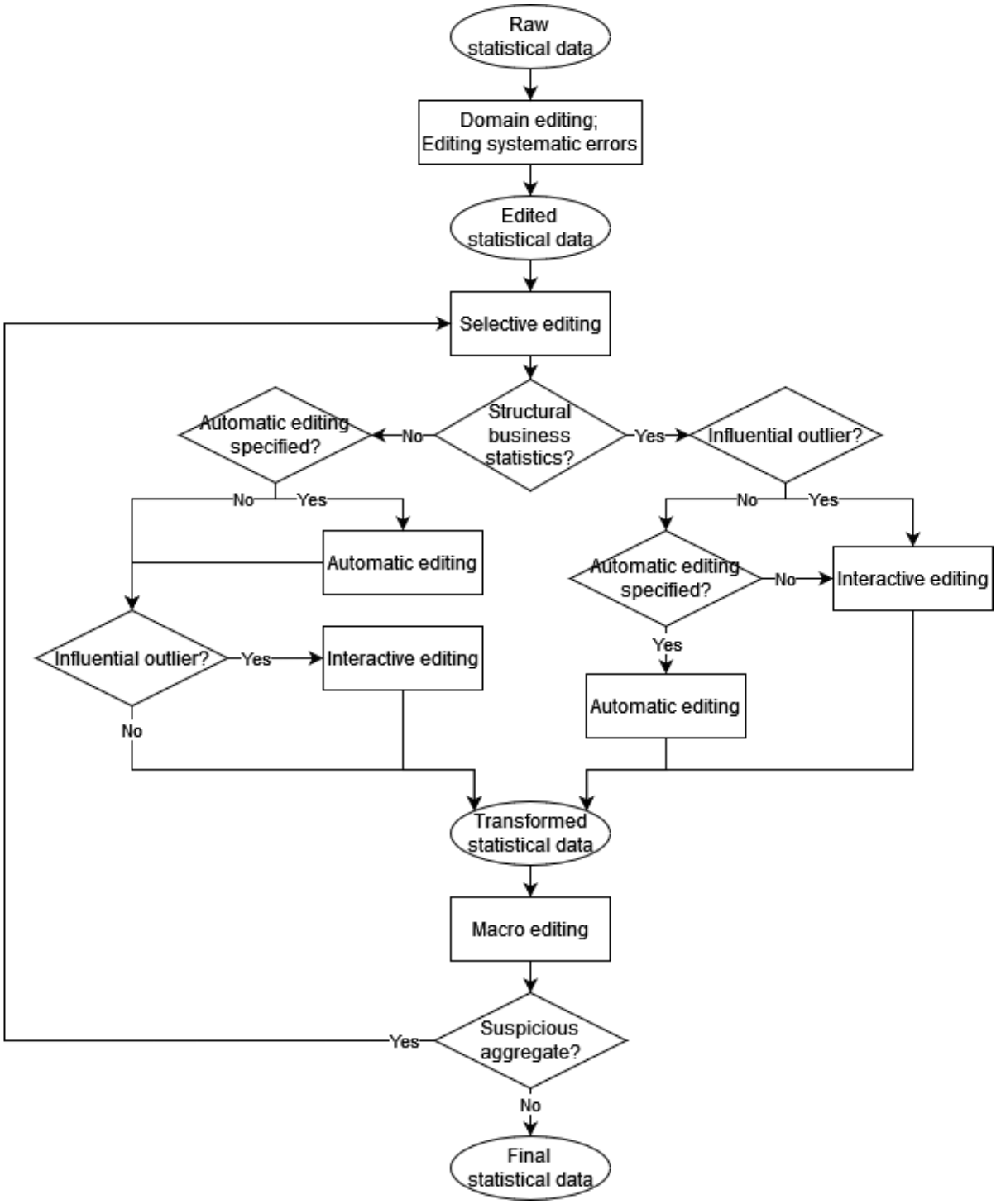

Figure A1: The envisioned statistical data E&I flow model for business statistics surveys.