UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert Meeting on Statistical Data Editing

3-7 October 2022, (virtual)

Automatic selective editing approach using machine learning: an application to VAT data

Benjamin Vasquez, Oscar Quintana, Javier Larrain (Central Bank of Chile, Chile)

bvasquez@bcentral.cl, oquintana@bcentral.cl, jlarrainq@bcentral.cl

I. Introduction

1. Data editing is one of the most time-consuming tasks of the National Statistics Institutes (NSIs), which face increasing demand for statistical products along with budget constraints (Menghinello, Faramondi, & Laureti, 2020; UNECE's Task Force on Value of Official Statistics, 2018). Administrative registers, as well as any data sources, are prone to errors. Facing an increasing volume of data available and more accessible computing power, the automation of the data editing process becomes unavoidable.

2. Recent advances in data editing have incorporated the use of machine learning (ML) in the detection of outliers, with promising results but unclear advantages regarding traditional (non-ML) methods. Compared to traditional methods, ML methods require more computational power, which may frustrate some implementation efforts.

3. In this working paper, we compare conventional and machine learning methods for outlier detection in a dataset of monthly sales from value added tax (VAT) declarations of Chilean firms. To establish a fair assessment process, we created a labeled dataset and evaluated its performance on standard classification metrics and computational cost. Our preliminary findings suggest that machine learning models outperform traditional approaches in the task of outlier detection as shown by performance metrics.

4. To take advantage of these novel methods we propose an automatic selective editing approach: focusing the use of ML methods on the most influential entities.

II. Comparison of outlier detection methods

A. Dataset

5. Value added tax (VAT) form is one of the most important source of information for sales of the firms in Chile, feeding high frequency indicators for national accounts. The low size of the informal economy in Chile (Medina & Schneider, 2018) makes this a good source of information. VAT declarations provides monthly data for more than a million of tax contributors.

6. For this exercise we grouped the firms by industry[1] and firm size[2], considering from each group the top 25% of the firms with most presence in the period between January 2013 to August 2021. From this subset, we performed a random selection of 1,200 firms, obtaining a sample of 109,044 records.

B. Labeling the dataset

7. A labeled dataset provide a benchmark with ground truth annotation, necessary to perform an evaluation of outlier detection methods.

8. The sampled dataset was revised and labeled by industry experts, who tagged outliers in the sales data in a firm-by-firm time-series approach. This exercise identified 717 outliers, corresponding to a 0.7% of the sample.

C. Outlier detection methods

9. In this work, we selected different methods and compared their performance using standard metrics. The selected methods are described as follows:

(a) Interquartile distances method (IQ)

Interquartile distances method (IQ) defines boundaries for monthly and annual ratios using the percentiles 25th, 50th (median) and 75th for both ratios. The monthly and annual ratios of the sales data for a firm i at the month t are defined as $r_{i,t}^{m} = x_{i,t}/x_{i,t-1}$ and $r_{i,t}^{a} = x_{i,t}/x_{i,t-12}$, respectively, where $x_{i,t}$ is the sales of the firm i at the period t. The percentiles were estimated over the 2 years of previous information for each firm, whereas the lower and upper bounds for the monthly and annual ratios of each firm are defined as follows:

$$\begin{aligned}\text{Lower limit } &= r_{p50} - k * (r_{p50} - r_{p25}) \\ \text{Upper limit } &= r_{p50} + k * (r_{p75} - r_{p50})\end{aligned}$$

Where k value defines the range of the limits. The authors suggest a value of 3 for this parameter. These data values $x_{i,t}$ whose ratios $r_{i,t}^{m}$ and $r_{i,t}^{a}$ lie outside the bounds are considered outliers.

(b) HB method

Hidiroglou and Berthelot (1986) presented a method “proven successful by experience” (Norberg, 2016) which considers the ratio and the relative size of a variable and define robust boundaries transforming the data and calculating median and quartiles.

From the same definitions used for describing the monthly and annual ratios in the IQ method, the first transformation is as follows:

$$s_i = \begin{cases} 1 - r_{p50}/r_i \text{ , if } 0 < r_i < r_{p50} \\ r_i/r_{p50} - 1 \text{ , if } r_{p50} \leq r_i \end{cases}$$

The second transformation considers the size of the data and control its importance with the parameter $U \in [0,1]$:

$$E_i = s_i * \left(Max\left(x_{i,t-1}, x_{i,t}\right)\right)^{U}$$

[1] We considered the following industries: Wholesale and retail trade, Information services, Construction, Electricity, gas and water supply, Education and health services, Manufacturing, Administrative and professional services, Transport and storage and Accommodation and food services.

[2] We set up two groups: large and non-large enterprises. We used the definition established in the Chilean law, which considers a company as “large” if their annual sales is over 100,000 UF, where UF refers to a daily indexation indicator that is CPI-adjusted. For reference, in 2021 100,000 UF equals 3.93 MM USD and 3.32 MM EUR.

Finally, the method considers as outliers all the values E_i outside the interval defined as follows:

$$\begin{aligned}\text{Lower limit } &= E_{p50} - Cd_{Q1}\\ \text{Upper limit } &= E_{p50} + Cd_{Q3}\end{aligned}$$

Where $d_{Q1} = Max(E_{p50} - E_{p25}, |AE_{p50}|)$ and $d_{Q3} = Max(E_{p75} - E_{p50}, |AE_{p50}|)$. The authors suggest a value of 0.05 for A, while for the U and C values we use 0.5 and 2 respectively.

(c) IQ & HB (HB anchored) method

We also used a combination of the two outliers' detection methods described previously. It is known that the HB method alone may flag too many small values as outliers (Statistics New Zealand, 2009). To prevent the over-detection of outliers we used the IQ method as an "anchor" for the HB method. This hybrid method identifies an observation as outlier if it is tagged as such from both methods, reducing the chance of false positives and preserves further the original data. This method is currently uses at the Central Bank of Chile for editing the sales (VAT) data.

(d) DBSCAN

Lastly, we considered Density-based spatial clustering of applications with noise (DBSCAN) as an outlier detection method (Ester, Kriegel, Sander, & Xu, 1996), a very well-known non-supervised machine learning algorithm. This clustering algorithm is based on the density of the data grouping together the points that are closely packed together, i.e., points with many nearby neighbors, tagging as outliers those points that lie alone in low-density regions. We used this algorithm in three versions: with sales (one variable), with sales and costs of sales, and with sales and the ratio sales/costs.

D. Assessment metrics

10. Due to the nature of this exercise -a binary classification problem- a confusion matrix and the metrics derived from it are useful for evaluating these methods and algorithms. A confusion matrix intersects the actual classes (rows) with the predicted classes (columns), as follows:

Table 1. Confusion matrix

		Predicted condition	
	Total = P + N	Positive	Negative
Actual condition	Positive	True positive (TP)	False negative (FN)
	Negative	False positive (FP)	True negative (TN)

11. Three of the metrics derived from the confusion matrix are precision, true positive rate (TPR) and false positive rate (FPR). Precision or positive predictive value is the proportion of positive results that are true positive. TPR, recall or sensitivity is the proportion of actual positives that are predicted as such. Finally, FPR is the proportion of actual negatives that are misclassified. These metrics allowed to build up more complex ones.

$$precision = \frac{TP}{TP + FP}$$

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

12. We also used other very well-known metrics that derives from the confusion matrix for evaluating the performance of the outlier detection techniques.

(a) Area under the curve of the receiver operating characteristic curve (AUC ROC)

The receiver operating characteristic curve (ROC) plots the FPR (x-axis) against the TPR (y-axis) for different probability threshold. A perfect classifier will yield the coordinate (0,1), representing no false negatives and no false positives, and the area under the curve (AUC) will be 1. A random classifier will result in a curve close to the diagonal and an AUC ROC value of 0.5.

(b) Area under the curve of the precision-recall curve (AUC PRC)

AUC ROC is a popular metric for the evaluation of a binary classifier, however for imbalanced datasets a more appropriate measure is the AUC of the precision-recall curve (PRC) (Saito & Rehmsmeier, 2015). The PRC plots the precision and TPR (recall) for different thresholds. Higher AUC PRC implies more accurate results.

E. Results

13. The performance metrics are shown in the Figure 1 and in the Table 2. In the Table 2 we also included a metric called RATIO_N that measures the number of predicted outliers vs the actual number of outliers. For example, if we have a RATIO_N of 2 it means that the model is predicting twice the amount of the actual outliers.

Figure 1. Comparison of the performance metrics

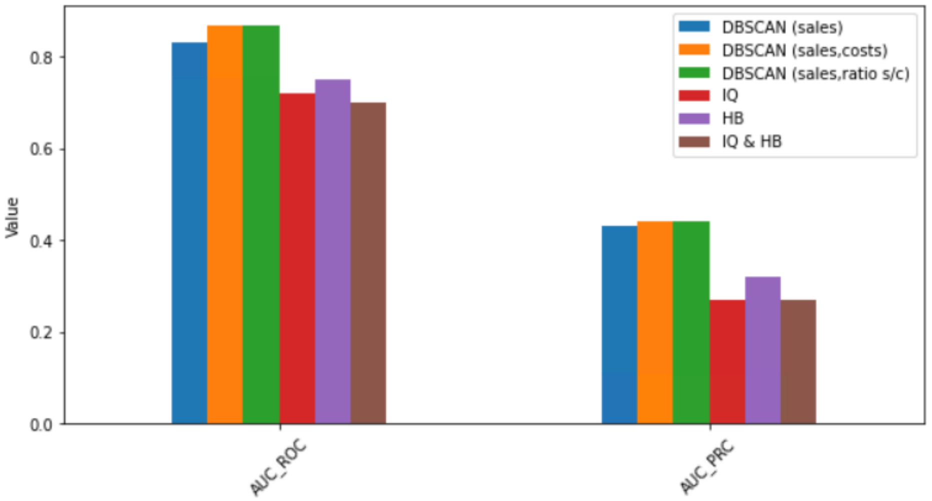


Table 2. Detailed comparison of the performance metrics

Method	AUC_ROC	AUC_PRC	TPR	FPR	PRECISION	RATIO_N
DBSCAN (sales)	0.83	0.43	0.68	0.02	0.19	3.69
DBSCAN (sales, costs)	0.87	0.44	0.84	0.09	0.06	14.64
DBSCAN (sales, ratio s/c)	0.87	0.44	0.86	0.12	0.05	18.62
IQ	0.72	0.27	0.47	0.04	0.07	6.38
HB	0.75	0.32	0.63	0.13	0.03	20.02
IQ & HB	0.70	0.27	0.42	0.02	0.12	3.38

14. DBSCAN maximizes the AUC ratios over conventional methods (IQ, HB, IQ & HB), meaning better performance. However, DBSCAN with two variables detects a high number of outliers and have therefore a higher FPR, a higher RATIO_N and a lower Precision compared to IQ & HB.

15. In summary, DBSCAN with one variable has the advantage of a ML predictor in accuracy and the efficiency of the IQ & HB method in terms of not flagging too many outliers.

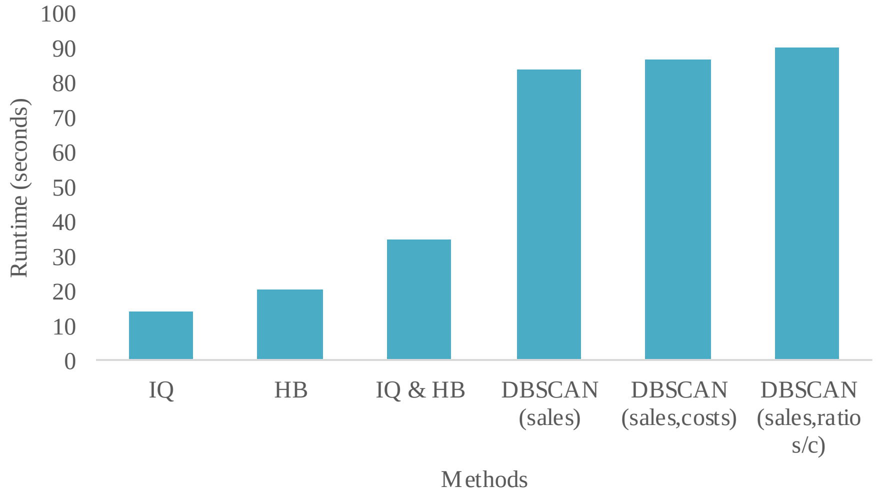
III. Automatic selective editing

A. The computing problem of machine learning (ML) methods

16. ML methods by nature require more computing power[3], implying more computing time. This can be troublesome when working with large datasets, making some of these novel methods very difficult to implement.

17. The methods presented in this working paper were executed on a server with 2 Intel Xeon Gold 6142 processors running at 2.6 GHz each and 1 TB of RAM. The resulting process time (CPU time + system time) for each method is displayed in the following graph:

Figure 2. Process time by method



18. The standard methods in average were run in 22.9 seconds of process time, while DBSCAN used around 86.6 of process time. This implied that the ML method used 3.8 times more process time than the average of the non-ML methods, illustrating the problem mentioned above.

Table 3. Detailed process time and wall-clock time by method

	Time in seconds	
Method	**Process Time**	**Wall-Clock Time**
DBSCAN (sales)	83.4	86.6
DBSCAN (sales,costs)	86.6	89.0
DBSCAN (sales,ratio s/c)	89.8	92.4
IQ	13.8	17.5
HB	20.4	23.3
IQ & HB	34.6	37.4
Mixed	45.9	48.1

[3] DBSCAN has a run time complexity of $O(n * \log(n))$, where n are the points of the database, making it a reasonable reference for other ML methods that may have higher complexity.

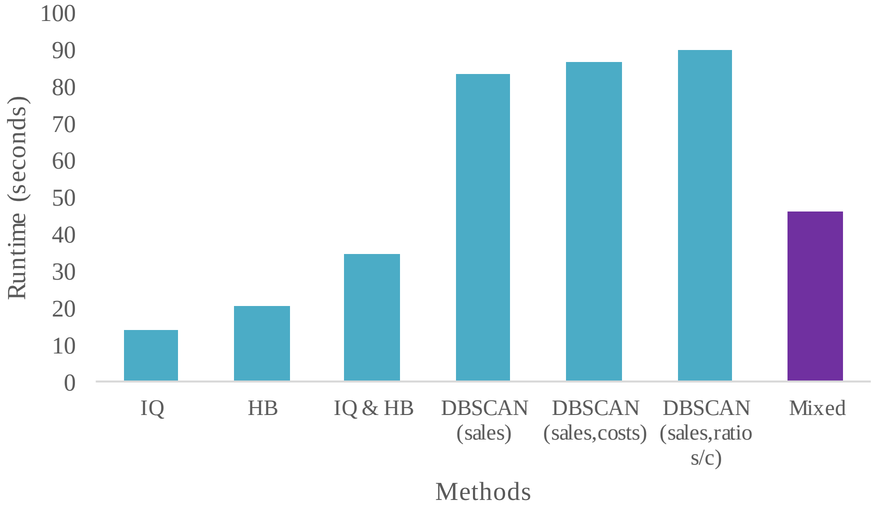
B. Automatic selective editing approach

19. To overcome this problem, we propose an approach based on the selective editing paradigm (Wall, 2013), restricting the implementation of the ML methods to the most influential entities, while the bigger set of less influential records are examined with less computational-costly standard methods. This "automatic selective editing" approach can eventually be extended to the amendment (imputation) phase.

C. Performance gains

20. Restricting DBSCAN to the top 56.3% most influential[4] records, we can reduce the time almost by half (47.3% specifically), as we can see in the Figure 3.

Figure 3. Process time by method



IV. Conclusions and future work

21. In this work, we presented a standard evaluation scheme for outlier detection methods and its implementation. We compared some methods, using a labeled dataset and standard metrics, finding better performance in DBSCAN –a non-supervised ML method- against conventional methods.

22. Nonetheless, in a broad sense ML methods require more computational power and process time compared to conventional methods, which are simpler to compute. This problem grows along with the size of the dataset. To get over this situation, we propose focusing its implementation on the most influential entities, while maintaining the conventional methods for the rest of the records.

24. This evaluation can be improved by adding and assessing more outlier detection methods. Additionally, we can take advantage of the labeled dataset and train supervised ML models that can perform outlier detection tasks. This can be addressed in a future work.

Acknowledgments

We would like to thank the contributions made by Joaquin Perez-Lapillo, who pave the way for this document. We also thank the National Accounts Department and the Experimental Statistics Group for their dedicated effort in labeling the dataset. Finally, we are grateful to Francisco Ruiz, Josué Pérez, Sebastián Rébora, Erika Barrera and Ricardo Flores for their reading of our manuscript and their insightful comments and suggestions.

[4] For simplicity, we considered as influential the group of large enterprises (see note 2 above).

References

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*.

Hidiroglou, M., & Berthelot, J. (1986). Statistical editing and imputation for periodic business surveys. *Survey Methodology*, 73-83.

Medina, L., & Schneider, F. (2018). Shadows economies around the world: what did we learn over the last 20 years? *IMF Working Paper*.

Menghinello, S., Faramondi, A., & Laureti, T. (2020). The future role of official statistics in the business data arena. *Statistical Journal of the IAOS*, 519-533.

Norberg, A. (2016). SELEKT - a generic tool for selective editing. *Journal of Official Statistics*, 209-229.

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall plot Is more Informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*.

Statistics New Zealand. (2009). Automated editing and imputation system for administrative financial data in New Zealand. *Work session on statistical data editing, UNECE.*

UNECE's Task Force on Value of Official Statistics. (2018). *Recommendations for promoting, measuring and communicating the value of official statistics.* New York and Geneva: United Nations.

Wall, T. d. (2013). Selective editing: a quest for efficiency and data quality. *Journal of Official Statistics*, 473-488.