ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/368536531

Who Are You? Cartel Detection Using Unlabeled Data

Article *in* International Journal of Industrial Organization · January 2023

DOI: 10.1016/j.ijindorg.2023.102931

CITATION
1

READS
172

4 authors, including:

Douglas Silveira
University of Alberta
23 PUBLICATIONS **47** CITATIONS
SEE PROFILE

Eduardo Fiuza
Instituto de Pesquisa Econômica Aplicada - IPEA
27 PUBLICATIONS **130** CITATIONS
SEE PROFILE

Daniel O. Cajueiro
University of Brasília
134 PUBLICATIONS **4,473** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Do Antidumping Measures Increase Market Power? Evidence From Latin American Countries View project

All content following this page was uploaded by Douglas Silveira on 15 February 2023.

The user has requested enhancement of the downloaded file.

Who Are You?
Cartel Detection Using Unlabeled Data

Douglas Silveira

Department of Economics, University of Alberta, Canada

Territorial and Sectoral Analysis Laboratory – LATES, Brazil.

Lucas B. de Moraes

Institute for Applied Economic Research (Ipea), Brazil

Eduardo P. S. Fiuza

Institute for Applied Economic Research (Ipea), Brazil

Daniel O. Cajueiro*

Department of Economics, University of Brasilia, Brazil

National Institute of Science and Technology for Complex Systems (INCT-SC), Brazil

Machine Learning Laboratory in Finance and Organizations (LAMFO), Brazil

January 26, 2023

Abstract

We propose a data-driven machine learning approach to flag bid-rigging cartels in the Brazilian road maintenance sector. First, we apply a clustering algorithm to group the tenders based on their attributes. Second, we use the labels created by the clustering algorithm as a target variable to predict them using a classifier. We rank the screens according to their relevance to decrease the number of false positive (detecting cartel when it does not exist) and false negative (not detecting cartel when it does exist) predictions. Our results shed light on the need to use a range of screens to recognize the vast profile of strategies practiced by bid-rigging cartels, such as misleading competitive dynamics, bid combination, and cover bidding behavior. Our method can improve cartels' deterrence in different economic sectors, especially when labeled data are not available. In a controlled environment with a simulated labeled dataset, the overall average accuracy of the algorithm is 99.33%. In a real-world cartel case with a labeled dataset, the overall average accuracy is 80.25%. When applied to the road maintenance unlabeled dataset, our model identified a group containing 273 (31% of the total) suspicious tenders. We conclude by offering a policy prescription discussion for antitrust authorities.

Keywords: Cartel screens, bid-rigging cartels, unsupervised learning, clustering

*I am the corresponding author (danielcajueiro@gmail.com). Department of Economics, University of Brasilia, Campus Universitario Darcy Ribeiro - FACE, Brasilia, DF 70910-900.

1 Introduction

One of the fundamental issues in Industrial Organization and competition policy is detection, punishment, and deterrence of collusive behavior. Although it is a consensus that collusive agreements reduce social welfare, firms have incentives to coordinate their decisions and increase their returns (Levenstein and Suslow 2006). The increasing amount of electronic databases has been crucial for developing methods that integrate economic analysis and data-driven approaches to combat such behavior effectively.

In this paper, we propose a Machine Learning (ML) approach to identify bidding patterns consistent with bid-rigging cartels related to road maintenance in Brazil. The analysis is based on an unlabeled dataset from open tenders between 2012 to 2020, covering 891 tenders from all states and the Capital District of Brazil. We combine ingredients from unsupervised and supervised ML with statistical and economic analysis. First, we apply a clustering algorithm (i.e., unsupervised ML) to group the tenders according to their attributes.[1] Second, we use the labels created by the clustering algorithm as a target variable and try to predict these labels using a classifier (i.e., supervised ML). Third, we find out the most relevant input variables to characterize each cluster and the typical values of these variables. Fourth, based on the attributes that shape these clusters of tenders, we identify the ones that present behavior consistent with the existence of bid-rigging cartels. Thereby, our last step is strongly dependent on previous evidence about the detection of cartels using statistical screens.

Certain supervised ML methods have been proposed to detect cartels using statistical screens (Huber and Imhof 2019; Imhof and Wallimann 2021). They depend on two basic steps: (1) labeled data collection; (2) model parameters estimations. However, labeled data are frequently unavailable. In general, they are costly and depend on previous in-

1. In ML terminology, the following are synonyms: attributes, features, and explanatory/input variables. In our context, our attributes are summary statistics we use as statistical screens for detecting potential collusive patterns in the data. More details on the statistical screens are given in Section 4.1.

vestigations of the antitrust authority. In addition, even after the cartel investigation, the pieces of evidence collected may not be conclusive. In the case of the tenders related to road maintenance activities in Brazil, the available data is insufficient to train supervised machine learning models. Our ML approach intends to identify collusive patterns without labeled data.

We use the Gaussian Mixture Clustering Model (GMCM) for the unsupervised ML stage and the Quadratic Discriminant Analysis (QDA) for the supervised ML stage. The choice of the GMCM explicitly assumes that the data-generating process of the open tenders attributes comes from a Gaussian mixture. Thus, due to the finiteness property of the first and second moments of each mode of the Gaussian mixture, the GMCM tends to group tenders with similar attributes. QDA also assumes that each class follows a Gaussian distribution and uses the expected value of each input variable as a classification criterion. The Permutation Importance (PI) technique is used in supervised ML to assess the relationship between the input variables and the target variable (the labels generated via the clustering analysis in our unsupervised ML stage). This step in our approach finds out the most "skilled variables" to separate the clusters into groups of tenders with higher and lower probabilities of anti-competitive behavior. Furthermore, we generate (additional) results to support our findings using different combinations of the GMCM and K-means (unsupervised ML stage) and QDA and logistic regression (supervised ML stage).[2]

We first validate our approach using Monte Carlo simulations to generate labeled data in a controlled environment. Then, we provide a similar analysis using the gasoline cartel labeled dataset due to Silveira et al. (2022) and compare our predictions with their actual labels. Since the Monte Carlo simulations exercise presents a data-generating process close to the one used in our model, the model accuracy is around 100%. The overall average accuracy to detecting the correct label in the gasoline cartel dataset is 80.25%.

Then we investigate an unlabeled dataset of open tenders related to road maintenance

2. The additional estimations and results are available in Appendixes A and B, respectively.

activities in Brazil. We use six statistical screens in our model. These are grouped into: (a) "elementary" screens (total number of firms, the total number of bids, the average number of bids, and the number of firms offering single bids) to detect possible misleading competitive dynamics; (b) "variance-based" screen (coefficient of variation) to capture bid combination strategies; (c) "cover bidding" screen (skewness) to identify situations where the cartel members submit bids in excess of the cartel member designated to win the contract. Explanatory variables similar to (a), (b), and (c) have been used in works dedicated to studying bid-rigging strategies, either using labeled (Porter and Zona 1999; Tóth et al. 2014; Huber and Imhof 2019) or unlabeled data (Bajari and Ye 2003; Chassang et al. 2019; Kawai and Nakabayashi 2022). Our findings suggest the need to use a range of statistical screens to recognize the vast profile of strategies practiced by bid-rigging cartels, such as misleading competitive dynamics, bid combination, and cover bidding behavior. Guided by the patterns found by these statistical screens, our approach indicates a cluster of bidding data indicative/suspicious of anti-competitive practices in approximately 31% of public tenders distributed across all Brazilian states. The replication package is available at the Zenodo repository (Silveira et al. 2023).

In summary, this work combines economic and statistical analysis with a data-driven framework to investigate collusive market behavior without using labeled datasets. Even (indirectly) relying on patterns found in previous studies on the strategies of bid-rigging cartels, this screening approach can be a valuable tool for competition authorities to identify potential collusion in different markets and circumstances.

The remainder of this paper is organized as follows. Sections 2 – 4 review the literature, detail the methodology, describe the dataset, and introduce the screens, respectively. Sections 5 and 6 present and discuss our main findings, respectively. Section 7 concludes.

2 Literature Review

Harrington (2008) differentiates the literature between structural and behavioral screens for collusive patterns. Structural screens aim at identifying sectors with a high risk of anti-competitive practices. Typically, it assesses the structure of industries that would potentially favor collusive agreements (explicit or tacit), i.e., making it more profitable and easy to reach and sustain (Abrantes-Metz 2013). In contrast, behavioral screens use variables such as prices, quantities, and market share to identify patterns that match the cooperative conduct between firms (Abrantes-Metz and Bajari 2009). They have different levels of complexity to compare the behavior of colluding firms with those potentially behaving competitively.

One vein of the empirical literature on behavioral methods uses econometric and counterfactual analysis to detect collusion in auctions. This approach benefits from well-known episodes of bid-rigging cartels to test (*ex post*) the properties of the bidding functions. *Ex post* analysis may be suitable as a method of testing specific allegations of collusion and for validation purposes. Porter and Zona (1993, 1999), Pesendorfer (2000), and Clark et al. (2018) are examples of this strand of literature. Other works use econometric tests without any previous knowledge (*ex ante*) or suspicions of bid-rigging cartels (Bajari and Ye 2003; Baldwin, Marshall, and Richard 1997; Ishii 2009; Abrantes-Metz et al. 2012; Chassang et al. 2019; Kawai and Nakabayashi 2022). There are also simpler behavioral screens to design effective and preventive tools to fight bid-rigging cartels (Feinstein, Block, and Nold 1985). They are typically guided by descriptive statistics and exploratory analysis. Abrantes-Metz et al. (2006) and Imhof, Karagök, and Rutz (2018) are examples of simple screens used in *ex post* and *ex ante* analysis, respectively.

Our paper relates to the recent literature that integrates economics, statistics, and machine learning methods to detect cartels (Huber and Imhof 2019; Wallimann, Imhof, and Huber 2022). Supported by the evaluation of *ex post* cartel cases, these works offer screens

based on descriptive statistics capable of adapting to different markets and jurisdictions. This literature uses labeled data, in which the competition authority has already collected hard evidence about collusion and revealed information on the strategies adopted by the cartel members – allowing for supervised learning algorithms to recognize collusive patterns (Huber, Imhof, and Ishii 2022; Silveira et al. 2022).[3]

Finally, some readers might wonder whether a purely predictive exercise (such as ours) could be valuable for Industrial Organization researchers and competition policy. We believe it does. We draw the reader's attention to the distinction between causal inference analysis and "prediction policy" problems (Athey and Imbens 2019; Kleinberg et al. 2015). In the former, competition policy are based on understanding and constructing counterfactuals, where the differences between scenarios with and without a given policy would guide the policymakers' decisions. By contrast, our "predictive" analysis is purely descriptive, suggestive, and data-driven. If our predictive outputs suggest signs of collusive behavior, this result may encourage competition authorities to investigate firms involved in the tenders.

3 Methods

We summarize our screening method in four steps. The first step depends on a clustering algorithm to group the tenders based on a set of relevant attributes. The second step uses the attributes within a classifier algorithm to predict the labels (generated by the clustering algorithm). The third step finds out the most relevant features of our method. The fourth step uses the most relevant features of each cluster to identify the ones in which the behavior is consistent with the existence of cartels.

In order to run the first step, we implement the GMCM and use the Bayesian Informa-

3. Our work also dialogues with a field of the literature that analyses the behaviors of co-bidding groups (Conley and Decarolis 2016). Similar to our approach, they do not need labeled data, but different from ours, they use explicitly co-bidding data for grouping firms that bid together in the same auctions.

tion Criterion (BIC) to measure the model's ability to group the data.[4] In the second step, we use the QDA. Note that the GMCM and QDA algorithms share the same assumption about the distribution of the data (Gaussian mixture) – which increases the consistency of our approach. We then use in the third step the PI method to find the most important features to describe the clusters.

Subsection 3.1 reviews the GMCM (McLachlan 1982; Dempster, Laird, and Rubin 1977; Izenman 2008). Subsection 3.2 presents the QDA classifier (Williams 1982). Subsection 3.3 reviews the PI method (Altmann et al. 2010). Subsection 3.4 provides interesting simulated examples to elucidate our approach. Finally, in Subsection 3.5, we apply our method to a labeled database of retail gasoline cartels detected in Brazil without using the labels offered by Silveira et al. (2022). Using real-world data – such as the Brazilian gasoline cartels – we can measure the performance of our method by comparing our findings to the known (actual) labels. In Appendix B.2, we explore different combinations of the algorithms employed in both the unsupervised and supervised stages. In the Appendix B.2.1, we replace the GMCM with the K-means clustering algorithm and consider the logit and QDA in the supervised phase. We provide a comparative predictive analysis in Appendix B.2.2 to legitimate our choices regarding the combination of GMCM and QDA.

3.1 Gaussian Mixture Clustering Model (GMCM)

Suppose we have a dataset of points $X = \{x_1, \cdots, x_N\}$ consisting of N observations of a D-dimensional variable (i.e., X_N is a D-vector), which we want to partition in K clusters (where $K << N$). The GMCM assumes that the random variable x_n is a Gaussian mixture of K-components. Therefore, we describe the elements of each partition $k \in \{1, \cdots, K\}$ of elements of X by a Gaussian distribution and we model the entire dataset by a mixture of Gaussian distributions such as

4. The lower the BIC, the better the clustering/labeling of the data and, therefore, the closer the GMCM will be to the true (and unknown) distribution.

$$p(x_n) = \sum_{k=1}^{K} \pi_k p(x_n|\mu_k, \Sigma_k) \tag{1}$$

where $\pi_1, \cdots, \pi_K$ are the mixing probabilities (the unconditional probability of the class k in the dataset), $p\left(x_n|\mu_k, \Sigma_k\right)$ is the Gaussian likelihood,

$$p\left(x_n|\mu_k, \Sigma_k\right) = \frac{1}{(2\pi)^{D/2}|\Sigma_k|^{1/2}} \exp -\frac{1}{2}(x_n - \mu_k)^T \Sigma_k^{-1}(x_n - \mu_k), \tag{2}$$

μ_k is a D-dimensional mean vector, Σ_k is a $D \times D$ covariance matrix, π is the ratio of a circle's perimeter to its diameter (a parameter of the standard Gaussian distribution – not to be confused with π_k), $|\Sigma_k|$ is the determinant of Σ_k, and T denotes transpose.

We estimate this model using the Expectation–Maximization (EM) algorithm and the trick of creating an additional vector of "missing" dummy variables (Dempster, Laird, and Rubin 1977):

$$d_n = (d_{n,1}, \cdots, d_{n,K}), \tag{3}$$

which indicates the unknown cluster label that x_n belongs to. We can use it to augment each observation and to achieve a vector $\tilde{x}_n = (x_n, d_n)$. We assume that d_n is a single draw from a K-class multinomial distribution. Thus,

$$p\left(x_n|d_n\right) = \prod_{k=1}^{K} p(x_n|\mu_k, \Sigma_k)^{d_{n,k}} \tag{4}$$

is the likelihod of observing x_n given d_n, and the total log-likelihood of observing $\tilde{X}$ (the matrix composed of $\tilde{x}_n$ for $i = 1, \cdots, N$) given the parameters $(\pi_1, \cdots, \pi_K, \mu_1, \cdots, \mu_K, \Sigma_1, \cdots, \Sigma_K)$ is

$$l(\tilde{X}/\pi_1, \cdots, \pi_K, \mu_1, \cdots, \mu_K, \Sigma_1, \cdots, \Sigma_K) = \sum_{n=1}^{N} \sum_{k=1}^{K} d_{n,k} \log(\pi_k p_k(x_n|\mu_k, \Sigma_k)). \tag{5}$$

The E-step computes the expectation of the vector of dummy variables d_n by the posterior probability that x_n belongs to the cluster K using the Bayes Rule and the available parameters μ_k and Σ_k associated with each cluster k. The M-step uses the expected value of d_n to maximize the log-likelihood. The cluster $k \in \{1, \cdots, K\}$ – to which an observation x_n belongs – is the one that presents the largest value of the posterior probability.

The covariance matrix of a Gaussian distribution usually determines the directions and lengths of the axes of its density contours, admitting the following specifications[5]:

(a) *Full* means the components may independently adopt any position and shape. Therefore, we represent each cluster by a different ellipsoid and we need to estimate the full covariance matrix for each of them.

(b) *Tied* means the components have the same shape. Although we may use any type of ellipsoid to represent the clusters, all of them have the same shape. Therefore, we have to estimate only one covariance matrix that all of the clusters have to share and the mean vector is the only attribute that differentiates one of cluster from the other.

(c) *Diagonal* means that covariance matrices of the components are diagonal matrices and the contour axes are oriented along the coordinate axes. Thus, we have only to estimate D parameters for each covariance matrix of each cluster.

(d) *Spherical* means that the covariance matrices of the components are spherical and we need to estimate one parameter for each covariance matrix of each cluster. In this case, the clusters present circular contours.

3.2 Quadratic Discriminant Analysis (QDA)

Suppose we have a dataset of points $\{(x_1, y_1), \cdots, (x_N, y_N)\}$ consisting of N observations of a joint $D+1$-dimensional variable (x, y). QDA assumes that a point x_n belongs to class

5. See Banfield and Raftery (1993), Celeux and Govaert (1995) and Gan, Ma, and Wu (2020) for details.

k if it maximizes its posterior probability given by the Bayes rule

$$k = \arg\max_{l} p(x_n|\mu_l, \Sigma_l)\pi_l, \tag{6}$$

where we define π_l and $p(x_n|\mu_l, \Sigma_l)$ after Eq. (1). Therefore, we estimate $p(x_n|\mu_l, \Sigma_l)$ using the data of each class and π_l is the proportion of the class in the sample (Williams 1982). Note that this is mathematically consistent with the E-step of the GMCM.

3.3 Permutation Importance (PI)

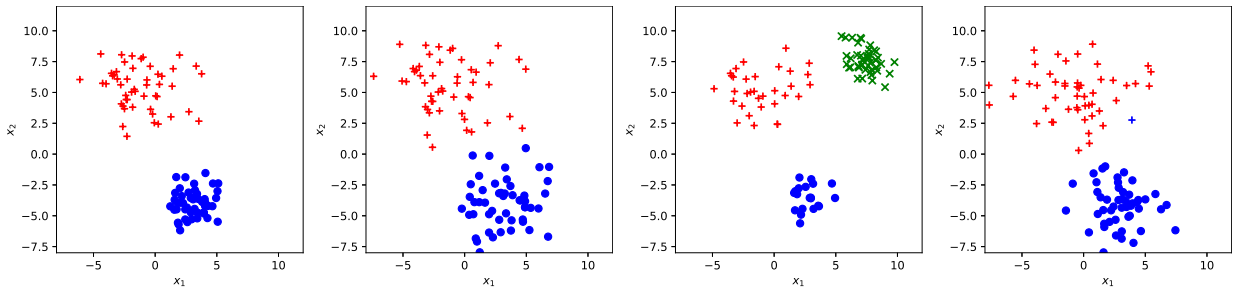
Suppose we have a supervised model such as QDA that depends on D attributes, and we want to identify the most "predictively important" of them. The idea behind the PI method is simple. The relevant attributes are the ones that may compromise the prediction ability of the model when replaced with a shuffled version of them. Therefore, we may summarize this method using the following steps (Altmann et al. 2010): (i) Randomly mix the data for that specific attribute while keeping the values of the other attributes constant; (ii) Generate new predictions based on the randomized values and rate the quality of new predictions; (iii) Rank the importance of the attributes following the decrease in the quality of new predictions compared to the original ones. After calculating PI scores for each attribute, we can rank them with respect to their predictive relevance.[6]

3.4 Simulated examples

In this subsection, we test our method in a controlled environment. We generate clusters using a known distribution and we apply our method to this data. GMCM identifies these

6. We are aware that Principal Component Analysis (PCA) is widely used in data science to order the relevance of screens and dimensionality reduction (Izenman 2008). It is generated as a linear combination of the original set of variables. Our scope of analysis considers different profiles of collusive bidding behavior, and the correct interpretation of such dimensionality-reduction technique may lead to a loss of essential information to identify anti-competitive patterns and generate inaccurate predictions.

clusters.[7] Then, we estimate the QDA using the labels associated with the clusters as the dependent variable. Then, we evaluate the PI of each variable and evaluate the out-of-sample accuracy of the classifiers. Table 1 summarizes the outcomes. Column [1] names the most relevant characteristic of each computational exercise. Column [2] presents the distribution mixture used to generate the clusters. Column [3] presents the parameters of this distribution. Column [4] presents the PI of the variables considered in the exercise using the QDA model. Column [5] presents the out-of-sample accuracy of the classifiers. Each row of this table explores a different situation. The "baseline" simulation in Figure 1a generates two clusters using a Gaussian mixture. The "larger variance" exercise (Figure 1b) is a minor variation of the baseline case, where the variance of the points in the clusters is the only difference. In this case, we investigate the variance effect on the variables used to generate the clusters – in line with the PI criteria – and, also, how it can influence the accuracy of the classifier. In the "irrelevant variables" and "correlated irrelevant variables" simulations, we add two variables x_3 and x_4 that we do not use to generate the cluster.



(a) Baseline (b) Larger variance (c) Three clusters (d) Other distribution

Figure 1: Simulated clusters. The "baseline" case also includes "irrelevant variables" and "correlated irrelevant variables". The"red" color represents cluster 1, "blue" cluster 2, and "green" cluster 3. Different markers mean different outputs of the training data of the classifier. "+" means cluster 1, "o" means cluster 2 and "x" means cluster 3.

The difference between the exercises is that, in the former, these variables are generated independently from the baseline variables (used to generate the clusters). In the latter case,

7. We use the *Full* specification for the covariance matrix in the GMCM in order to make clear that our results do not depend on the known correct specification, which is, for the case of the Gaussian generated clusters, the *Diagonal* one.

we generate variables that correlate with the baseline variables. In both cases the PI of variables x_3 and x_4 is smaller than the PI of variables x_1 and x_2. However, the PI is much smaller when generating the additional variables independently, but not so small in the second case due to the correlation among these variables. The "three clusters" simulation in Figure 1c presents the situation where the supervised learning is not binary (i.e., it is multinomial) – the task is to group the data into three (or more) clusters.

[1]	[2]	[3]			[4]				[5]
Name	**Distribution**	**Clusters**			**PI**				**Accuracy**
		1	2	3	x_1	x_2	x_3	x_4	
Baseline	Gaussian	Mean: $[-1, 5]$ Covariance: diag(4,2.56)	Mean: $[3, -4]$ Covariance: $1.21I$		2.600 ± 0.326	2.669 ± 0.359			100%
Larger variance	Gaussian	Mean: $[-1, 5]$ Covariance: diag(6.25,4)	Mean: $[3, -4]$ Covariance: $4I$		1.024 ± 0.156	1.052 ± 0.165			96%
Irrelevant variables	Gaussian	Mean: $[-1, 5]$ Covariance: diag(4,2.56)	Mean: $[3, -4]$ Covariance: $1.21I$		2.791 ± 0.393	3.112 ± 0.393	0.335 ± 0.075	0.063 ± 0.029	100%
Correlated irrelevant variables	Gaussian	Mean: $[-1, 5]$ Covariance: diag(4,2.56)	Mean: $[3, -4]$ Covariance: $1.21I$		3.113 ± 0.375	4.341 ± 0.495	0.634 ± 0.105	1.404 ± 0.164	100%
Three clusters	Gaussian	Mean: $[-1, 5]$ Covariance: diag(4,2.56)	Mean: $[3, -4]$ Covariance: $1.21I$	Mean: $[7.5, 7.5]$ Covariance: $1I$	1.964 ± 0.253	1.881 ± 0.186			100%
Other distribution	Logistic	Location: $[-1, 5]$ Scale: [1.69.1.21]	Location: $[3, -4]$ Scale: $[1, 1]$		0.875 ± 0.125	0.844 ± 0.143			100%

Table 1: Results of the computational exercise using the GMCM clustering in the unsupervised stage.

Notes: The sample size of all clusters is 100. The sample size to test the classifier's out of sample performance is 25. We use small samples because in many situations the samples available for this kind of empirical exercise is small. In the "irrelevant variables" exercise we use the Gaussian distribution with mean $[1, 2]$ and covariance matrix given by $1.69I$ to generate x_3 and x_4. In the "correlated irrelevant variables" simulation, we generate x_3 and x_4 using $[x_3, x_4]' = \rho[x_1, x_2]' + (1 - \rho)[x_3^p, x_4^p]'$, where we use $\rho = 0.25$ and $[x_3^p, x_4^p]$ are the values of $[x_3, x_4]$ in the previous exercise (mean $[1, 2]$ and covariance matrix given by $1.69I$).

The last exercise considers a logistic mixture (Figure 1d). This is equivalent to a situation where the assumption about the generating process of our random variables is wrong. Therefore, we demonstrate that our method works correctly in a controlled environment presenting the required ability to separate the points in clusters and correctly classify them.

The PI can highlight the relevance of each variable. Although the benchmark models K-means and logit present stronger modeling assumptions than the GMCM and the QDA classifiers[8], respectively, these results are preserved when GMCM is combined with logit and when we replace the GMCM with the K-means and combine it with the logit and the QDA classifiers (see Appendix B.1).

3.5 Labeled database: the gasoline cartels detected in Brazil

In this section, we apply our screening method to the retail gasoline cartels detected in Brazil (Silveira et al. 2022). The hard evidence collected by the Brazilian antitrust authority revealed explicit selling price coordination strategies in the following cities: Belo Horizonte, Brasília, Caxias do Sul and São Luís. In such cartel agreements, the average price typically increases – as the cartel members seek to raise their profit. Also, we commonly observe a lower – and persistent – price dispersion.[9] Silveira et al. (2022) uses weekly data to calculate the standard deviation (PriceSd) and the coefficient of variation (CV) of the gasoline sales price. The explanatory variable PriceSd aims to capture price rigidity. The CV is calculated as PriceSd divided by the arithmetic mean of the gasoline selling price. Variance-based screens suggests that we may observe lower PriceSd and CV during cartel periods. For that reason, we focus on these attributes as screens to evaluate the matching between the labels generated by our cartel detection methodology and the known (actual) labels.[10]

8. GMCM assumes clusters can be elliptical, while K-means considers clusters are spherical. In particular, we can derive a (soft) form of the K-means clustering approach as a Maximum Likelihood (ML) estimator with Gaussian distributions with equal and spherical covariance matrices (MacKay, Mac Kay, et al. 2003). On the other hand, while logistic regression uses a linear decision boundary to separate the classes, the QDA uses a quadratic one.

9. There is a vast literature that uses variance-based screens to show how collusive behavior affects price dispersion (Eckert and West 2004; Connor 2005; Abrantes-Metz et al. 2006; Harrington Jr and Chen 2006; Bolotova, Connor, and Miller 2008; Abrantes-Metz 2012; Perdiguero and Jiménez 2020; Silveira et al. 2021).

10. We acknowledge that cartels can affect other strategic variables rather than PriceSd and CV. To avoid them being missed, it is worth constructing a screening method based on a set of explanatory variables that may recognize the (different) strategic profiles cartels use to sustain the agreement.

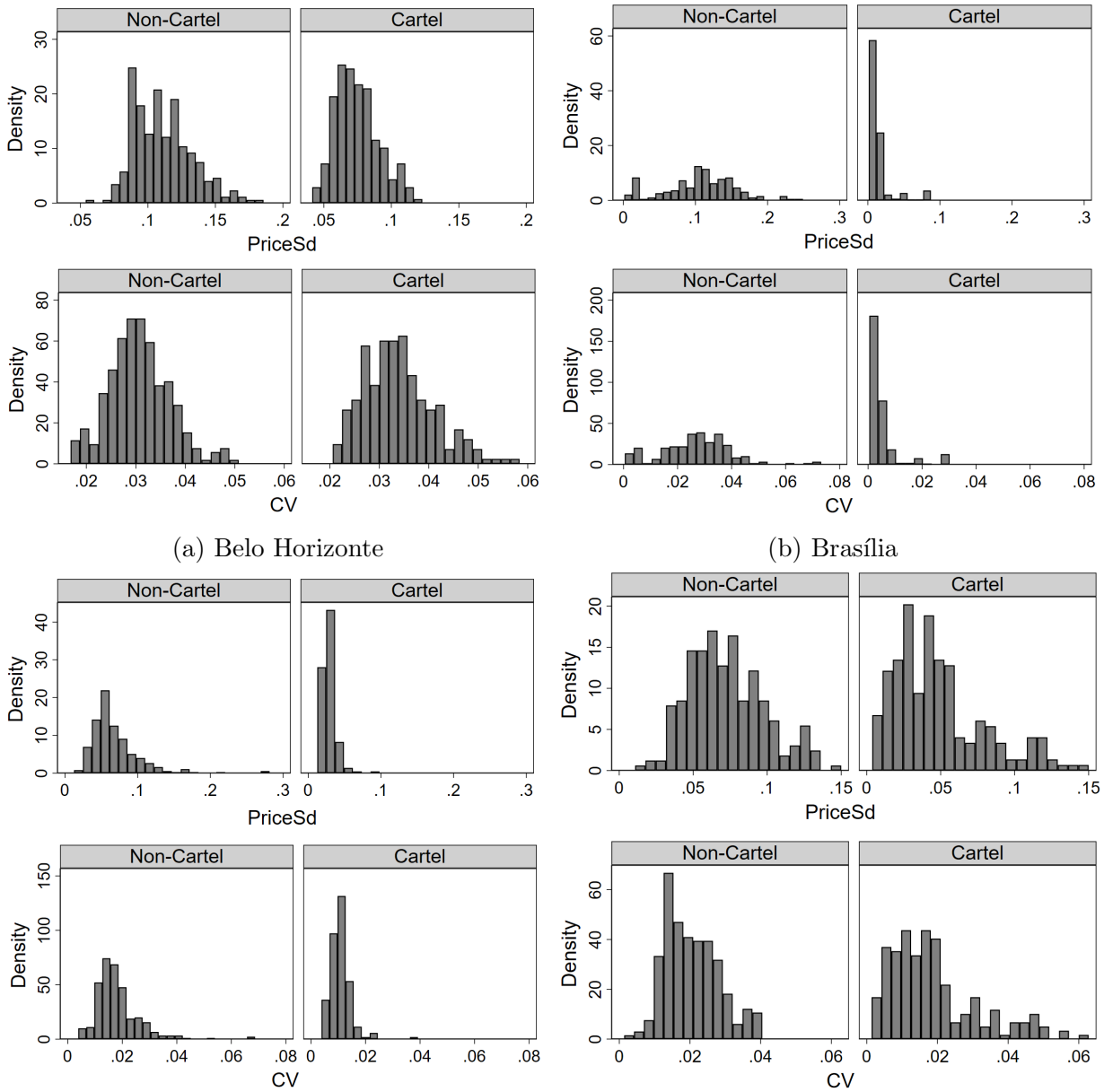
Panel (A) Belo Horizonte	Obs	Mean	Std. Dev.	Panel (B) Brasília	Obs	Mean	Std. Dev.
Cartel Period (01/2004–04/2008)				**Cartel Period** (11/2009–11/2015)			
PriceSd	220	0.074	0.016	PriceSd	311	0.016	0.016
CV	220	0.034	0.007	CV	311	0.005	0.006
Non-Cartel Period (01/2014–04/2019)				**Non-Cartel Period** (12/2015–04/2019)			
PriceSd	276	0.111	0.022	PriceSd	178	0.106	0.049
CV	276	0.031	0.006	CV	178	0.027	0.013
Panel (C) Caxias do Sul	Obs	Mean	Std. Dev.	Panel (D) São Luís	Obs	Mean	Std. Dev.
Cartel Period (01/2004–07/2007)				**Cartel Period** (01/2007–03/2011)			
PriceSd	178	0.028	0.009	PriceSd	213	0.048	0.031
CV	178	0.011	0.004	CV	213	0.019	0.013
Non-Cartel Period (03/2013–04/2019)				**Non-Cartel Period** (10/2014–04/2019)			
PriceSd	306	0.068	0.034	PriceSd	236	0.073	0.026
CV	306	0.018	0.008	CV	236	0.021	0.008

Table 2: Descriptive Statistics of the weekly PriceSd and CV in each evaluated city.

The problem considered in our paper is more challenging than the problem considered in Silveira et al. (2022). In both works, we want to predict whether the dynamics of retail gasoline prices are consistent with the presence of cartels, i.e., we want to predict y as "cartel" and "non-cartel" instances. Bearing this in mind, let X be the matrix of attributes (explanatory variables) associated with the price of gasoline, where each row represents one sample period and each column a different attribute. Let y be the information on whether there is an ongoing cartel agreement (or not) in each sample period.[11] The supervised ML method of Silveira et al. (2022) uses the joint distribution of (X, y). Instead, here we only use the distribution of X to separate the "cartel" from the "non-cartel" instances. Therefore, our method works well when the distribution of X in periods of "cartel" and "non-cartel" are different. The key point is to understand how meaningful – and aligned with economic intuition – these clusters are. Our method indicates the presence of cartels when, in one or more clusters, we find explanatory variables that behave in accordance with the pattern typically observed in collusive agreements. For example, the variance-based screens (PriceSd and CV) can indicate anticompetitive pricing strategies, such as price fixing agreements. In particular, if we find a specific cluster with low PriceSd and CV this can serve as evidence of cartel behavior. On the other hand, if the distribution of X

11. In the original data source constructed by Silveira et al. (2022), the information contained in y is based on the hard evidence collected by the Brazilian competition authority. While information about cartel versus non-cartel periods may not always be accurate, it still gives us valuable clues on firms' strategic behavior over these periods.

is not different in "cartel" and "non-cartel" situations, the act of clustering is meaningless. Table 2 presents the descriptive statistics of the explanatory variables used in the labeled dataset exercise.



(a) Belo Horizonte

(b) Brasília

(c) Caxias do Sul

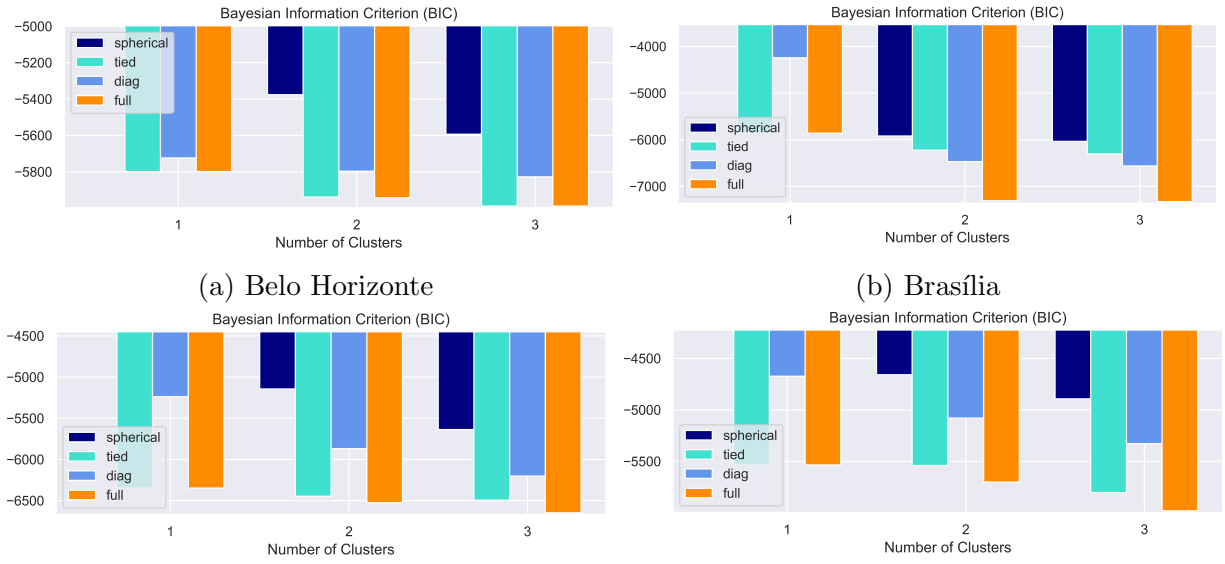
(d) São Luís

Figure 2: The histogram of weekly PriceSd and CV for non-cartel and cartel observations.

The histograms of the PriceSd and CV illustrated in Figure 2 give a clue of the cartels detected more accurately by our method. Comparing the observations labeled as non-cartel and cartel, respectively, notice that the distribution of the prices' standard deviation and

coefficient of variation in Brasília is different. The PriceSd in São Luís also shows a clear and visible different pattern during the cartel and non-cartel periods, respectively. The differences are less clear in Belo Horizonte and Caxias do Sul, especially for CV.

In the sequel, we show how the clustering algorithm splits the groups of data. The GMCM combined with QDA method shows a better performance to identify cartels in gasoline retail in Brasília and São Luís. We use the Bayesian information criterion (BIC) as the metric for fitting the best GMCM. To avoid overfitting, we penalize models with a large number of clusters.[12]



(a) Belo Horizonte

(b) Brasília

(c) Caxias do Sul

(d) São Luís

Figure 3: BIC analysis to group the data based on the dynamics of PriceSd and CV.

However, as illustrated by Figure 3, this is not always an unambiguous criterion (Rokach and Maimon 2005). When there is a small difference in the BIC, whether for choosing the number of clusters or determining the covariance matrix, we favor parsimony. We use the "full" covariance definition in Brasília (Figure 3b) and São Luís (Figure 3d), and the "tied" covariance type in Belo Horizonte (Figure 3a) and Caxias do Sul (Figure 3c). These definitions are parsimonious and preferable in the sense they are almost as lower

12. We evaluate the gradient of the score curve by the BIC: if two consecutive points have the same value, their gradient will be zero. If they have different values, their gradient becomes negative (positive) if the second point has a lower (upper) value. Its magnitude tells us how different the two values are.

as the lowest BIC, and they nicely shape each cluster, improving its visualization and interpretability.

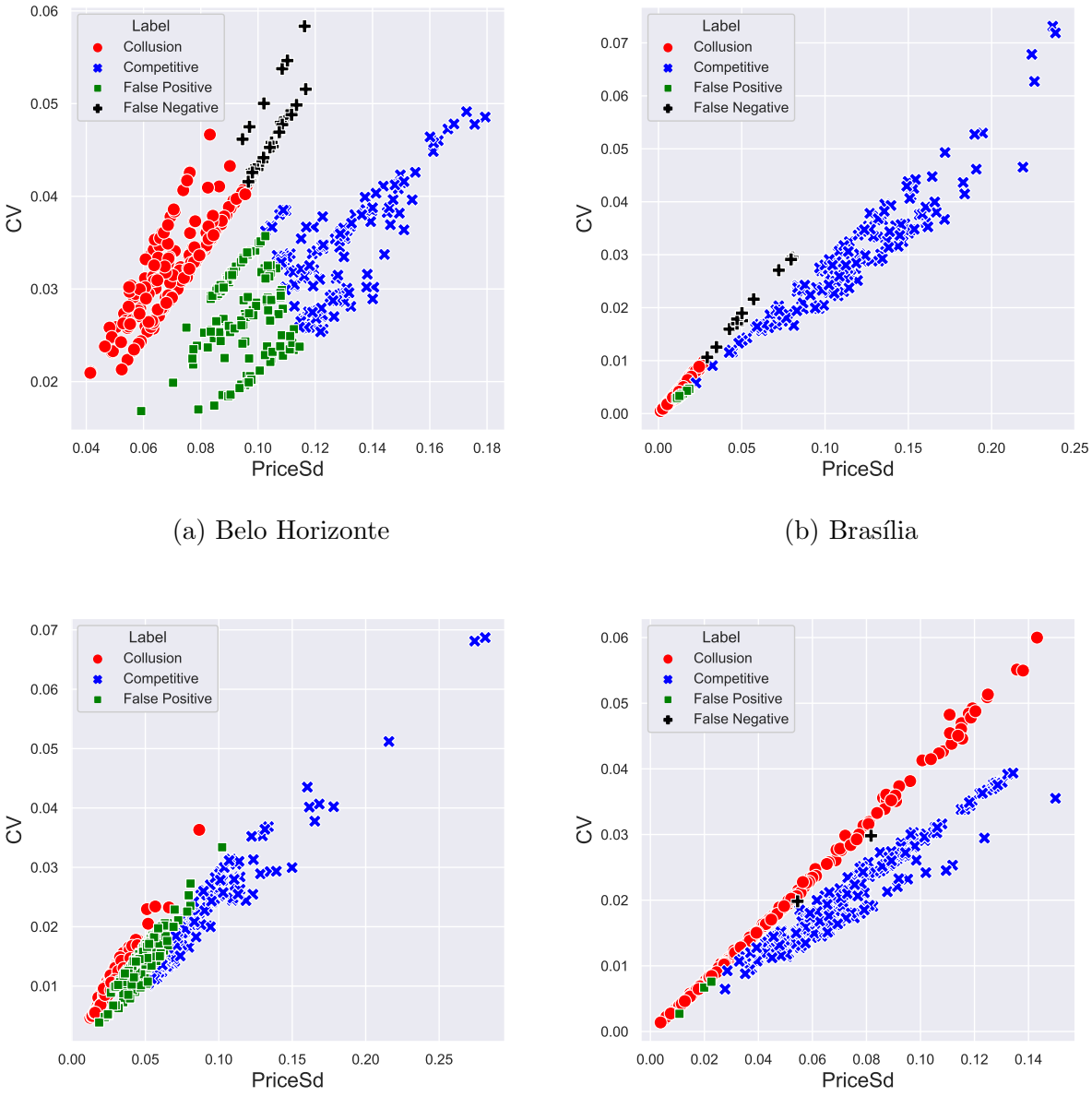
It is worth mentioning that the GMCM algorithm does not use the actual labels of the gasoline cartel data. The clustering step – where we adopt the GMCM (unsupervised ML) – labels the gasoline cartel data solely based on the two explanatory variables we use: PriceSd and CV. Then, we define the label generated by the GMCM as our target variable (y) to fit the QDA algorithm (supervised ML). This procedure allows us to use the PI technique to measure how much the GMCM labeling (and clustering) depends on each explanatory variable.

Table 3 presents the results of the PI technique when applied to the classification provided by the QDA algorithm (supervised ML). In the second column of each panel, we show the estimated increase in prediction error when we replace each statistical screen with its random shuffling counterpart. The drop in the model accuracy captures how much the identification of the labels (cartel versus non-cartel) depends on each statistical screen.

Panel (A) Belo Horizonte		Panel (B) Brasília	
Screen	**Estimated Increase in Prediction Error**	**Screen**	**Estimated Increase in Prediction Error**
PriceSd	0.4000 ± 0.0301	PriceSd	0.4027 ± 0.0315
CV	0.0711 ± 0.0201	CV	0.4014 ± 0.0375
Panel (C) Caxias do Sul		**Panel (D) São Luís**	
Screen	**Estimated Increase in Prediction Error**	**Screen**	**Estimated Increase in Prediction Error**
PriceSd	0.3918 ± 0.0419	PriceSd	0.4978 ± 0.0525
CV	0.1671 ± 0.0834	CV	0.4963 ± 0.0629

Table 3: Estimated Increase in Prediction Error for the statistical screens used in the gasoline cartels.

In Brasília (Panel B) and São Luís (Panel D), we see a balance in the importance of PriceSd and CV. This is not observed in Belo Horizonte (Panel A) and Caxias do Sul (Panel C). Thus, based on this table, we may suspect (*ex ante*) that the accuracy of the predictions in Belo Horizonte and Caxias do Sul are not as good as the ones found in Brasília and São Luís – since, in the former, the forecasts are primarily based on PriceSd.



(a) Belo Horizonte

(b) Brasília

(c) Caxias do Sul

(d) São Luís

Figure 4: GMCM clustering based on PriceSd and CV.

Considering that we propose an exercise based on a labeled dataset, we may evaluate our findings by comparing our predictions with the actual labels. It allows us to discuss the predictive power of our model in the light of the false positive versus false negative rates, which is crucial for competition authorities – and would not be possible to assess with unlabeled datasets. A false positive (negative) corresponds to the Type I (II) error. There is a well-known trade-off between Type I and Type II errors. Positive outcomes

typically require proactive actions. Thus, when there is Type I errors, the (wrong) action taken by the competition agency may compromise its reputation and cause budget waste (Huber and Imhof 2019). Bearing this in mind, we see through Figures 4a–4d that in the cities of Brasília and São Luís, the occurrence of type I error is quite low. However, it is more frequent in the cities of Belo Horizonte and Caxias do Sul.

Panel (A) Belo Horizonte				**Panel (B) Brasília**			
	Precision	Recall	f1-score		Precision	Recall	f1-score
non-cartel	0.84	0.50	0.63	non-cartel	0.86	0.89	0.88
cartel	0.58	0.88	0.70	cartel	0.94	0.92	0.93
accuracy			0.67	accuracy			0.91
Panel (C) Caxias do Sul				**Panel (D) São Luís**			
	Precision	Recall	f1-score		Precision	Recall	f1-score
non-cartel	1.00	0.44	0.61	non-cartel	0.99	0.99	0.99
cartel	0.51	1.00	0.67	cartel	0.99	0.99	0.99
accuracy			0.64	accuracy			0.99

Table 4: The performance of the GMCM in each evaluated city

In Table 4, we summarize the performance of GMCM in each city evaluated. *Precision* quantifies the number of positive class (cartel) predictions that belong to the actual positive class. Thus, high *Precision* is associated with a low incidence of Type I errors. On the other hand, *Recall* measures the number of positive class predictions made up of all positive observations in the dataset. A high *Recall* is associated with a low incidence of Type II errors. To achieve maximal *Precision* (no false positives) and *Recall* (no false negatives) there needs to be an absence of type I and II errors, respectively. The *f1-score* provides a single score that balances (via the harmonic mean) the concerns of both *Precision* and *Recall* in the same measure. We use the classification accuracy score to measure the ability of the GMCM to identify the cartel and non-cartel instances. The classification accuracy is given by the proportion of correct predictions – true positives and true negatives - among the total sample. The overall average classification accuracy considering all four cities is 80.25

Overall, the analysis of the gasoline cartels detected in Brazil contributes to the literature based on variance-based screens, telling us that prices typically have a lower standard

deviation and, potentially, a lower coefficient of variation during cartels based on price coordination agreements. Regarding the accuracy of our methodology, we found a very high performance in the case of Brasília and São Luís, suggesting that the distribution of the explanatory variables is consistent with the presence or not of cartels. In the case of Belo Horizonte and Caxias do Sul, we have still found an acceptable performance of our approach. However, we have observed the unfortunate situation of a higher incidence of false positives.[13]

4 Data Source and Outline

From now on, our objective is to apply our data-driven ML approach introduced in Section 3 to detect bidding patterns consistent with bid-rigging cartels related to road maintenance services in Brazil. This empirical exercise is completely different from those presented in Sections 3.4 and 3.5. In Section 3.4, we evaluate our approach using simulated data we generate by assuming we know its "real" distribution. Differently, Section 3.5 explores our methods on the gasoline cartels detected in Brazil. The datasets used in those analyses have labels – indicating the situations where cartels occurred. Notice that we do not use their actual labels to estimate the ML models. More precisely, we compare the actual labels with those generated (predicted) with our ML method to measure the performance of our approach. In summary, sections 3.4 and 3.5 serve for the sake of didactics and testing the validity of our ML method.

The data we explore in this section does not have labels. Thus, it illustrates the typical situation where our approach may be useful. The dataset comes from the Brazilian Federal

13. In Appendixes B.2, we offer additional analysis (and results) considering other benchmark models, such as the logit classifier and the K-means clustering algorithm, respectively. These models are less flexible: the K-means specify equal (spherical) covariance matrices to shape each cluster, and the logit classifier relies on a generalized linear model. These more restrictive specifications may explain their inferior performance when dealing with real-world data (Appendix B.2.1). The additional findings using the (labeled) Brazilian gasoline cartels datasets presented in Appendix B.2.2 show how our approach may achieve low false positive rates.

Public Procurement portal (*Data Warehouse Comprasnet*). We evaluate 891 unlabeled open tenders from 27 states, including the Capital District of Brazil, from 2012 through 2020.[14]

Section 4.1 presents the statistical screens derived from the bids' distribution in each tender. Section 4.2 summarizes the descriptive statistics for each variable we use to identify potential collusive behavior in the public procurement auction data.

4.1 Statistical Screens

In this section, we introduce statistical screens based on the distribution of bids in each tender. Each input variable we use as screen captures a different aspect of the bidding distribution and potentially allows us to accommodate distinct strategies related to bid manipulation.

Section 4.1.1 presents the screens derived from the discussion in Bhargava, Jenamani, and Zhong (2005) and Tóth et al. (2014), which call attention to two distinct behaviors of deliberately losing firms with a decorative role (so-called "superfluous bidders") in the procurement. The first refers to the high number of active firms in the market who only submit one initial bid in the same tender and then drop out. It is one of the tactics used by the cartel to simulate a competitive dynamic insofar as a significant number of firms participate in the same procurement but offer only a single bid. The second behavior refers to the case in which cartel members artificially bid on open tenders, sometimes discouraging rivals with tight and continuous undercuts and, later on, dropping out.

Section 4.1.2 presents the economic intuition of the variance-based and cover bidding screens. As shown by Huber and Imhof (2019), the coefficient of variation (variance-based) screen may capture bid combination strategies within the same tender. The skewness statistics can work as a "cover bidding" screen. It allows us to potentially identify situations

14. As our data is not labeled, we cannot assume complete (all-inclusive) bid-rigging cartels.

where the cartel members submit bids in excess of the cartel member designated to win the contract.

4.1.1 Elementary Screens

The first statistic we consider as an elementary screen is TotalSingleBids_t, which is the total number of bidders per tender t who placed only a single bid – that is, who only submitted the preliminary sealed bid without actively engaging in the open bidding phase. The statistical screen TotalSingleBids_t is an adaptation of the formula by Tóth et al. (2014) meant to keep track of the so-called *bid suppression*, whereby most bidding ring members abstain from bidding against the member designated to be the winner. In other words, tenders with high rates for the screen TotalSingleBids_t could potentially indicate collusion strategies that emulate competitive dynamics by populating the tender with the so-called fictitious, superfluous, or "frequent loser" bidders.[15] TotalSingleBids_t considers whether a firm may be willing enough to submit the initial sealed bid, thus inflating the number of bidders, and yet abstains from submitting further bids. Let SingleBid_{it} be given by:

$$\text{SingleBid}_{it} = \begin{cases} 1 & \text{if} \quad \text{the firm } i \text{ offered a single bid in tender } t \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

We can calculate TotalSingleBids_t as follows:

$$\text{TotalSingleBids}_t = \sum_i \text{SingleBid}_{it}. \tag{8}$$

Other indicators of superfluous bidders are TotalFirms_t, TotalBids_t, and AverageBids_t. The first corresponds to the number of all firms competing in the same procurement. A low number of bidders in a tender can facilitate coordination between them (or could indicate

15. See Secretariat of Economic Law (2009) for hard evidence of this fake bids in bid-rigging cartels in Brazil.

bid "suppression"). However, tenders with many firms do not necessarily guarantee a competitive environment. The second statistic we use as an elementary screen is the total number of bids offered by each company participating in the open tender. Both the tenders with a high frequency of single bids and those containing a relatively high number of bids placed by the same company may indicate an anomaly – also arising from the strategy of (artificially) emulating competitive bidding behavior. We calculate the AverageBids_t screen as follows:

$$\text{AverageBids}_t = \frac{\text{TotalBids}_t}{\text{TotalFirms}_t} \tag{9}$$

That is, AverageBids_t is a ratio of the total number of bids to the total number of companies participating in the same tender. This potentially captures other patterns related to strategies used by bid-rigging cartels to emulate competitive behavior.

4.1.2 Variance-based and cover bidding screens

Coordinated actions in bid-rigging cartels may increase the mean – as the members of the collusive agreement submit higher bids to raise profits – and affect the dispersion of bids in a given tender. Thus, we consider as a screen the bids' coefficient of variation, i.e., the ratio of the bids' standard deviation to their mean:

$$\text{CV}_t = \frac{\sigma_t}{\mu_t}. \tag{10}$$

The average bid in a tender t (μ_t) necessarily increases as cartel participants submit higher bids to increase their profit. Therefore, the evolution of σ_t determines this effect on CV_t. The scale-invariant property of CV_t is a desirable attribute of "variance-based" screens to detect bid-rigging cartels – as it allows comparison of bidding behavior in tenders where contract values vary substantially (Feinstein, Block, and Nold 1985; Abrantes-Metz

et al. 2006; Imhof, Karagök, and Rutz 2018). Imhof (2019) suggests that σ_t decreases as a result of the communication across cartelized companies. Low values for the coefficient of variation may indicate a bid-rigging cartel.

The difference between the losing bids and the difference between the lowest (winning) bid and the second-lowest bid can suggest a "cover bidding" strategy. Bid-rigging cartels that use cover bidding mechanisms may act as follows. Cartel members not selected to win the contract offer distinguishing higher prices than the designated winner. This strategy can both ensure that the contract is awarded to the member designated by the cartel and suggest that the winning bid is a result of competition between bidders (Pesendorfer 2000). Assuming that the difference between the first and second lowest bids is higher than the difference between the losing bids, Huber and Imhof (2019) calculate the following skewness statistic for the collection of all bids (b_t) in each tender t to validate their argument:

$$\text{Skewness}\,(b_t) = \frac{n}{(n-1)\,(n-2)} \sum_{i=1}^{n} \left(\frac{b_{it} - \mu_t}{\sigma_t} \right)^3. \tag{11}$$

If the differences in "cover bidding" are lower and the difference between the first and second lowest bids are significant, skewness will be more outstanding. Therefore, the "cover bidding" strategy adopted by bid-rigging cartels may affect the distribution of the bids within the same tender t – transforming it into a negatively skewed distribution.

4.2 Summary Statistics

This section summarizes the statistical behavior of each screen and illustrates their descriptive statistics and distributions.

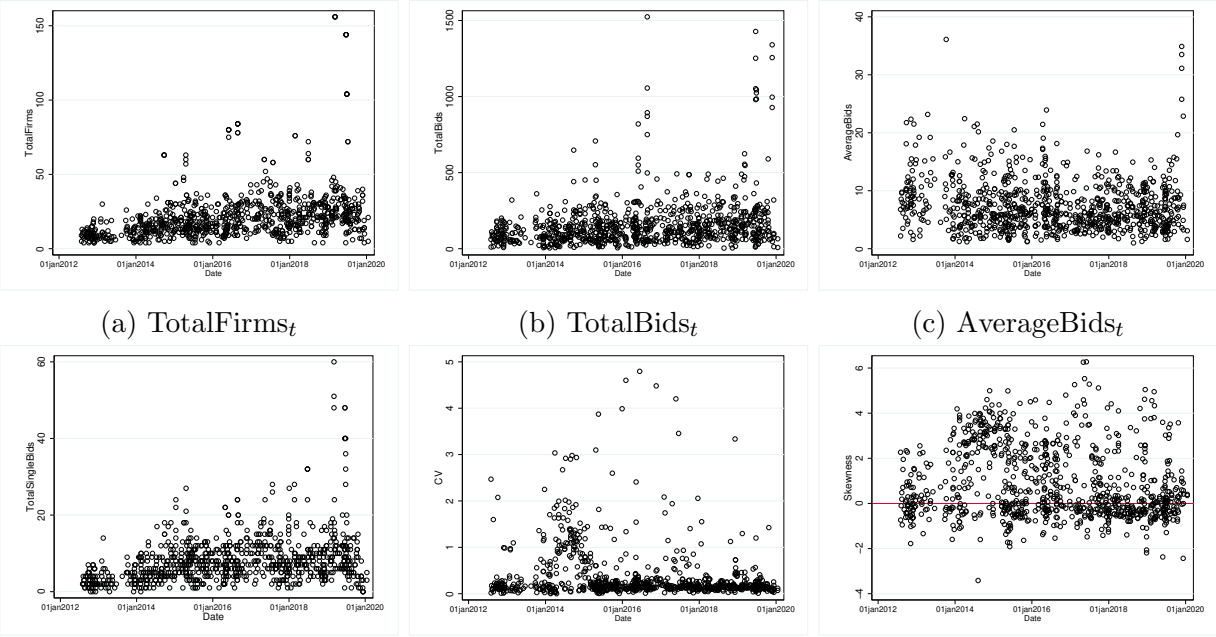
Table 5 details the descriptive statistics for each screen presented in Section 4.1. Figure 5 illustrates the dynamics observed by each screen. Figures 5a, 5b, and 5c show the total number of firms, bids, and the average number of bids between 2012 and 2020, respectively.

We observe, over the years, an increase in the total number of firms and the total number

Screen	Obs	Mean	Std. Dev.	Min	Max
TotalFirms_t	891	22.23	18.15	4	156
TotalBids_t	891	161.05	165.86	4	1524
AverageBids_t	891	7.47	4.46	1	36.1
TotalSingleBids_t	891	8.16	6.38	0	60
CV_t	891	0.38	0.59	0.00	4.80
$\text{Skewness}(b_t)$	891	0.89	1.59	-3.42	6.28

Table 5: Summary of the Descriptive Statistics

of bids offered in the same tender. AverageBids_t, on average, shows a homogeneous behavior across the sample. Figure 5d shows an increase in the number of firms offering a single bid on the same purchase, especially between 2018 and 2019. Figure 5e reveals a significant number of tenders with a low CV_t (closer to zero). In Figure 5f, we notice a consistent number of tenders displaying negative skewness over the years 2012 and 2020.



(a) TotalFirms_t (b) TotalBids_t (c) AverageBids_t

(d) TotalSingleBids_t (e) CV_t (f) $\text{Skewness}(b_t)$

Figure 5: Statistical Screens

5 Results

All the screens considered in our analysis underwent a normalization procedure. This is a well-established and necessary step in clustering tasks when the data scale varies widely

because their range can affect the results (Revelle 1979). We standardize the screens by removing the mean and scaling to unit variance, both of them taken from the whole sample.[16]

Figure 6 illustrates how we choose the optimal number of clusters. The BIC reaches its minimum value when the number of components is four. Then, the BIC takes slightly higher values when the number of clusters is five and six, respectively. In short, this plot suggests dividing the tenders into four groups of tenders would be sufficiently informative.

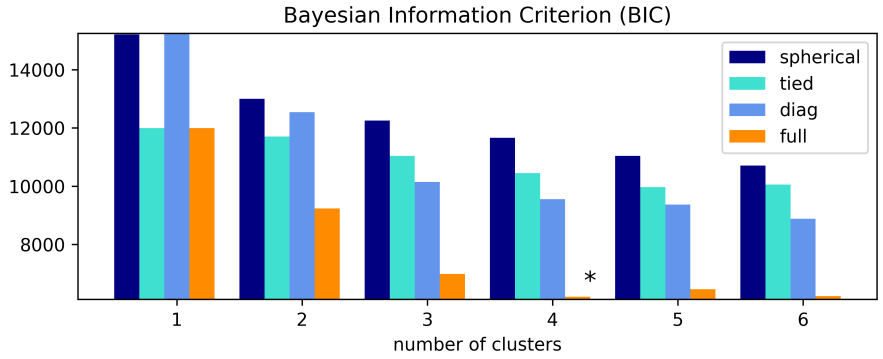


Figure 6: BIC analysis

While Table 6 summarizes the descriptive statistics of the screens for the tenders grouped in each of the four clusters generated by the GMCM, Figure 7 complements this analysis by showing the boxplots of the statistical screens of each cluster.

These results suggest that the statistical screens in each cluster present a particular pattern. We start assessing the difference among these clusters. In the sequel, we show that these differences are statistically significant.

Cluster 0 contains more tenders. It combines a relatively high average number of TotalFirms$_t$ and positive Skewness(b_t). Cluster 1 comprises the higher average values for CV$_t$ and Skewness (b_t). Cluster 2 contains relatively few tenders. For TotalFirms$_t$, TotalBids$_t$, and TotalSingleBids$_t$, the significantly higher average number in cluster 2 stands out. Cluster 3 holds the higher average for the screen AverageBids$_t$, and is – quite explicitly

16. We implement our computation in Python 3.9. For the details of normalization, see Python documentations for sklearn.preprocessing.StandardScaler and sklearn.preprocessing.normalize.

– also a locus of a joint incidence of low CV_t and negative Skewness(b_t).

Among the clusters we have discussed so far, the patterns of CV_t and Skewness(b_t) observed in cluster 3 matches the features consistently observed in the bid-rigging cartels (Huber and Imhof 2019). In addition, the higher average for the screen AverageBids$_t$, in this context, suggests that a small number of firms per tender may submit, in relative terms, a significant number of bids to convey a false impression of high competition.

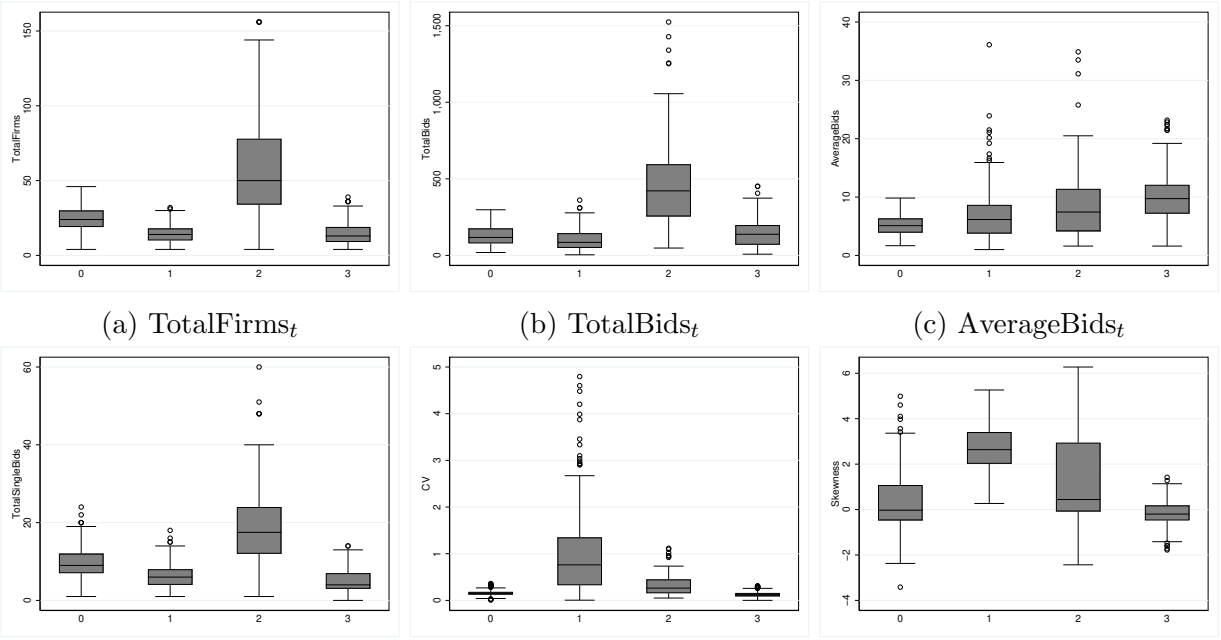
	Mean	**Std. Dev.**	**Min**	**Max**		**Mean**	**Sdt. Dev.**	**Min**	**Max**
cluster 0					*cluster 1*				
TotalFirms$_t$	24.74	7.83	4	46	TotalFirms$_t$	14.24	6.04	4	32
TotalBids$_t$	132.24	65.67	19	298	TotalBids$_t$	99.38	67.70	4	361
AverageBids$_t$	5.23	1.68	1.67	9.83	AverageBids$_t$	7.03	4.62	1	36
TotalSingleBids$_t$	9.67	3.84	1	24	TotalSingleBids$_t$	6.13	3.33	1	18
CV_t	0.16	0.06	0.01	0.36	CV_t	0.98	0.90	0.005	4.80
Skewness(b_t)	0.38	1.24	-3.42	4.99	Skewness(b_t)	2.64	1.03	0.27	5.26
cluster 2					*cluster 3*				
TotalFirms$_t$	60.49	35.09	4	156	TotalFirms$_t$	14.72	7.26	4	39
TotalBids$_t$	488.99	346.99	48	1524	TotalBids$_t$	147.03	93.10	8	452
AverageBids$_t$	9.07	6.66	1.60	34.89	AverageBids$_t$	9.84	4.30	1.60	23.17
TotalSingleBids$_t$	19.46	12.04	1	60	TotalSingleBids$_t$	4.81	2.92	0	14
CV_t	0.34	0.26	0.05	1.12	CV_t	0.13	0.06	0.00	0.31
Skewness(b_t)	1.35	2.07	-2.43	6.28	Skewness(b_t)	-0.16	0.56	-1.78	1.42

Table 6: Descriptive Statistics by each clusters. We have 304 observations (obs) in Cluster 0; 232 obs in Cluster 1; 82 obs in Cluster 2; and 273 obs in Cluster 3.

Although cluster 2 contains the public tenders with the higher incidence (on average) of single bids (TotalSingleBids$_t$) and the total number of bids (TotalBids$_t$), it is not trivial to associate it with a clear – competitive or collusive – behavioral pattern. This is justified by the average value observed for the other screens, such as CV_t and Skewness(b_t). While the expected behavior for potential bid-rigging cartels would result in a combination of low coefficient of variation and negative skewness, comparatively, cluster 2 shows us the opposite. Cluster 0 and Cluster 1 do not present a combination of low CV_t and negative Skewness(b_t).

An issue that arises is if the input variables we use as screens show similar values (patterns) in each of the clusters identified by GMCM. We may use the one-way "analysis of variance" (ANOVA) to compare whether the means of two samples are significantly

different. Through this analysis, we can assess similarities between the clusters. The null hypothesis of the one-way ANOVA is that all the clusters come from the same distribution (with the same mean values). If the statistical screens' means from each cluster come from populations with the same mean values, the observed variance between the clusters' means tends to be smaller. Variations in the screens' average behavior in each cluster imply that samples come from populations with different average values.[17] This evaluation is valuable as it helps to distinguish the patterns captured by the statistical screens' within each group of tender.



(a) TotalFirms$_t$ (b) TotalBids$_t$ (c) AverageBids$_t$

(d) TotalSingleBids$_t$ (e) CV$_t$ (f) Skewness(b_t)

Figure 7: Boxplot by cluster for each screen

Table 7 informs that, for the same screen, at least two clusters are different. However, as we have a total of 4 clusters, it is relevant to determine which ones differ from each other using a posthoc test. We apply Tukey's test to complement the heterogeneity analysis between the groups.[18]

17. In the machine learning literature related to cluster analysis, the one-way ANOVA has been established as an alternative strategy for the selection of attributes (Elssied, Ibrahim, and Osman 2014; Palaniappan, Sundaraj, and Sundaraj 2014).

18. We follow Kotsiantis, Pierrakeas, and Pintelas (2004), who argue that this is the appropriate analytical procedure to use together with the one-way ANOVA.

Screen	**Adj R-squared**	**F-Statistics**	**Prob>F**
TotalFirms_t	0.5173	318.93	< 0.0001
TotalBids_t	0.4067	204.39	< 0.0001
AverageBids_t	0.1840	67.88	< 0.0001
TotalSingleBids_t	0.4164	212.71	< 0.0001
CV_t	0.3724	177.05	< 0.0001
$\text{Skewness}(b_t)$	0.4907	286.88	< 0.0001

Table 7: *One-way* ANOVA

Table 8 shows that, for the following screens, there is no statistically significant difference between some clusters: TotalFirms_t (clusters 3 vs 1), TotalBids_t (clusters 3 vs 0), AverageBids_T (clusters 3 vs 2) and CV_t (clusters 3 vs 0).

TotalFirms_t				**TotalSingleBids_t**			
clusters	Std. Err.	t- statistics	P > t	clusters	Std. Err.	t- statistics	P > t
1 vs 0	0.06	-9.55	<0.0001	1 vs 0	0.07	-8.31	<0.0001
2 vs 0	0.09	22.78	<0.0001	2 vs 0	0.10	16.14	<0.0001
3 vs 0	0.06	-9.53	<0.0001	3 vs 0	0.06	-11.94	<0.0001
2 vs 1	0.09	28.55	<0.0001	2 vs 1	0.10	21.27	<0.0001
3 vs 1	0.06	0.42	0.975	3 vs 1	0.07	-3.03	0.013
3 vs 2	0.09	-28.82	<0.0001	3 vs 2	0.10	-23.85	<0.0001
TotalBids_t				**CV_t**			
clusters	Std. Err.	t- statistics	P > t	clusters	Std. Err.	t- statistics	P > t
1 vs 0	0.07	-2.95	0.017	1 vs 0	0.07	20.18	<0.0001
2 vs 0	0.10	22.44	<0.0001	2 vs 0	0.10	3.13	0.01
3 vs 0	0.06	1.39	0.507	3 vs 0	0.07	-0.76	0.871
2 vs 1	0.10	23.74	<0.0001	2 vs 1	0.10	-10.66	<0.0001
3 vs 1	0.07	4.18	<0.0001	3 vs 1	0.07	-20.42	<0.0001
3 vs 2	0.10	-21.26	<0.0001	3 vs 2	0.10	-3.60	0.002
AverageBids_t				**$\text{Skewness}(b_t)$**			
clusters	Std. Err.	t- statistics	P > t	clusters	Std. Err.	t- statistics	P > t
1 vs 0	0.08	5.11	<0.0001	1 vs 0	0.06	22.84	<0.0001
2 vs 0	0.11	7.66	<0.0001	2 vs 0	0.09	6.88	<0.0001
3 vs 0	0.08	13.7	<0.0001	3 vs 0	0.06	-5.72	<0.0001
2 vs 1	0.12	3.95	<0.0001	2 vs 1	0.09	-8.83	<0.0001
3 vs 1	0.08	7.8	<0.0001	3 vs 1	0.06	-27.64	<0.0001
3 vs 2	0.11	1.51	0.433	3 vs 2	0.09	-10.59	<0.0001

Table 8: Tukey's Test

Now, using the vector of clusters' labels (provided by the GMCM) as the dependent variable and the statistical screens (used to build the clusters) as independent variables, we run the QDA and evaluate the PI of each screen associated with this model.

Table 9 presents the PI of each screen summarizing its contribution to the proper mapping of clusters. The most important statistical screens for the data clustering are those based on the detection of "superfluous bidders".

Screen	**Estimated Increase in Prediction Error**
TotalFirms_t	0.5179 ± 0.0385
TotalBids_t	0.4836 ± 0.0373
AverageBids_t	0.3806 ± 0.0267
$\text{Skewness}(b_t)$	0.2373 ± 0.0322
CV_t	0.2343 ± 0.0322
TotalSingleBids_t	0.1299 ± 0.0296

Table 9: Permutation Importance

The use of the PI technique can increase the competition authority's likelihood of detecting cartels via joint analysis of the screens. In light of what the PI technique reveals, a behavioral pattern associated with the performance of a bid-rigging cartel might be the following: few firms in the same tendering process could be simulating a competitive dynamic by increasing the number of bids. Consequently, it would shift up the average bids per firm. In addition, remember that tenders with few firms can ease communication between them. It can lead to a bidding pattern with low CV and a negative skewness – which may reflect the cartels' bid-rigging strategies.

In other words, the regulator can optimize its screening strategy (with a fixed budget constraint) by jointly considering the variance-based and cover bidding screens. Recall that cluster 3 combines the occurrence of low CV in tenders with a relatively low number of firms and negative skewness. We must remember that the patterns found with the statistical screens we used in this study match the standard bid-rigging behavior recognized by these same statistical screens in previous works (Porter and Zona 1999; Abrantes-Metz et al. 2006; Tóth et al. 2014; Huber and Imhof 2019; Wallimann, Imhof, and Huber 2022).[19]

19. It is worth mentioning that cartel screens applied to retail markets – such as the gasoline cartels in Brazil we discussed in Section 3.5 – also use the low price coefficient of variation as a standard "variance-based" screen to detect (potential) price-fixing agreements.

Figure 8 shows, based on cluster 3, the suspected cartelized tendering distributed across all Brazilian states.

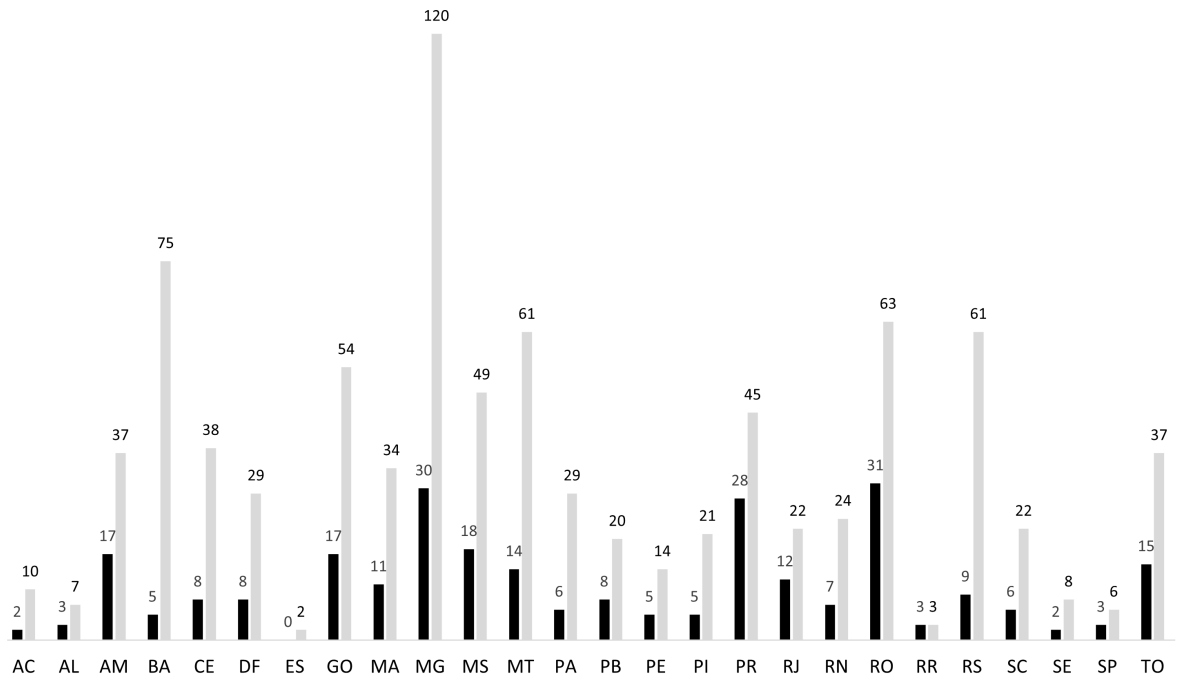


Figure 8: The black bars represent the tenders in cluster 3 by each Brazilian state. The bars in gray aggregate all tenders of all clusters.

Of this set, the states of Amazonas (AM), Goiás (GO), Minas Gerais (MG), Mato Grosso do Sul (MS), Mato Grosso (MT), Paraná (PR), Rio de Janeiro (RJ), Roraima (RO), and Tocantins (TO) stand out. Paraná, Amazonas, and Roraima are the states with the highest proportion of suspected cases. Therefore, the proposed method suggests that the competition authority needs to pay more attention to these public procurements.

6 Discussions

We discuss in this section the ideas that drive our work, the results we have found, the limitations, and possible improvements. We split this section as follows. In Subsection 6.1, we review our methodological choices. In Subsection 6.2, we highlight the limitations of our data-driven approach.

6.1 Our Methodological Choices

Our *ex ante* analysis depends on four different choices. First, we choose screens that may indicate collusive strategic profiles. To seek bidding patterns consistent with bid-rigging cartels related to road maintenance services in Brazil, we use statistical screens previously adopted in empirical studies based on behavioral screens as input variables (attributes) in our proposed method. The combination of "elementary", "variance-based" and "cover bidding" screens may help competition authority to recognize (ex-ante) the vast profile of strategies practiced by bid-rigging cartels. Second, we need to select the clustering model to group the observations. In theory, we can use any clustering algorithm. In this matter, we choose the GMCM due to its simplicity and flexibility regarding (technical) attributes and specifications.[20] Third, we need to choose a supervised classifier to replicate the labels of the clustering algorithm. In this step, we select the QDA due to its flexibility and simplicity. Conveniently, it shares the same assumptions as the GMCM.[21] Fourth, to identify the most important statistical screens, we choose PI as an evaluation criterion because it is the "benchmark" in interpretable machine learning approaches (Molnar 2020).

6.2 Limitations

The variables' statistical properties may impose some limitations on our method. More precisely, when the distribution of the majority of the selected explanatory variables does not allow us to distinguish between collusive and competitive market strategies, we may end up with clusters with ambiguous patterns. This can be a source of inaccurate predictions, harming competition authorities' reputation. One way to overcome this drawback is to use economic analysis and intuition to choose screens that would present different

20. In Appendix B, we offer a comparative analysis with the K-means, one of the most widely used clustering algorithms that rely on less flexible attributes and specifications than GMCM.

21. Also in Appendix B, we make a comparison with logit, a widespread model. In ideal cases, the QDA and logit models behave similarly. In other specific situations, logit loses attractiveness, as its decision frontier is linear, while QDA is quadratic.

distributions in the presence of cartels. However, the choice of our statistical screens and their interpretation is guided by results (and patterns) widely discussed in previous works that sought empirical evidence on cartel performance strategies – whether in retail markets or public tenders. Thus, the need for (empirical or economic-based) evidence to choose the statistical screens also bounds our work.

7 Conclusion

We design a data-driven method to detect collusive behavior in situations where we do not have labeled data. The replication package is available at the Zenodo repository (Silveira et al. 2023). Aware that antitrust practitioners and competition authorities must often anticipate the cartels' movements, our unsupervised ML method – relying on the cartel identification ability of previously studied statistical screens – can provide an additional tool for the detection and prosecution of cartels.

Acknowledgement

This work extends part of the research undertaken by the authors while Douglas Silveira was a Research Associate at Ipea under the Executive Program of Cooperation between Ipea and the UN Economic Commission for Latin America and the Caribbean (ECLAC) − contracts 2500253107 and 2500287988. Daniel O. Cajueiro thanks CNPQ (grant 302629/2019-0) for partial financial support. The authors also thank the editor for the thoughtful comments and suggestions and the two anonymous referees who helped us to improve the manuscript.

References

Abrantes-Metz, Rosa, and Patrick Bajari. 2009. "Screen for Conspiracies and Their Multiple Applications." *Antitrust* 24:66.

Abrantes-Metz, Rosa M. 2012. "Screens for conspiracies and their multiple applications." *Competition Policy International* 8 (1).

———. 2013. "Roundtable on Ex Officio Cartel Investigations and the Use of Screens to Detect Cartels." *Available at SSRN 2343465.*

Abrantes-Metz, Rosa M, Luke M Froeb, John Geweke, and Christopher T Taylor. 2006. "A variance Screen for Collusion." *International Journal of Industrial Organization* 24 (3): 467–486.

Abrantes-Metz, Rosa M, Michael Kraten, Albert D Metz, and Gim S Seow. 2012. "Libor Mnipulation?" *Journal of Banking & Finance* 36 (1): 136–150.

Altmann, André, Laura Toloşi, Oliver Sander, and Thomas Lengauer. 2010. "Permutation Importance: A Corrected Feature Importance Measure." *Bioinformatics* 26 (10): 1340–1347.

Athey, Susan, and Guido W Imbens. 2019. "Machine Learning Methods that Economists Should Know About." *Annual Review of Economics* 11 (1): 685–725.

Bajari, Patrick, and Lixin Ye. 2003. "Deciding Between Competition and Collusion." *Review of Economics and Statistics* 85 (4): 971–989.

Baldwin, Laura H, Robert C Marshall, and Jean-Francois Richard. 1997. "Bidder Collusion at Forest Service Timber Sales." *Journal of Political Economy* 105 (4): 657–699.

Banfield, Jeffrey D, and Adrian E Raftery. 1993. "Model-based Gaussian and non-Gaussian Clustering." *Biometrics,* 803–821.

Bhargava, Bharat, Mamata Jenamani, and Yuhui Zhong. 2005. "Counteracting shill bidding in online English auction." *International Journal of Cooperative Information Systems* 14 (02n03): 245–263.

Bolotova, Yuliya, John M Connor, and Douglas J Miller. 2008. "The impact of collusion on price behavior: Empirical results from two recent cases." *International Journal of Industrial Organization* 26 (6): 1290–1307.

Bradley, Andrew P. 1997. "The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms." *Pattern recognition* 30 (7): 1145–1159.

Celeux, Gilles, and Gérard Govaert. 1995. "Gaussian Parsimonious Clustering Models." *Pattern Recognition* 28 (5): 781–793.

Chassang, Sylvain, Kei Kawai, Jun Nakabayashi, and Juan M Ortner. 2019. "Data Driven Regulation: Theory and Application to Missing Bids." *NBER Working Paper Series,* no. 25654.

Clark, Robert, Decio Coviello, Jean-François Gauthier, and Art Shneyerov. 2018. "Bid Rigging and Entry Deterrence in Public Procurement: Evidence from an Investigation into Collusion and Corruption in Quebec." *The Journal of Law, Economics, and Organization* 34 (3): 301–363.

Conley, Timothy G, and Francesco Decarolis. 2016. "Detecting Bidders Groups in Collusive Auctions." *American Economic Journal: Microeconomics* 8 (2): 1–38.

Connor, John M. 2005. "Collusion and price dispersion." *Applied Economics Letters* 12 (6): 335–338.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society. Series B (Methodological)* 39 (1): 1–38.

Eckert, Andrew, and Douglas S West. 2004. "Retail gasoline price cycles across spatially dispersed gasoline stations." *The Journal of Law and Economics* 47 (1): 245–273.

Elssied, Nadir Omer Fadl, Othman Ibrahim, and Ahmed Hamza Osman. 2014. "A novel feature selection based on one-way anova f-test for e-mail spam classification." *Research Journal of Applied Sciences, Engineering and Technology* 7 (3): 625–638.

Feinstein, Jonathan S, Michael K Block, and Frederick C Nold. 1985. "Asymmetric Information and Collusive Behavior in Auction Markets." *The American Economic Review* 75 (3): 441–460.

Gan, Guojun, Chaoqun Ma, and Jianhong Wu. 2020. *Data clustering: Theory, Algorithms, and Applications.* SIAM.

Gnanadesikan, Ram. 2011. *Methods for Statistical Data Analysis of Multivariate Observations.* John Wiley & Sons.

Harrington, Joseph E. 2008. "Detecting Cartels." *Handbook of Antitrust Economics,* 213–258.

Harrington Jr, Joseph E, and Joe Chen. 2006. "Cartel Pricing Dynamics with Cost Variability and Endogenous Buyer Detection." *International Journal of Industrial Organization* 24 (6): 1185–1212.

Huber, Martin, and David Imhof. 2019. "Machine Learning with Screens for Detecting Bid-rigging Cartels." *International Journal of Industrial Organization* 65:277–301.

Huber, Martin, David Imhof, and Rieko Ishii. 2022. "Transnational Machine Learning with Screens for Flagging Bid-rigging Cartels." *Journal of the Royal Statistical Society Series A, Royal Statistics Society* 185 (3): 1074–1114.

Imhof, David. 2019. "Detecting Bid-rigging Cartels with Descriptive Statistics." *Journal of Competition Law & Economics* 15 (4): 427–467.

Imhof, David, Yavuz Karagök, and Samuel Rutz. 2018. "Screening for Bid Rigging–Does It Work?" *Journal of Competition Law & Economics* 14 (2): 235–261.

Imhof, David, and Hannes Wallimann. 2021. "Detecting Bid-rigging Coalitions in Different Countries and Auction Formats." *International Review of Law and Economics* 68:106016.

Ishii, Rieko. 2009. "Favor Exchange in Collusion: Empirical Study of Repeated Procurement Auctions in Japan." *International Journal of Industrial Organization* 27 (2): 137–144.

Izenman, Alan Julian. 2008. "Modern Multivariate Statistical Techniques." *Regression, Classification and Manifold Learning* 10:978–.

Kawai, Kei, and Jun Nakabayashi. 2022. "Detecting Large-scale Collusion in Procurement Auctions." *Journal of Political Economy* 130 (5): 1364–1411.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction Policy Problems." *American Economic Review* 105 (5): 491–95.

Kotsiantis, Sotiris, Christos Pierrakeas, and Panagiotis Pintelas. 2004. "Predicting students' performance in distance learning using machine learning techniques." *Applied Artificial Intelligence* 18 (5): 411–426.

Levenstein, Margaret C, and Valerie Y Suslow. 2006. "What Determines Cartel Success?" *Journal of Economic Literature* 44 (1): 43–95.

Lloyd, Stuart. 1982. "Least Squares Quantization in PCM." *IEEE transactions on information theory* 28 (2): 129–137.

MacKay, David JC, David JC Mac Kay, et al. 2003. *Information theory, inference and learning algorithms.* Cambridge university press.

MacQueen, J. 1967. "Classification and Analysis of Multivariate Observations." In *In 5th Berkeley Symposium on Mathematical Statistics and Probability,* 5:281–297. 1.

McLachlan, G.J. 1982. "The Classification and Mixture Maximum Likelihood Approaches to Cluster Analysis." In *Classification Pattern Recognition and Reduction of Dimensionality,* 2:199–208. Handbook of Statistics. Elsevier.

Molnar, Christoph. 2020. *Interpretable Machine Learning: A Guide for Making Black Box Models Interpretable.* Lulu.com.

Palaniappan, Rajkumar, Kenneth Sundaraj, and Sebastian Sundaraj. 2014. "A comparative study of the svm and k-nn machine learning algorithms for the diagnosis of respiratory pathologies using pulmonary acoustic signals." *BMC bioinformatics* 15 (1): 1–8.

Perdiguero, Jordi, and Juan Luis Jiménez. 2020. "Price coordination in the Spanish oil market: the monday effect." *Energy Policy,* 112016.

Pesendorfer, Martin. 2000. "A Study of Collusion in First-price Auctions." *The Review of Economic Studies* 67 (3): 381–411.

Porter, Robert H, and J Douglas Zona. 1993. "Detection of Bid Rigging in Procurement Auctions." *Journal of Political Economy* 101 (3): 518–538.

———. 1999. "Ohio School Milk Markets: An Analysis of Bidding." *RAND Journal of Economics* 30 (2): 263–288.

Revelle, William. 1979. "Hierarchical cluster analysis and the internal structure of tests." *Multivariate Behavioral Research* 14 (1): 57–74.

Rokach, Lior, and Oded Maimon. 2005. "Clustering Methods." In *Data Mining and Knowledge Discovery Handbook,* 321–352. Springer US.

Secretariat of Economic Law, Ministry of Justice. 2009. *Fighting Cartels: Brazil's Leniency Program.* 3rd Edition.

Silveira, Douglas, Lucas B. de Moraes, Eduardo P.S. Fiuza, and Daniel O. Cajueiro. 2023. "Who Are You? Cartel Detection Using Unlabeled Data." *Zenodo Repository.* URL: https://doi.org/10.5281/zenodo.7111547.

Silveira, Douglas, Silvinha Vasconcelos, Paula Bogossian, and Joaquim Neto. 2021. "Cartel Screening in the Brazilian Fuel Retail Market." *EconomiA* 22 (1): 53–70.

Silveira, Douglas, Silvinha Vasconcelos, Marcelo Resende, and Daniel O Cajueiro. 2022. "Won't get fooled again: A supervised Machine Learning Approach for Screening Gasoline Cartels." *Energy Economics* 105:105711.

Tóth, Bence, Mihály Fazekas, Ágnes Czibik, and István János Tóth. 2014. "Toolkit for Detecting Collusive Bidding in Public Procurement. With Examples from Hungary." *Corruption Research Center Budapest: Working Paper Series* CRCB-WP/2014:02.

Wallimann, Hannes, David Imhof, and Martin Huber. 2022. "A Machine Learning Approach for Flagging Incomplete Bid-rigging Cartels." *Computational Economics* (forthcoming), https://doi.org/10.1007/s10614-022-10315-w.

Williams, Byron K. 1982. "A Simple Demonstration of the Relationship between Classification and Canonical Variates Analysis." *The American Statistician* 36 (4): 363–365.

Appendix A Benchmark Models

In this appendix, we review the K-means (unsupervised) and logit (supervised) algorithms. It allows us to compare our main findings with benchmark models used in unsupervised and supervised frameworks. These models rely on less flexible assumptions than the respective models used in the main text, namely GMCM – used in the unsupervised stage –, and QDA – used in the supervised stage –, as we describe below.

Subsection A.1 *K*-means

K-means clustering is a method that aims at partitioning n observations into K ($\leq n$) clusters, assuming each observation belongs to the cluster with the nearest mean (MacQueen 1967; Lloyd 1982; Gnanadesikan 2011). In order to build the set of clusters C, we have to run the algorithm due to Lloyds presented in Figure A.1:

procedure LLOYDS
  Choose K points to the initial clusters
  while C changes **do**
    for all $x \in X$ **do**
      Find cluster C_k with center c_k that is the closest to x (using the distance)
      Add x to C_k
    end for
    for all Cluster k **do**
      Recalculate c_k as the average of all the members of C_k
    end for
  end while
end procedure

Figure A.1: The Lloyds algorithm.

Although it is not explicitly formulated, it is worth mentioning that the K-means algorithm favors equal spherical covariance matrices for each cluster. To determine the optimal number of clusters K necessary to run the K-means algorithm, we may use the Silhouette score (Izenman 2008). It allows us to measure the goodness of fit for the K-means clustering algorithm. For a particular clustering, C_K, of a dataset containing K clusters,

the silhouette score of the ith item is calculated using the mean intra-cluster distance (a_i) and the mean nearest-cluster distance (b_i):

$$s_{i\,K} = \frac{b_i - a_i}{\max\{a_i,\, b_i\}},$$

so that $-1 \leq s_{i\,K} \leq 1$. Values close to 1 indicate that the ith item is well-clustered, and values around -1 indicate poor clustering.

Subsection A.2 Logit

Suppose that y_i, $i = 1, \cdots, n$ comes from a sample of n independent Bernoulli variables. The logit model assumes that the probability $P(y_i = 1)$ subjected to a vector of explanatory variables x_i is given by a generalized linear model as in:

$$P(y_i = 1) = \frac{e^{x_i \beta}}{1 + e^{x_i \beta}}.$$

In order to estimate the vector β of coefficients of this model, we have to maximize the log-likelihood function expressed by

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} \Big[y_i\, x_i\, \beta - \log\left(1 + e^{x_i\, \beta}\right) \Big]. \tag{A.1}$$

Appendix B Additional results using combinations of the benchmark models

We present additional results exploring different combinations of the models employed in the unsupervised (GMCM and K-means) and supervised (logit and QDA) stages. In the following sections, namely Subsections B.1, B.2, and B.3, we revisit respectively the simulated data generated in Subsection 3.4, the gasoline cartel data explored in Subsection

3.5, and the road maintenance data considered in Section 5. Concisely, we provide a comparative analysis of the proposed models to complement the main discussion of the paper.

Subsection B.1 Simulated database

We start this section by exploring our approach using the data generated in Section 3.4 and replacing the QDA model with the logit model. The accuracy obtained from the combination of GMCM with logit in the simulated data exercise is the same obtained by the GMCM with QDA. The only difference noted is the Permutation Importance weights in the simulated scenario with three clusters. Table B.1 informs us that, although with a different magnitude, qualitatively, the results obtained via GMCM and logit converge to those generated with GMCM and QDA (see Table 1).

[1]	[2]	[3]			[4]		[5]
Name	**Distribution**	**Clusters**			**PI**		**Accuracy**
		1	2	3	x_1	x_2	
Three clusters	Gaussian	Mean: $[-1, 5]$ Covariance: diag(4,2.56)	Mean: $[3, -4]$ Covariance: $1.21I$	Mean: $[7.5, 7.5]$ Covariance: $1I$	3.009 ± 0.374	2.910 ± 0.271	100%

Table B.1: GMCM clustering (unsupervised stage) and logit (supervised stage).

We now assess the K-means in a controlled environment. We generate clusters using the same distributions and specifications used in the GMCM's evaluation. Then, we estimate the logit and QDA using the labels associated with the clusters as the dependent variable. In sequence, we evaluate the PI of each variable and evaluate the out-of-sample accuracy of the classifier. Table B.2 summarizes the outcomes. Notice in column [4] that the magnitude of the PI varies in relation to the GMCM (see Table 1). However, the accuracy of the K-means in the first stage is the same as found in the GMCM, qualitatively preserving the results. Therefore, since the data generating process is very well-behaved in controlled environments where we use simulated data, we may conclude that, even with simpler models, we have

found similar results.

[1]	[2]	[3]			[4]				[5]
Name	**Distribution**	**Clusters**			**PI**				**Accuracy**
		1	2	3	x_1	x_2	x_3	x_4	
Baseline	Gaussian	Mean: $[-1, 5]$ Covariance: diag(4,2.56)	Mean: $[3, -4]$ Covariance: $1.21I$		925.909 ± 116.669	954.843 ± 129.114			100%
Larger variance	Gaussian	Mean: $[-1, 5]$ Covariance: diag(6.25,4)	Mean: $[3, -4]$ Covariance: $4I$		963.627 ± 152.356	992.467 ± 159.154			96%
Irrelevant variables	Gaussian	Mean: $[-1, 5]$ Covariance: diag(4,2.56)	Mean: $[3, -4]$ Covariance: $1.21I$		925.926 ± 116.691	948.993 ± 130.806	1.305 ± 1.650	2.666 ± 1.934	100%
Correlated irrelevant variables	Gaussian	Mean: $[-1, 5]$ Covariance: diag(4,2.56)	Mean: $[3, -4]$ Covariance: $1.21I$		927.779 ± 116.872	1256.66 ± 162.138	80.201 ± 15.432	242.255 ± 31.888	100%
Three clusters	Gaussian	Mean: $[-1, 5]$ Covariance: diag(4,2.56)	Mean: $[3, -4]$ Covariance: $1.21I$	Mean: $[7.5, 7.5]$ Covariance: $1I$	698.514 ± 78.958	674.913 ± 62.764			100%
Other distribution	Logistic	Location: $[-1, 5]$ Scale: [1.69.1.21]	Location: $[3, -4]$ Scale: $[1, 1]$		632.648 ± 81.131	624.001 ± 109.837			100%

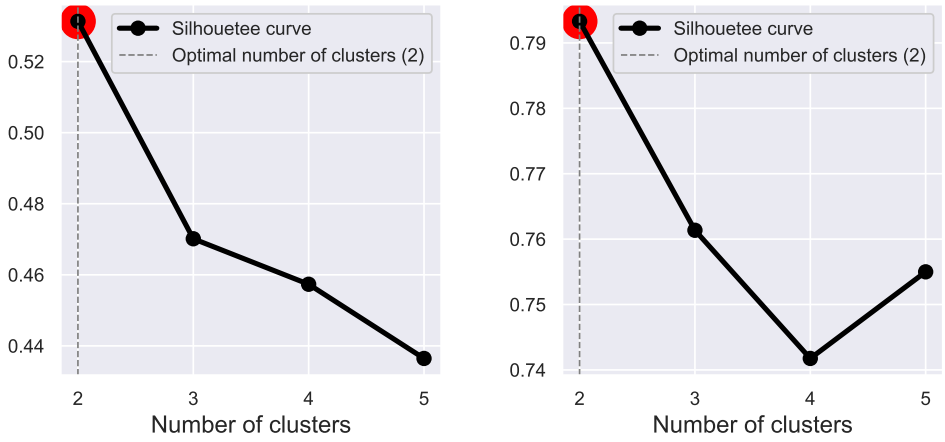
Table B.2: Results: K-means clustering in the unsupervised stage.

Subsection B.2 Gasoline cartels

In this section, we explore different combinations of our approach using the gasoline cartel data considered in Section 3.5. In Subsection B.2.1 we evaluate the performance of the K-means algorithm in clustering the gasoline cartels. It allows us to see how well it separates the data between collusion and competitive observations. Also, based on the PI outcomes, we compare its performance with the GMCM algorithm. In Subsection B.2.2, we compare the predictive power of different combinations of our data-driven approach. It is an important step since our goal is to increase competition authorities' capacity to resolve prediction policy problems. To reach that aim, we use the ROC (Receiver Operating Characteristic) curve analysis to assess the true positive and false positive rates.

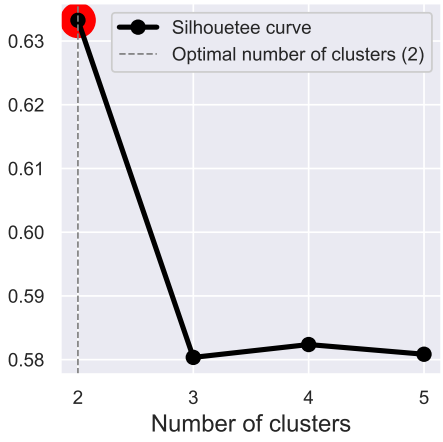
B.2.1 *K*-means *vs* GMCM clustering in the unsupervised stage

Based on the Silhouette score, where values close to 1 indicates a well-clustered data, Figure B.1 shows that 2 is the optimal number of clusters in all four gasoline cartel cases.



(a) Belo Horizonte

(b) Brasília

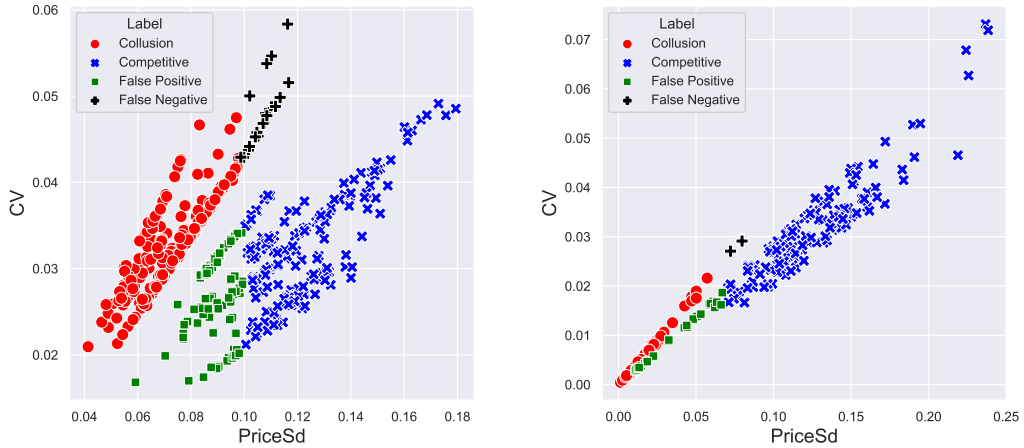


(c) Caxias do Sul

(d) São Luís

Figure B.1: Silhouette score to set the optimal number of clusters in the K-means analysis.

Figure B.2 illustrates the observations correctly identified as collusion and competitive and the number of false negative and false positive for each city analyzed. Table B.3 presents the outputs generated by the K-means clustering in each city evaluated. The K-means outperforms the GMCM in the clustering analysis of the Belo Horizonte gasoline cartel. However, considering the other cities, the K-means overall average accuracy is 72%, while the GMCM achieved 80.25%.



(a) Belo Horizonte

(b) Brasília

(c) Caxias do Sul

(d) São Luís

Figure B.2: K-means clustering based on the dynamics of PriceSd and CV.

Table B.4 presents the PI considering the all possible combinations of the algorithms employed in both the unsupervised and supervised stage. In Belo Horizonte and Caxias do Sul, we see more relevance to the screen PriceSd. It holds for both the combinations using the K-means and GMCM with logit and QDA models.

In Brasilia and Sao Luis, we can observe similar outcomes for both K-means and GMCM combined with the logit algorithm. In contrast, although with different magnitudes, the combination of both the GMCM and K-means with QDA attributes more homogeneous weights to the screens.

Panel (A) Belo Horizonte				**Panel (B) Brasília**			
	Precision	Recall	f1-score		Precision	Recall	f1-score
non-cartel	0.90	0.64	0.75	non-cartel	0.92	0.80	0.85
cartel	0.67	0.91	0.77	cartel	0.89	0.96	0.92
accuracy			0.76	accuracy			0.90
Panel (C) Caxias do Sul				**Panel (D) São Luís**			
	Precision	Recall	f1-score		Precision	Recall	f1-score
non-cartel	0.99	0.33	0.50	non-cartel	0.72	0.53	0.61
cartel	0.46	0.99	0.63	cartel	0.60	0.77	0.67
accuracy			0.57	accuracy			0.65

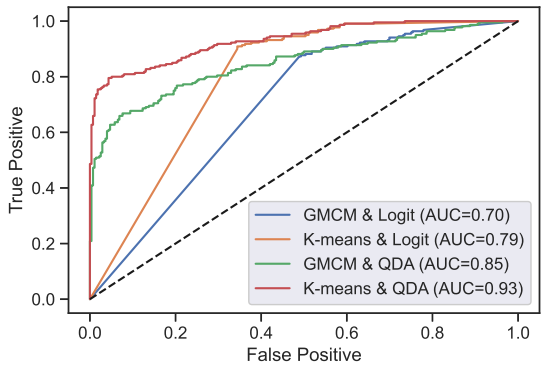
Table B.3: The performance of the K-means in each evaluated city

	***K*-means Clustering**		**GMCM Clustering**	
	Logit	**QDA**	**Logit**	**QDA**
Belo Horizonte				
PriceSd	0.4805 ± 0.0485	0.4725 ± 0.0499	0.3973 ± 0.0178	0.4 ± 0.0301
CV	0.0040 ± 0.0066	-0.004 ± 0.0137	0.0913 ± 0.0234	0.0711 ± 0.0201
Brasilia				
PriceSd	0.4830 ± 0.0430	0.3184 ± 0.0459	0.5782 ± 0.0344	0.4027 ± 0.0315
CV	0.0068 ± 0.0086	0.2966 ± 0.0460	0.0599 ± 0.0264	0.4014 ± 0.0375
Caxias do Sul				
PriceSd	0.3370 ± 0.0485	0.2603 ± 0.0513	0.4836 ± 0.0320	0.3918 ± 0.0419
CV	0.0096 ± 0.0140	0.1247 ± 0.0340	0.1452 ± 0.0329	0.1671 ± 0.0834
Sao Luis				
PriceSd	0.4252 ± 0.0518	0.3496 ± 0.0716	0.5570 ± 0.0645	0.4978 ± 0.0525
CV	0.0178 ± 0.0119	0.2770 ± 0.0753	0.3274 ± 0.0433	0.4963 ± 0.0629

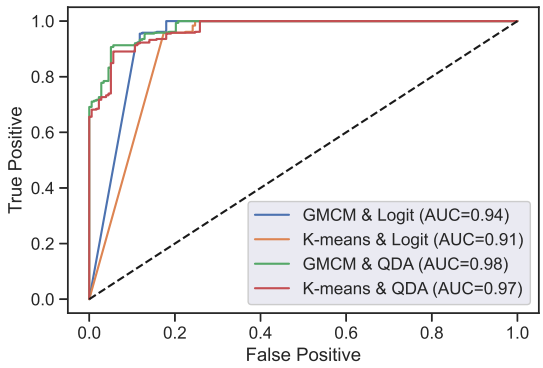
Table B.4: Permutation Importance based on the K-means and GMCM algorithms

B.2.2 Predictive analysis

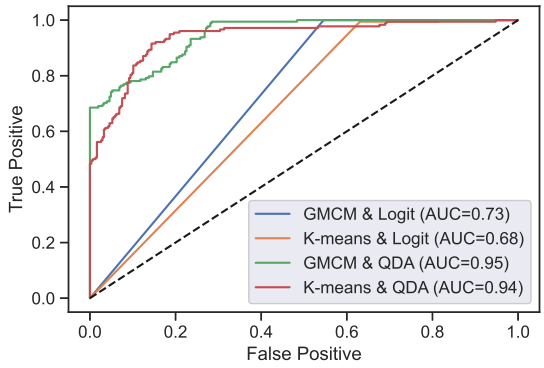
We offer in this section an extra validation step based on the model's predictive power. Using the data from gasoline cartels, we do the following forecasting exercise. First, we train the supervised models (QDA and logit) based on the labels generated by the unsupervised models (K-means and GMCM). Then, we test their performance against the original/actual labels. To this end, we use the ROC curve analysis and the area under the curve (AUC) metrics (Bradley 1997). In summary, the AUC captures the correspondence between the rates of true positives and false positives. Perfect predictions result in an AUC = 1, and an AUC $\leq$ 0.5 indicate low-quality predictions.



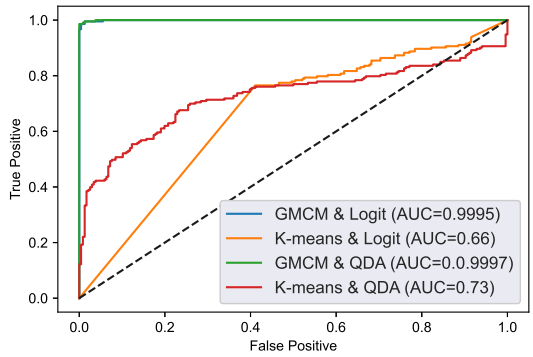
(a) Belo Horizonte



(b) Brasília



(c) Caxias do Sul



(d) São Luís

Figure B.3: ROC Curve and AUC for different configurations of the data-driven approach.

Figure B.3 illustrates the rates of correct predictions for the positive class – collusion, on the vertical axis – versus the fraction of errors for the negative class – competition, on the horizontal axis. In the best scenario, the rates of correct positive and incorrect negative class predictions would be null, respectively. Then, a perfect prediction would occupy the coordinate (0,1) in the upper-left corner of the graph. Furthermore, the ROC curve captures the true and false positive rates for different classification threshold probabilities. Poor quality predictions form a diagonal (dashed in black) line from coordinate (0,0) to coordinate (1,1). Along this dashed line, the algorithms predict all observations as collusion on competitive classes. Predictions lying below this dashed line have little or no ability to help competition authorities distinguish collusion from competitive behavior.

From Figures B.3a – B.3d, we can rank the performance of the different combinations

by the *overall average area under the curve* ($\overline{\text{AUC}}$) criterion as follows: (i) GMCM and QDA ($\overline{\text{AUC}} = 0.94$); (ii) K-Means and QDA ($\overline{\text{AUC}} = 0.89$); (iii) GMCM and logit ($\overline{\text{AUC}} = 0.84$); (iv) K-means and logit ($\overline{\text{AUC}} = 0.76$).

Subsection B.3 Road Maintenance

Table B.5 reports the PI based on the GMCM clustering applied to the road maintenance dataset. In addition to the differences in magnitudes, the ranking obtained with the logit model is slightly different from the QDA. Note that the TotalBids$_t$ screen has the lowest relevance for logit, while it is the second most relevant screen for QDA.

	Logit		**QDA**	
		rank		**rank**
TotalFirms$_t$	0.2619 ±	0.0351 (1st)	0.5179 ±	0.0385 (1st)
TotalBids$_t$	0.0455 ±	0.0145 (6th)	0.4836 ±	0.0373 (2nd)
AverageBids$_t$	0.1657 ±	0.0385 (2nd)	0.3806 ±	0.0267 (3rd)
Skewness(b_t)	0.1403 ±	0.0174 (3rd)	0.2373 ±	0.0322 (4th)
CV$_t$	0.1552 ±	0.0319 (4th)	0.2343 ±	0.0322 (5th)
TotalSingleBids$_t$	0.0799 ±	0.0138 (5th)	0.1299 ±	0.0296 (6th)

Table B.5: Permutation Importance based on the GMCM clustering

In addition, we provide analysis based on the K-means (unsupervised stage) and the logit and QDA classifiers (supervised stage). Figure B.4 shows that 2 is the optimal number of clusters in the K-means applied to the road maintenance database. Table B.6 provides information regarding the behavior of each screen in the two distinct groups of public procurements in the road maintenance sector identified by the K-means clustering. Notice that cluster 0 concentrates 842 tenders, representing almost 95% of the total. We can observe differences in the (average) behavior of each screen. However, it does not reveal a pattern we may unambiguously associate with collusive behavior.

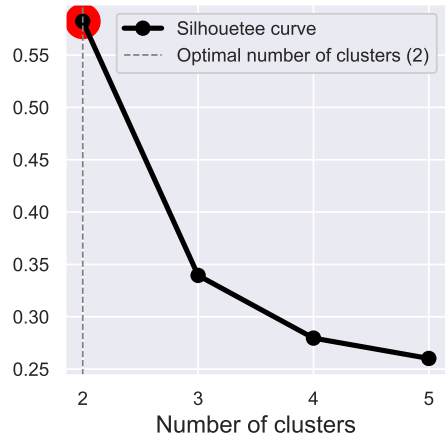


Figure B.4: Silhouette Score to set the optmial number of clusters in the K-means analysis

	Mean	**Std. Dev.**	**Min**	**Max**		**Mean**	**Sdt. Dev.**	**Min**	**Max**
cluster 0 (842 obs)					*cluster 1* (49 obs)				
TotalFirms$_t$	18.92	9.16	4	48	TotalFirms$_t$	79.01	33.89	32	156
TotalBids$_t$	133.30	87.45	4	590	TotalBids$_t$	638	361.03	162	1524
AverageBids$_t$	7.36	4.21	1	36.1	AverageBids$_t$	9.26	7.45	2.13	34.89
TotalSingleBids$_t$	7.24	4.24	0	27	TotalSingleBids$_t$	24.02	13.01	4	60
CV$_t$	0.38	0.61	0.001	4.80	CV	0.31	0.28	0.051	1.12
Skewness(b_t)	0.87	1.57	-3.42	5.29	Skewness(b_t)	1.28	1.91	-0.95	6.28

Table B.6: Descriptive Statistics by each cluster generated by the K-means.

In Table B.7, we report the Mann-Whitney (MW) and the Kolmogorov-Smirnov (KS) test for the screens used in the road maintenance dataset. The MW test informs us whether two independent samples are derived from the same distribution (population), i.e., it tests the hypothesis of a zero-median difference between two independently sampled populations. The KS is a nonparametric test to compare the probability distribution of two samples. The null hypothesis in both MW and KS assumes samples derived from the same distribution. The differences observed between the two clusters obtained via the K-means algorithm are statistically significant at the 1% level for the following screens: TotalFirms$_t$, TotalBids$_t$, and TotalSingleBids$_t$. CV$_t$ and Skewness(b_t) are only significant at 5% for the Kolmogorov-Smirnov test. The screen AverageBids$_t$ does not behave (statistically) differently in the two clusters.

	z-statistic	**p-value MW**	**Ksa**	**p-value KS**
TotalFirms_t	-11.66	< 0.0001	0.93	< 0.0001
TotalBids_t	-11.14	< 0.0001	0.84	< 0.0001
AverageBids_t	-1.39	0.1652	0.13	0.398
TotalSingleBids_t	-9.69	< 0.0001	0.74	< 0.0001
CV_t	-1.23	0.2178	0.22	0.025
Skewness_t	-1.72	0.0857	0.21	0.029

Table B.7: MW and KS tests for the screens in each cluster generated by the K-means.

Finally, Table B.8 reports the PI based on the K-means clustering applied to the road maintenance database. We observe subtle differences in the ordering and magnitude obtained by the PI. However, qualitatively, the results obtained by logit and QDA in the supervised stage show a satisfactory degree of agreement when combined with the K-means in the unsupervised stage.

	Logit		**QDA**	
		rank		**rank**
TotalFirms_t	0.0799 $\pm$	0.0112 (1st)	0.2396 $\pm$	0.0166 (1st)
TotalBids_t	0.0119 $\pm$	0.0099 (3th)	0.1888 $\pm$	0.0154 (2nd)
TotalSingleBids_T	0.0172 $\pm$	0.0101 (2nd)	0.0799 $\pm$	0.0180 (3rd)
AverageBids_t	0.0007 $\pm$	0.0030 (4rd)	0.0769 $\pm$	0.0112 (4th)
$\text{Skewness}(b_t)$	-0.0037 $\pm$	0.0000 (6th)	0.0060 $\pm$	0.0037 (5th)
CV_t	-0.0015 $\pm$	0.0037 (5th)	0.0045 $\pm$	0.0110 (6th)

Table B.8: Permutation Importance based on the K-means clustering

Compared to the results presented in Table B.5, there is a noticeable difference in the magnitude and order of the screens. It may be a consequence of the distinct nature of clustering algorithms. The GMCM supports a wide range of fine adjustments to the model parameters (covariance format, for example). In turn, K-means imposes more restrictions, assuming an equal spherical covariance matrix for the partition of each cluster.

View publication stats