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

Benjamin Vásquez
Central Bank of Chile

UNECE Expert Meeting on Statistical Data Editing. 3 – 7 October 2022.

I. Context and Motivation

Motivation

- Recent advances in data editing have incorporated the use of machine learning (ML) algorithms in the detection of outliers
- Results are promising, although in some cases it is unclear how much better is the performance of ML compared to conventional methods

Questions

- How can we evaluate and compare outlier detection algorithms in a standardized way?
- How can we take advantage of machine learning (ML) methods in a computational adverse scenario (i.e., complex ML methods or large datasets)?

Work objective

- Compare outlier detection methods with **standard metrics**
  - Generate a labeled dataset
  - Evaluate and compare **conventional and machine learning methods** in their ability to detect outliers using standard metrics
- Propose an approach to take advantage of the most **computationally complex** methods

II. Methodology

Dataset (1)

Value added tax (VAT) data

- The data comes from the Chilean Tax Agency (SII) and consists of **monthly sales** of Chilean firms
- ~ 700,000 firms per month from year 2013 and August 2021

	Jan. 2013	Feb. 2013	…	Aug. 2021
Firm 1	sales	sales	…	sales
Firm 2	sales	sales	…	sales
…	…	…	…	…
Firm N	sales	sales	…	sales

Dataset (2)

Sampled dataset

- We grouped the firms by industry and size. For each group we considered the top 25% firms with the highest sales
- From this dataset we randomly selected 1,200 firms => 109,044 records (firm x time period)
- Sampled subset labeled by industry experts

	date	sales	…	outlier
Firm 1	‘2013-01-01’	sales	…	0
Firm 1	‘2013-02-01’	sales	…	0
…	…	…	…	…
Firm 1200	‘2022-08-01’	sales	…	0

717 outliers in total
0.7% of the sample

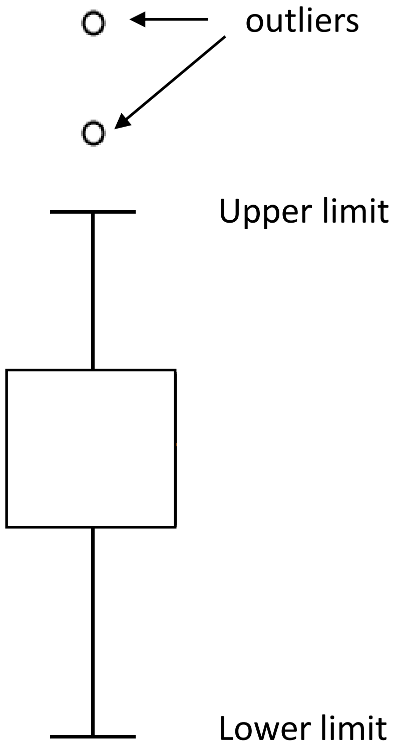
Outlier detection methods (1)

Interquartile distances (IQ)

- We use the sales variable to form **two ratios** to evaluate in this method: the annual and monthly ratios
- Then we apply **Interquartile distances** to define the upper and lower limits

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

- If both ratios of the firm f are outside de bounds, then it is tagged as outlier

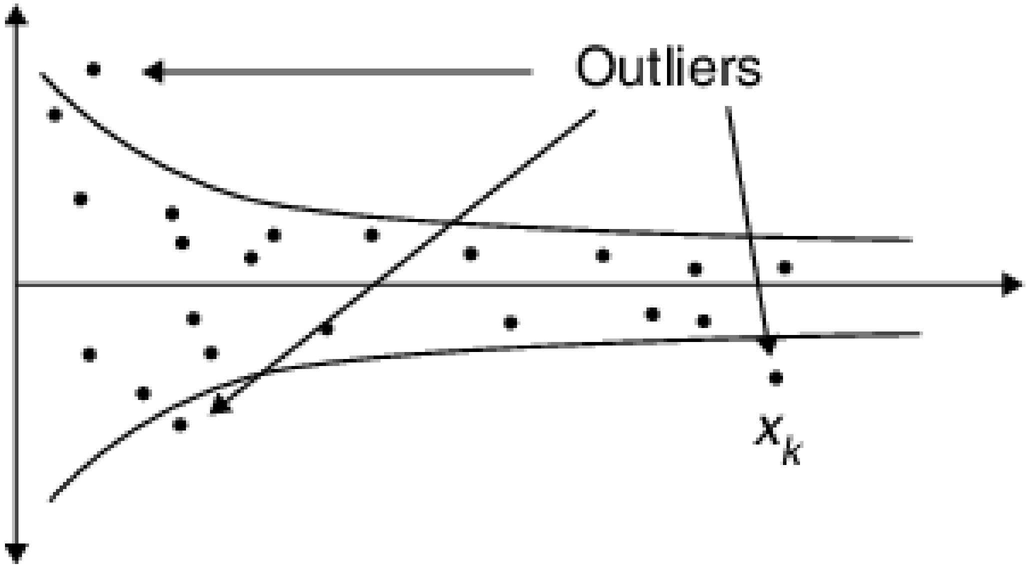


Outlier detection methods (2)

Hidiroglou and Berthelot (HB)

Hidiroglou and Berthelot (1986) considers the **ratio and the relative size** of a variable and define robust boundaries transforming the data and calculating median and quartiles.

This method deals with distributions not normally distributed and overcomes the masking effect and the problem of large variability of small business.



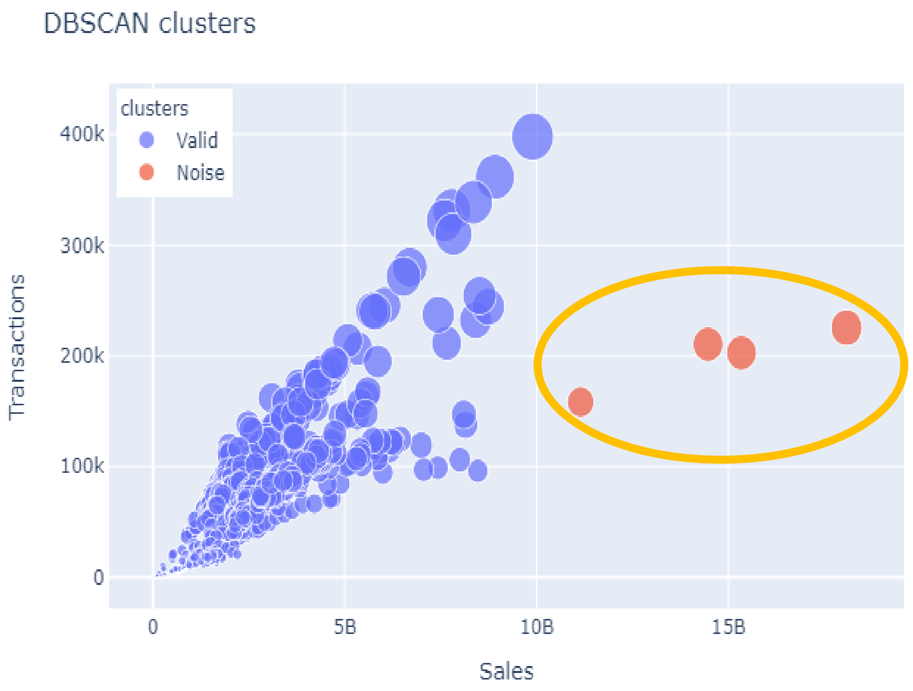
Lower limit $= E_{p50} - Cd_{Q1}$

Upper limit $= E_{p50} + Cd_{Q3}$

Outlier detection methods (3)

DBSCAN

- **DBSCAN** is an unsupervised clustering algorithm specially designed to **identify noise**
- The algorithm identifies **dense regions** based on Euclidean distances. Any observation outside the valid clusters will be **marked as an outlier**
- We use the two dimensions available in the dataset, to form a distribution for **each firm**



DBSCAN(date, sales)
DBSCAN(date, sales and costs)
DBSCAN(date, sales and sales/costs)

Assesments metrics

- Due to the nature of this exercise -a binary classification problem- a confusion matrix and the metrics derived from it are useful

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)

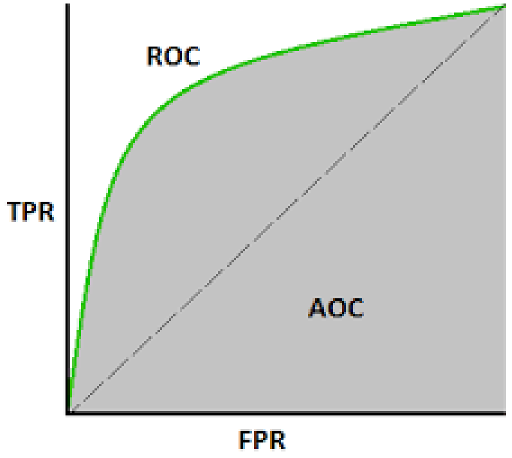
Assesments metrics

- **Precision**: Ability of the model to predict outliers
- True Positive Rate (**TPR**): Proportion of outliers correctly predicted
- False Positive Rate (**FPR**): Proportion of not outliers incorrectly predicted
- Area Under the Curve of the Receiver Operating Characteristic curve (**AUC ROC**): FPR against TPR. It tells us how much the model is able to distinguish the classes
- Area Under the Curve of the Precision-Recall Curve (**AUC PRC**): More appropriate for imbalanced datasets

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

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

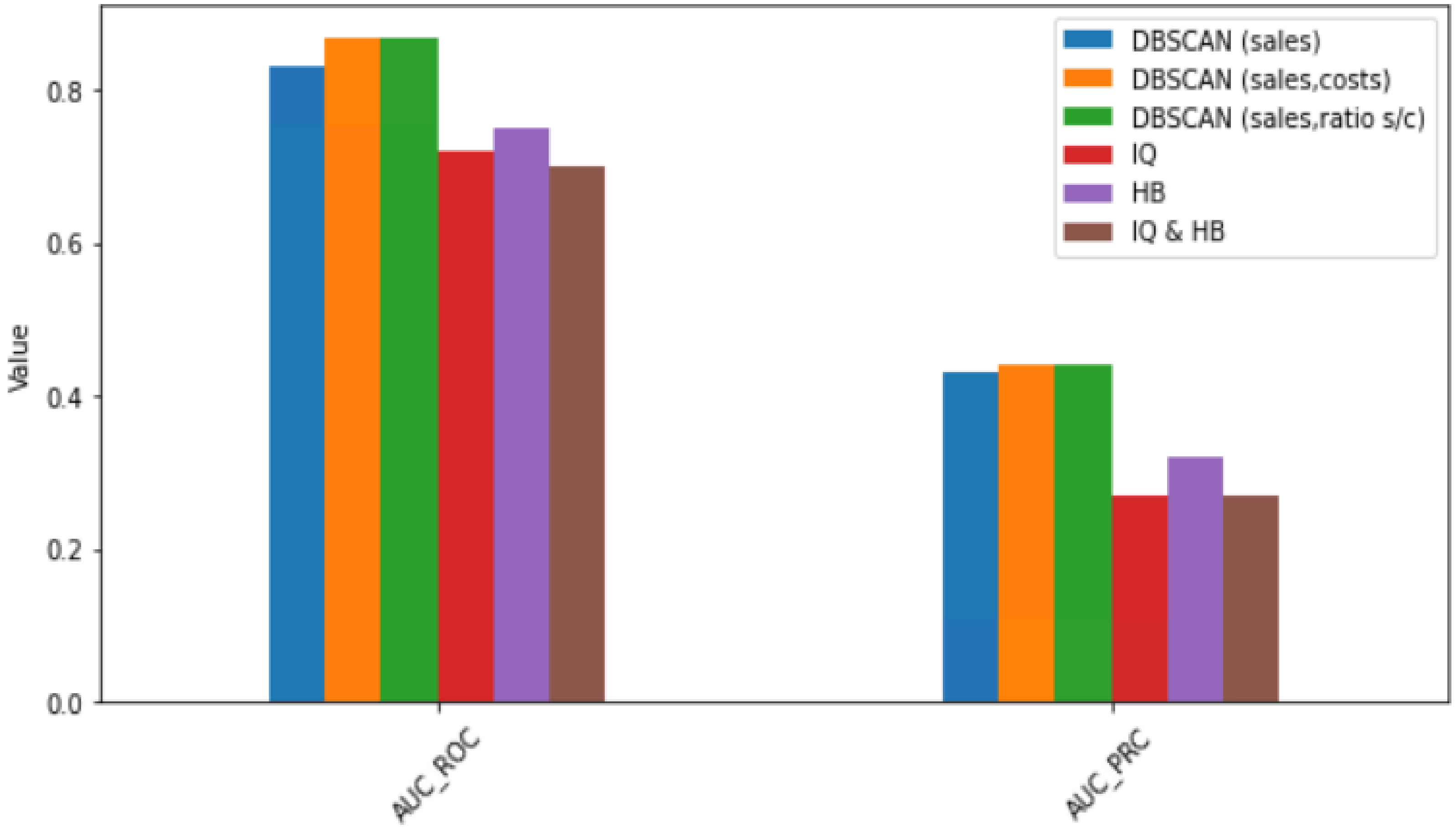
$$FPR = \frac{FP}{FP + TN}$$



III. Results

DBSCAN models have better results

AUC ROC and AUC PRC

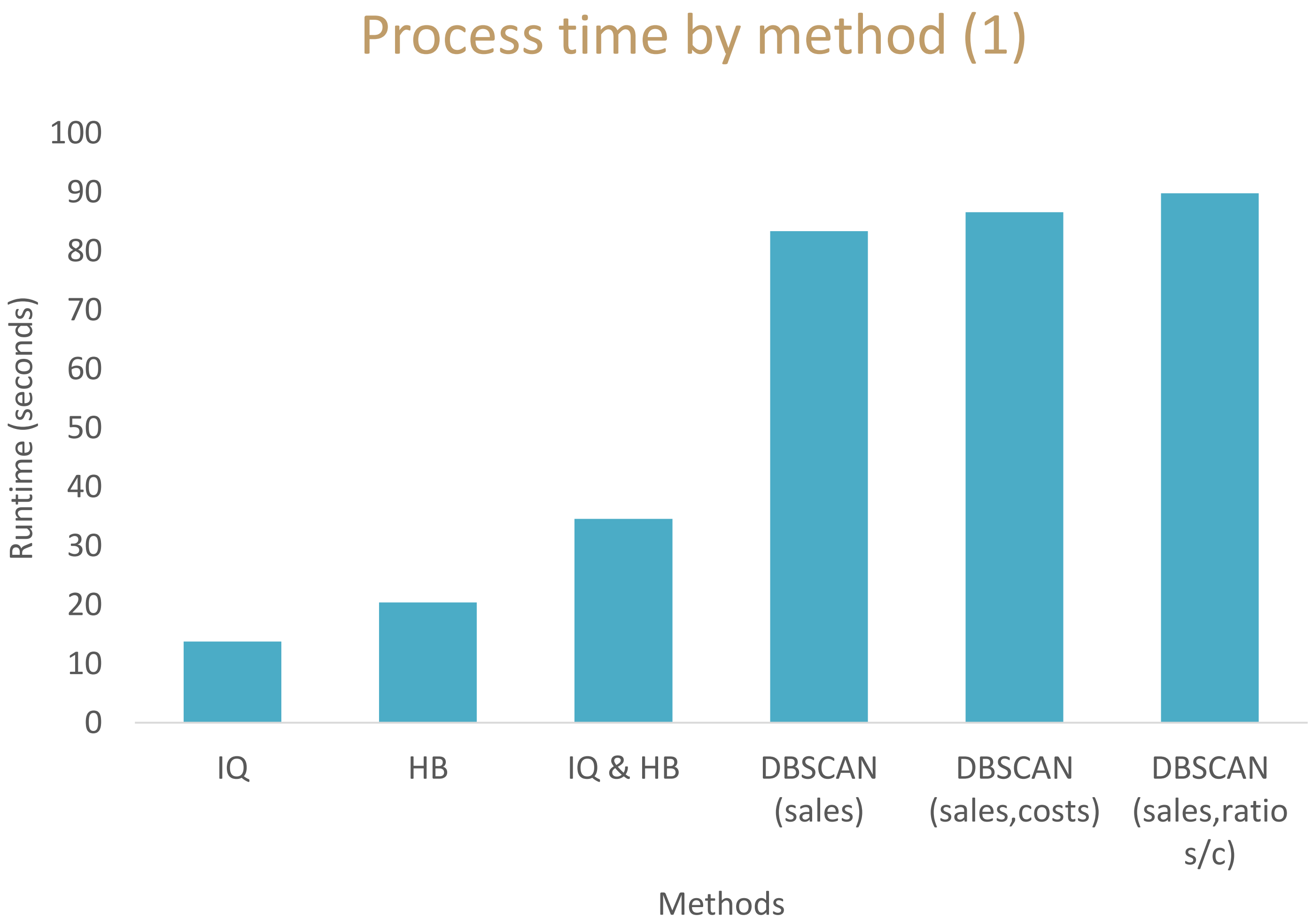


DBSCAN(date, sales) has overall better results.
It tags less records as outliers => it has less False Positives

Detailed results

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

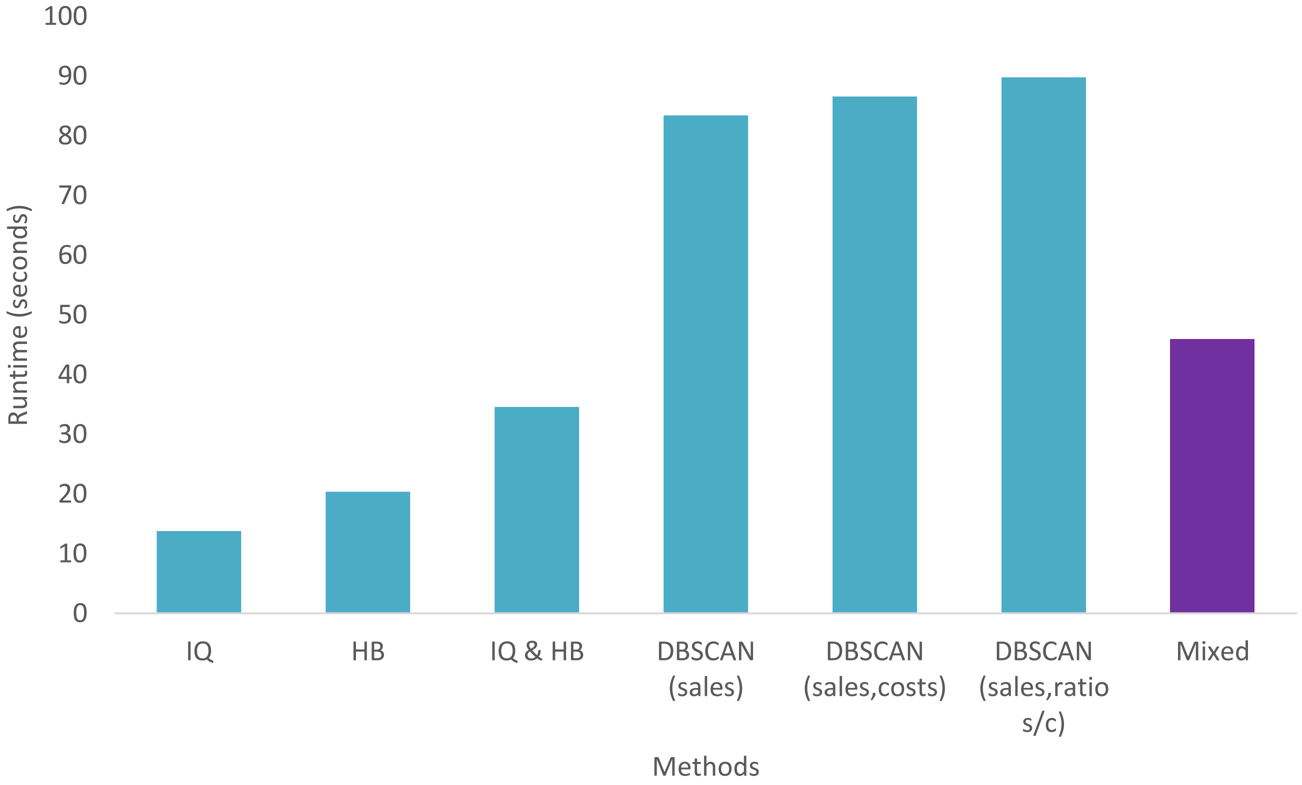
However, DBSCAN models require more computing power



banco central Chile

We propose a automatic selective editing approach: focusing the use of ML methods on the most influential entities

Process time by method (2)



Mixed method applies DBSCAN to the 56% of the firms with more sales and IQ&HB to the rest of the firms

Conclusions and future steps

Conclusions

- We compared some methods for outlier detection, using a labeled dataset and standard metrics, finding **better performance in DBSCAN** –a non-supervised ML method- against conventional methods
- We propose an automatic selective editing approach, focusing its implementation on the most influential entities (highest sales) to overcome the computational problem

Future steps

- Apply the latter approach to all the datset
- Evaluate more models and add more precise dimensions (variables) to the DBSCAN. Number of workers, financial statements, among others

banco central Chile

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

Benjamin Vásquez
Central Bank of Chile

UNECE Expert Meeting on Statistical Data Editing. 3 – 7 October 2022.