Received: 10 October 2020 | Accepted: 4 January 2022

DOI: 10.1111/rssa.12811

ORIGINAL ARTICLE

Journal of the Royal Statistical Society Series A Statistics in Society

Transnational machine learning with screens for flagging bid-rigging cartels

Martin Huber[1] | **David Imhof**[2] | **Rieko Ishii**[3]

[1]Department of Economics, University of Fribourg, Fribourg, Switzerland

[2]Competition Commission (COMCO), Department of Economics, University of Fribourg, Unidistance, Fribourg, Switzerland

[3]Department of Economics, Shiga University, Hikone, Japan

Correspondence
David Imhof, Department of Economics and Unidistance, Swiss Competition Commission, University of Fribourg, Hallwylstrasse 4, 3003 Bern, Switzerland.
Email: david.imhof5@gmail.com

Abstract

We investigate the transnational transferability of statistical screening methods originally developed using Swiss data for detecting bid-rigging cartels in Japan. We find that combining screens with machine learning (either a random forest or an ensemble method consisting of six different algorithms) to classify collusive versus competitive tenders entails (depending on the model) correct classification rates of 88%–97% when training and testing the method on the Okinawa bid-rigging cartel. As in Switzerland, bid rigging in Okinawa reduced the variance and increased the asymmetry in the distribution of bids. When training the models in data from one country to test their performance in the data from the other country, imbalance increases between the correct prediction of truly collusive and competitive tenders for all machine learners and classification rates go down substantially when using the random forest as machine learner, due to some screens for competitive Japanese tenders being similar to those for collusive Swiss tenders. Demeaning the screens reduces such distortions due to institutional differences across countries such that correct classification rates based on training in one and testing in the other country amount to 85% and to 90% when using the ensemble method as machine learner, which generally outperforms the random forest.

KEYWORDS

Bid rigging, ensemble methods, machine learning, random forest, screening methods

1 | INTRODUCTION

Bid rigging arises in many different markets related for instance to construction, highway maintenance, cement, timber, milk, rice, seafood procession or finance (see Abrantes-Metz et al., 2006; Abrantes-Metz et al., 2012; Asker, 2010; Bajari & Ye, 2003; Baldwin et al., 1997; Banerji & Meenakshi, 2004; Bergman et al., 2019; Feinstein et al., 1985; Hueschelrath & Veith, 2014; Ishii, 2009; Ishii, 2014; Lee & Hahn, 2002; Pesendorfer, 2000; Porter & Zona, 1993; Porter & Zona, 1999, among others papers). It represents a significant share of cartel enforcement in many countries.[1] The OECD estimates that the elimination of bid rigging could help reduce procurement prices by 20% or more.[2] Since public procurement represents approximately 13% of the gross domestic product in OECD countries and 29% of government expenditure,[3] the potential damage of bid rigging can be enormous. For this reason, the OECD recommends promoting pro-active methods for uncovering cartels (OECD, 2014). Responding to such a need for pro-active statistical methods, Abrantes-Metz et al. (2006), Bolotova et al. (2008), Harrington (2007), Abrantes-Metz et al. (2012), Imhof et al. (2018), Huber and Imhof (2019), Imhof (2019) and Chassang et al. (2020) have proposed different methods for dismantling cartels based on descriptive statistics called screens. A legitimate question is whether one can successfully transfer such approaches that have been developed based on observations from one specific case or country to other institutional settings and countries to fight bid-rigging cartels in a transnational context.

In this study, we apply the screening methods suggested by Imhof et al. (2018), Huber and Imhof (2019), Imhof (2019) and Wallimann et al. (2020) in the context of bid rigging in Switzerland to the Okinawa bid-rigging cartel dismantled by the Japanese Fair Trade Commission (hereafter: JFTC) in June 2005. Since the identity of convicted bidders is known, one can categorize tenders as collusive when the respective bidders had participated in the tendering process before the opening of the JFTC's investigation and as competitive after JFTC had sentenced the involved cartel participants. Our data cover three periods: the pre-inspection period including all tenders before the opening of the JFTC investigation in June 2005; the post-inspection period including all tenders between the opening of the JFTC investigation and the amendment of Japanese competition laws in January 2006; and the post-amendment period including all tenders after the amendment of Japanese competition laws. The JFTC sentenced and sanctioned the involved cartel participants in the beginning of the post-amendment period in March 2006.

We show that combining statistical screens with machine learning is capable of dismantling the Okinawa bid-rigging cartel when using the Japanese data for training and testing the statistical models for classifying collusive vs. competitive tenders. Qualitatively, bid rigging affects the distribution of bids in Japanese tenders in the same way as in Switzerland, by reducing the variance and increasing the asymmetry in bids. As machine learners, we consider the random forest, see Breiman (1996), and the 'SuperLearner' ensemble method (a weighted average of several machine learners, namely bagged decision trees, Bayesian additive regression trees, random forest, lasso regression, support vector machines and neural nets), see van der Laan et al. (2008). The overall correct classification rate varies between 88% and 97%, depending on the model (i.e. the number of predictors) and the machine learner considered. This suggests that our method can be an effective tool for promoting competition in public procurement in Japan.

[1]See the OECD report Fighting Bid Rigging in Public Procurement, 2016, page 6, available at the following internet page: http://www.oecd.org/daf/competition/Fighting-bid-rigging-in-public-procurement-2016-implementation-report.pdf.
[2]See OECD Internet page: http://www.oecd.org/competition/cartels/fightingbidrigginginpublicprocurement.htm.
[3]See the OECD report Fighting Bid Rigging in Public Procurement, 2016: http://www.oecd.org/daf/competition/Fighting-bid-rigging-in-public-procurement-2016-implementation-report.pdf.

In a next step, we use the data from one country for training predictive models and those from the other country for testing model performance. The correct classification rates decrease to unacceptably low rates for the random forest whereas the ensemble method continues to have correct classification rates of 82%–87%. Both algorithms exhibit imbalances concerning sensitivity and specificity that is concerning the correct classification rates across truly collusive and competitive tenders. Even for the ensemble method, which clearly outperforms the random forest when training in one and testing in the other country, the imbalance amounts to roughly 20 percentage points in some cases.

This imbalance across collusive and competitive tenders is due to some screens having comparable values for collusive tenders in Switzerland and competitive tenders in Japan, which thus reduce the overall predictive performance of the trained model. A likely reason for this is that the Okinawa Prefecture Government (hereafter: OPG) announces the reserve price or its cost estimate for a contract in the tendering process. Since bidders are disqualified if they bid above the reserve price, the distribution of bids in the tenders procured by the OPG is truncated, implying a lower variance of bids even under competition. Since such announcements do not exist in Switzerland, the distribution of bids is not truncated and has a higher variance.

Our results could suggest that a country's particularities in terms of institutions, for example the judicial and legislative environment as well as the procurement rules and procedures, imply that a predictive model trained in one country does not necessarily perform well when applied across borders in a different cultural context. To mitigate the effect of institutional differences, we in a next step centre the screens within countries such that they have a zero mean. As a result, the ensemble method, which clearly outperforms the random forest in this context, yields overall correct classification rates of 85%–90% and the imbalance in the rates across collusive and competitive tenders disappears. Furthermore, demeaned screens also increase the performance of the random forest in some cases. However, even with demeaned screens, the ensemble method clearly outperforms the random forest when transferring the screening method from one country to the other.

To complement our machine learning approach, we further analyse whether bid rigging affects particular screens similarly across countries based on a regression analysis. While cartel formation influences some screens differentially in Japan and Switzerland, a subset of screens (even if different in levels across countries) experiences more comparable changes across collusive and competitive tenders. This may imply that insights drawn from such screens in one country can be transferred to another country, which appears useful for the development of screening strategies in an international context.

Finally and similarly to Imhof et al. (2018), we perform an ex-ante or screening analysis based on the Japanese data as it might be conducted by a competition agency in order to flag suspicious tenders. To this end, we use all episodes for which collusive and competitive tenders can be distinguished to train the machine-learning based models, in order to predict bid rigging in the remaining Japanese data for which no information on the incidence of collusion is available. We then define a binary variable for the conspicuousness of contracts, taking the value one if both the random forest and the ensemble method classify a tender as collusive. A geographical cluster analysis shows that bidders in the North more frequently participate in tenders classified as conspicuous by our algorithms than firms situated in the centre or in the South of Okinawa. A region-specific investigation of the evolution over time of (i) the coefficient of variation (CV), a screen related to the variance and average of bids in a tender, as well as (ii) the magnitude of the bids normalized by a tender's reserve price corroborates this finding. In general, the bidding patterns as measured by the CV and the normalized bids remain rather stable across periods in the North, even after the JFTC sanctioned the previous cartel participants, while the behaviour of firms changes importantly towards a more competitive pattern in other regions of Okinawa. Those findings support the predictions of the machine learners and point to potentially large gains of the application of machine learning with screens to promote competition in public procurement.

Our analysis contributes to a growing literature developing and implementing screening methods (see Abrantes-Metz et al., 2006; Abrantes-Metz et al., 2012; Bolotova et al., 2008; Froeb et al., 2014; Harrington, 2007; Jimenez & Perdiguero, 2012; OECD, 2014). More specifically, it is related to studies using screens for detecting bid-rigging cartels or analysing the distribution of bids and the statistical patterns created in the distribution when bid rigging occurs (see Chassang et al., 2020; Feinstein et al., 1985; Huber & Imhof, 2019; Imhof, 2019; Imhof et al., 2018; Wallimann et al., 2020). Those papers contrast with papers which suggest and apply econometric tests for detecting bid-rigging cartels (see Aryal & Gabrielli, 2013; Bajari & Ye, 2003; Baldwin et al., 1997; Banerji & Meenakshi, 2004; Bergman et al., 2019; Chotibhongs & Arditi, 2012a,b; Conley & Decarolis, 2016; Imhof, 2017; Jakobsson, 2007; Pesendorfer, 2000; Porter & Zona, 1993; Porter & Zona, 1999). Such tests require data at the firm level, which are typically not readily available. Gathering such data might attract the attention of cartel participants, who might try to hide evidence prior to the opening of an investigation by a competition agency. Moreover, Imhof (2017) shows that econometric tests produce too many false negative results on the Ticino bid-rigging cartel, while screens perform considerably better. More broadly, our paper is also related to further studies discussing bid-rigging cartels or rings (see Asker, 2010; Baldwin et al., 1997; Banerji & Meenakshi, 2004; Conley & Decarolis, 2016; Hueschelrath & Veith, 2014; Ishii, 2009, 2014; Lee & Hahn, 2002; Porter & Zona, 1999).

The remainder of the paper is organized as follows. Section 2 describes the Okinawa bid-rigging cartel and introduces the data. Section 3 discusses the statistical screens and machine learning algorithms, namely the random forest and an ensemble method. Section 4 applies the screens and machine learners for training and testing predictive models for collusion in the Okinawa data. It also analyses the performance when training models in the Japanese data and testing model performance in the Swiss data and vice versa. Furthermore, it investigates whether bid rigging affects screens similarly or differentially across countries by means of a linear regression. Finally, it presents the results of a screening analysis when predicting bid rigging in parts of the Japanese data without information on the actual incidence of collusion. Section 5 concludes.

2 | BID-RIGGING AND DATA

2.1 | The Okinawa bid-rigging cartel and procurement

In our analysis, we consider procurement data from April 2003 to March 2007 obtained from the Okinawa Prefectural Government (hereafter: the OPG), that mainly contain tenders for civil engineering and building construction. In June 2005, the JFTC filed a bid-rigging investigation against a large number of firms involved in those tenders. The construction market in Okinawa exhibits several features facilitating collusion. First, the Okinawa Prefecture consists of 47 islands including Okinawa Main Island, which is the largest island with 1200 km^2. Such geographical conditions make it difficult to enter the market from outside of the prefecture. For the same reason, it is difficult to enter the market of one island from other islands. Hence, it appears inevitable that bidders repeatedly meet when they participate in the tenders.

Second, the OPG used an invitation procedure to procure construction works during the whole period. Under this procedure, the buyer chooses the companies allowed to submit a bid for a contract, which limits the entry to procurement tenders. Furthermore, the OPG announced the identity of the invited bidders prior to each tendering procedure until it changed its bidding system in January 2006. This enabled cartel participants to coordinate their bids and to detect whether an outsider of the bid-rigging cartel was invited in the tendering procedure. Third, the OPG systematically segmented the market and classified its contracts into five ranks from A+ and A to D with regard to the reserve price of each contract. The OPG also classified the bidders into five ranks according to the firm's score of qualification.[4] In the invitation procedure, the OPG selected bidders whose rank of qualification matches the rank of the contract. For instance, for a contract of rank A, the OPG invited bidders of rank A to submit a bid for that contract. However, bidders were often invited to submit bids for contracts with different ranks. For example, bidders with rank A+ were frequently invited to tenders with rank A.

Two major events took place in the market during our data window. First, the JFTC launched an inspection on 8 June 2005 against firms participating in tenders for civil engineering and building construction. In March 2006, the JFTC sanctioned 152 firms that were involved in bid-rigging conspiracies. Penalties included fines and the suspension of bidding in public tenders for 1–6 months and most of the suspended firms were qualified as A+ contractors. 100 cartel participants rigged tenders in civil engineering, 103 participants tenders in building construction. The JFTC investigation revealed that the bid-rigging cartel had started in April 2002 at the latest. Bid coordination took place in 94% and 98% of the civil engineering and construction building tenders, respectively, for contracts of rank A+ in the cartel period.

According to the JFTC, the process of bid coordination took the following form. Prior to each tender, the invited cartel participants met and those interested in the contract expressed their interest. If only one firm was interested, it would be the designated winner of the tender. Otherwise, the interested firms would engage in negotiations for determining the designated winner. During the negotiations, cartel participants considered factors such as the proximity of the contract location to each firm and the amount of each firm's order backlog for allocating contracts among them. When they failed to reach an agreement, the designated winner was chosen by voting among bidders not interested in the contract. Cartel participants then agreed on the winning price and after that, each bidder other than the cartel winner independently calculated a phony bid that was higher than the bid of the designated cartel winner.

The second major event in our data window was the change of the Japanese competition law in January 2006. The Japanese Antimonopoly Act, amended and entering into force in January 2006, increased the fines by 50% and introduced a leniency program aiming at reducing the incentives to collude. Under the leniency program, penalties are reduced for companies providing specific and helpful information on collusion to the JFTC. Also the OPG changed its procurement system in January 2006, reacting to the bid-rigging cartel. First, the OPG increased the number of invited participants in each tendering procedure. For example, with such a change, the mode of invited bidders to contracts of rank A+ increased from 14 to 21. Second, the OPG also stopped announcing the identity of the invited bidders prior to bidding, which increased the difficulty of coordination among potential cartel participants. Under the new amended competition law and the new adapted procurement system, bid rigging became more difficult for potential cartel participants.

[4]In Japan, firms intending to bid in public construction tenders are inspected and qualified according to different criteria, as for example the financial status or the experience in the construction field.

The procurement system of the OPG operates as follows. Whenever the OPG procures a construction contract, engineers estimate the costs of the construction contract using an average firm as proxy. Based on this cost estimate, the OPG determines the reserve price for that contract, as well as its rank. The OPG then invites bidders to the tender with the same rank as the rank of the contract, which determines the number of invited bidders. The format of the tendering procedure follows a first-price sealed-bid auction with a reserve price and a lowest acceptable price. The bidder with the lowest bid wins the contract only if its bid remains between the lowest acceptable price and the reserve price. The OPG rejects bids if they are below the lowest acceptable price, or above the reserve price.[5]

A further change in the procurement system concerned the reserve price, which was announced prior to each auction until January 2006, but was kept secret afterwards. However, this modification did most likely not importantly affect the bidding behaviour, as after January 2006, the OPG instead announced its engineers' cost estimate for each tender, which is typically slightly higher than the reserve price. In contrast, the lowest acceptable price was not revealed to the bidders at any point. It was set to 0.8 times the reserve price in 75% of the tenders until the change in the bidding system in 2006. After that change, the lowest acceptable price randomly oscillated in an interval between 0.8 and 0.85 times the reserve price.

We refer to the first period from April 2003 until the JFTC inspection starting in June 2005 as the 'pre-inspection' period. The second period between the inspection and the amendment of the Antimonopoly Act in January 2006 is the 'post-inspection' period. The final period after the amendment of the law is the 'post-amendment' period, into which also fall the sentences by the JFTC in March 2006.

2.2 | Data

The data are obtained from the OPG's website. For each tender, the date of the tender, the name of the contract for which the auction is conducted, the type of construction (such as civil engineering), the project location, the winner, the winning price, the reserve price, the lowest acceptable price, the identity of each bidder and their bids are available for our analysis. During our data window, 17,798 invitations in 1408 tenders were sent to bidders by the OPG, with 686 invitations being declined. 1297 tenders concerned civil engineering and 111 building construction. We drop 74 bids in four tenders for civil engineering for which the reserve price is unavailable and therefore end up with 17,724 bids in 1404 auctions. The data contain 645, 307 and 452 tenders for the pre-inspection, post-inspection and post-amendment periods, respectively.

In total, the OPG invited 1,767 firms to bid at least once in our data. Based on the JFTC's documents, we identified 150 firms who received a bidding suspension due to being involved in bid rigging. We call these 150 firms 'suspended bidders' or 'cartel participants' and the other firms 'unsuspended bidders'. Unsuspended bidders are, however, not necessarily competitive bidders, especially in the pre-inspection period.

For illustration, Figure 1 shows the distribution of the ratio of the winning bid to the reserve price when only considering tenders with suspended bidders across periods. In the pre-inspection period, the ratio was between 0.95 and 1 in most tenders. In the post-inspection period, we find a greater dispersion of the ratio varying from 0.8 to 0.97. Such a bimodal distribution

[5]In Japan, the lowest acceptable price is imposed in public procurement in order to disqualify a bid which is too low and is unlikely to reflect the true cost of the bidder.

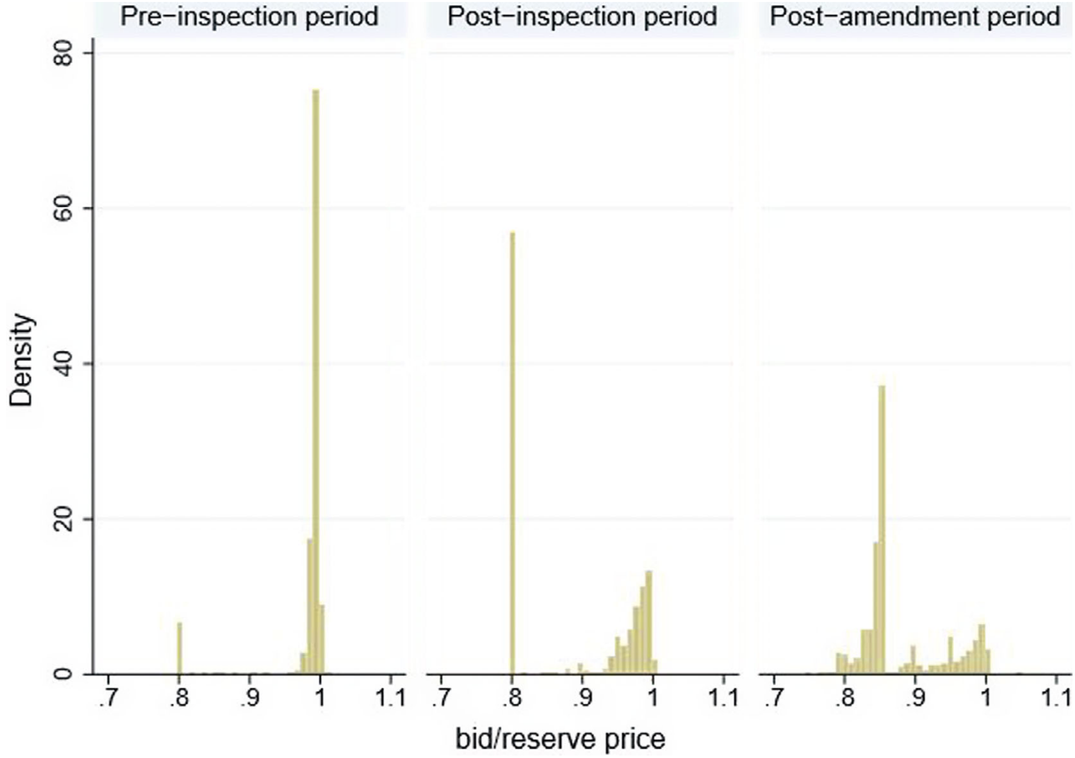

FIGURE 1 The distribution of the ratio of the winning bid to the reserve price among suspended bidders [Colour figure can be viewed at wileyonlinelibrary.com]

typically suggests that cartel participants rigged some contracts but competed on others. Therefore, the post-inspection period appears to be a transition phase. In the post-inspection period, many winning bids were clustered at 0.8 of the reserve price, which corresponded to the lowest acceptable price prior to the change of the procurement system in January 2006, pointing to competition among former cartel participants. We may presume that the 0.8 level for the lowest acceptable price seemed to be common knowledge even though it was officially a secret. In the post-amendment period, when the lowest acceptable price was generally higher and less uniform after the change in the procurement system, the bids were clustered at 0.85. As the lowest acceptable price became less predictable in the post-amendment period, 10% of the bids were rejected for being too low, which was the case for less than 1% of the bids in the earlier two periods.

3 | SCREENING AND MACHINE LEARNING

3.1 | Screens as predictors

We subsequently discuss three types of screens for describing the distribution of bids in tenders: screens for the variance of bids, screens for the asymmetry of bids and one screen for the uniformity of bids (see Huber & Imhof, 2019; Imhof, 2019; Wallimann et al., 2020). Our first variance screen is the CV, a scale-invariant statistic considered for instance in Imhof et al. (2018), Huber and Imhof (2019) and Imhof (2019), which is formally defined as follows:

$$\mathrm{CV}_t = \frac{s_t}{\bar{b}_t}, \tag{1}$$

where s_t is the standard deviation and $\bar{b}_t$ the mean of the bids in some tender t. Another screen related to the support of the bids is the spread (SPD), which is calculated as follows:

$$\mathrm{SPD}_t = \frac{b_{max,t} - b_{min,t}}{b_{min,t}}, \tag{2}$$

where $b_{max,t}$ denotes the maximum bid and $b_{min,t}$ the minimum bid in some tender t (see Wallimann et al., 2020).

In a bid-rigging cartel, participants not supposed to win the contract do not abstain from submitting bids for several reasons. First, a diminishing number of bidders could possibly raise suspicion, especially when the procurement agency invites firms to submit a bid for a contract as it was the case with the OPG. Second, it could be perceived as negative among the other participants not to cover a cartel fellow. Furthermore, the cost of a cover bid in terms of time and effort appears low as it needs not be based on an accurate calculation, but must simply meet the requirement to be higher than the bid of the designated winner. Such a manipulation of bids may induce a convergence of coordinated bids, which we aim to capture by the kurtosis statistic (KURTO):

$$\mathrm{KURTO}_t = \frac{n_t(n_t+1)}{(n_t-1)(n_t-2)(n_t-3)} \sum_{i=1}^{n_t} \left(\frac{b_{it} - \bar{b}_t}{s_t} \right)^4 - \frac{3(n_t-1)^3}{(n_t-2)(n_t-3)}, \tag{3}$$

where b_{it} denotes the bid i in tender t, n_t the number of bids in tender t, s_t the standard deviation of bids and $\bar{b}_t$ the mean of bids in that tender. As all tenders in our sample contain at least five bids, we are able to calculate the kurtosis for each tender.[6]

Concerning empirical evidence on the relevance of variance screens, Feinstein et al. (1985) noticed that the CV was considerably lower under bid rigging when analysing highway construction contracts in North Carolina. Bolotova et al. (2008) observed a reduced variance of prices for a lysine cartel (but not for a citric acid cartel), while Abrantes-Metz et al. (2006) did so for frozen perch when a bid-rigging cartel was in place. Considering the CV, Abrantes-Metz et al. (2012) provided evidence of manipulation in the daily bank quotes submitted to calculate the Dollar Libor. Jimenez and Perdiguero (2012) illustrated that markets with few competitors had a lower price variability and higher prices. In addition to our previous studies for Switzerland (see Huber & Imhof, 2019; Imhof, 2019; Imhof et al., 2018; Wallimann et al., 2020), also other competition agencies have found evidence for a reduced price variance in the case of collusion (see Esposito & Ferrero, 2006; Estrada & Vasquez, 2013; Mena-Labarthe, 2012; Ragazzo, 2012). See also Athey et al. (2004) and Harrington and Chen (2006) for two theoretical (rather than empirical) contributions demonstrating a decrease in the variance of prices when firms collude.

The reduction of the variance can be explained as follows. Under competition, the support of the distribution of bids is determined by the lowest and the highest bid in a tender. If a bid-rigging cartel wants to raise its profit, all cartel participants need to agree to bid higher than the lowest bid under competition, which changes the support of the distribution of bids. It implies a reduction of the variance if the increase above the lowest bid in a competitive situation is

[6]The KURTO can only be calculated for tenders with four bids or more.

substantial (in order to realize profits) and if cartel members conjecture that the procurement agency can likely approximate the highest reasonable bid, for example based on experiences in previous tenders. In this case, the cartel cannot raise the bids of its participants beyond an unreasonably high value without risking to attract the attention of the procurement agency. The manipulated bids are then distributed on a reduced support, implying a decrease in the variance of bids. In the specific case of the OPG, the support of the distribution of bids expressed in terms of the ratio of the bids to the reserve price vary most of the time from 0.95 to 1 in the pre-inspection period and from 0.80 to 1 in the post-amendment period. We therefore expect the variance to be lower in the pre-inspection period compared to the variance in the post-amendment period.

The coordination and manipulation of bids may also affect the symmetry of their distribution, which we aim to capture by specific screens also considered in Imhof et al. (2018), Huber and Imhof (2019), Imhof (2019) and Wallimann et al. (2020). These studies demonstrated that bid rigging increased the asymmetry in the distribution of bids in Swiss tenders. Such changes in the symmetry may be driven by differences between the two lowest bids in a tender or between losing bids, in line with theoretical findings by Marshall and Marx (2007) on manipulated bids in first-price auctions. Furthermore, Chassang et al. (2020) showed that a systematic appearance of substantial price gaps between the two lowest bids in tenders is at odds with competition and might imply bid-rigging conspiracies. At the same time, such gaps affect the symmetry of bids and the related screens.

One screen targeting manipulations in the differences between the two lowest bids is the percentage difference (DIFFP):

$$DIFFP_t = \frac{b_{2t} - b_{1t}}{b_{1t}}, \tag{4}$$

where b_{1t} is the lowest bid and b_{2t} the second lowest bid in some tender t. As an alternative difference measure, one may replace the denominator in (4) by the standard deviation of losing bids, which we refer to as relative distance (RD):

$$RD_t = \frac{b_{2t} - b_{1t}}{sd_{losingbids,t}}, \tag{5}$$

where $sd_{losingbids,t}$ is the standard deviation calculated among the losing bids. Yet another normalization of the difference is based on using the mean difference between all adjacent bids in a tender as denominator, referred to as normalized distance (RDNOR):

$$RDNOR_t = \frac{b_{2t} - b_{1t}}{\frac{\sum_{i=1}^{n_t-1}(b_{i+1,t} - b_{it})}{n_t - 1}}, \tag{6}$$

where n_t is the number of bids and $b_{i+1,t}$, b_{it} are adjacent bids (in terms of price) in tender t, with bids being ordered increasingly. As a variation of (6), we exclude the lowest bid to use the mean difference between adjacent losing bids as denominator, referred to as alternative distance (ALTRD):

$$ALTRD_t = \frac{b_{2t} - b_{1t}}{\frac{\sum_{i=2}^{n_t-1}(b_{i+1,t} - b_{it})}{n_t - 2}}, \tag{7}$$

where b_{1t} is the lowest bid, b_{2t} the second lowest bid, n_t the number of bids and $b_{i+1,t}$, b_{it} are adjacent losing bids in tender t, with bids being ordered increasingly. We also consider the simple (i.e. non-normalized) difference between the two lowest bids: $D_t = b_{2t} - b_{1t}$.

A further screen is the skewness statistic (SKEW) as a standard measure of symmetry in distributions:

$$SKEW_t = \frac{n_t}{(n_t-1)(n_t-2)} \sum_{i=1}^{n_t} \left(\frac{b_{it} - \overline{b}_t}{s_t} \right)^3, \tag{8}$$

where n_t denotes the number of the bids, b_{it} the i^{th} bid, s_t the standard deviation of the bids and $\overline{b}_t$ the mean of the bids in tender t. Finally, we compute the nonparametric Kolmogorov–Smirnov (KS) statistic for measuring uniformity in the distribution of bids (see Wallimann et al., 2020). Even though competitive bids are not necessarily uniformly distributed, we suspect that coordination makes the distribution of bids even less uniform, which can be captured by a change in the following KS statistic:

$$D_t^+ = max_i \left(x_{it} - \frac{i_t}{n_t+1} \right) D_t^- = max_i \left(\frac{i_t}{n_t+1} - x_{it} \right) KS_t = max(D_t^+, D_t^-), \tag{9}$$

where n_t is the number of bids in a tender, i_t the rank of a bid and x_{it} the standardized bid for the i^{th} rank in tender t. The standardized bids x_{it} are the bids b_{it} divided by the standard deviation of the bids in tender t for a normalized comparison of tenders with different contract values.

We have previously mentioned some explanations for why and how bid rigging could affect the screens in a particular way, but it is worth noting our machine learning approach can also be applied when we are fully agnostic about how the screens change from competition to collusion. The single hypothesis on which the power of our method relies is that bid rigging affects the distribution of bids in some (possibly unknown) way that materializes in our screens. If bid rigging did not affect the bidding distribution at all, it would be infeasible to detect bid-rigging cartels by statistical screens.

The question thus arises whether some cleverly managed cartel could beat our screening method by avoiding an impact on the distribution of bids and determining realistic bids for all cartel participants. The goal of the latter would be to impede an effective application of the screening methods as suggested in this paper or the econometric tests formalized by Bajari and Ye (2003). However, the necessity to collect realistic bids of cartel participants would involve substantially higher efforts in coordinating and exchanging the bids compared to the Swiss and Japanese bid-rigging cases. In both countries, the cartels only determined the bid of the designated winner for a contract. The other cartel participants merely submitted cover bids sufficiently high to ensure the contract allocation to the designated winner. In contrast, a sophisticated cartel would succeed in generating coordinated bids with the same properties as competitive bids. Assuming that the goal of such a cartel consists of increasing prices, the distribution of coordinated bids would translate to higher values without otherwise modifying its distributional properties. In this case, engineering cost estimates or benchmarks would represent the last opportunity to detect such an upper shift in the cleverly manipulated distribution of bids. But even then, one could argue that the price increase might be hard to detect if the cartel implements it only gradually rather than abruptly. However, such constraints like a gradual price increase make coordination and information exchange between cartel participants even more complicated, with the threat that these activities are discovered if an investigation is opened. This may jeopardize the stability of a cartel, because even if a cleverly managed cartel might be able to trick screening, there exist further means for dismantling cartels such as leniency programs, whistle-blowing or complaints from customers and procurement agencies, which complement statistical tools for flagging cartels.

3.2 | Application of two selected screens to the Okinawa bid-rigging cartel

In the following, we investigate how the bid-rigging cartel in Okinawa affected the distribution of bids and how cartel participants reacted to the opening of the investigation and the decision with sanctions from the JFTC. We apply the best-performing predictors in the study of Huber and Imhof (2019), namely the CV and the normalized distance, to test if they capture effects or patterns specific to collusion in the distribution of bids for those contracts of type A+ that have been rigged by the Okinawa cartel. This provides evidence on whether changes in the distributional pattern of bids are correlated with the opening of the investigation and the decision of the JFTC.

Figure 2 plots the CV calculated for the contracts A+. Table 1 provides descriptive statistics for the CV across competitive and collusive periods and Table 2 presents Mann–Whitney (MW) and KS tests for differences across those periods. We notice an important increase in the values of the CV in the post-amendment period, which already starts in the post-inspection period. Looking at Figure 2, we observe that the post-inspection period constitutes a transition period between the pre-inspection and the post-amendment period, in which the former cartel participants started to adapt their behaviour. In the post-amendment period, three tenders out of four exhibit a CV equal to or higher than 4.63 as indicated in Table 1. This is larger than the mean of the CV in

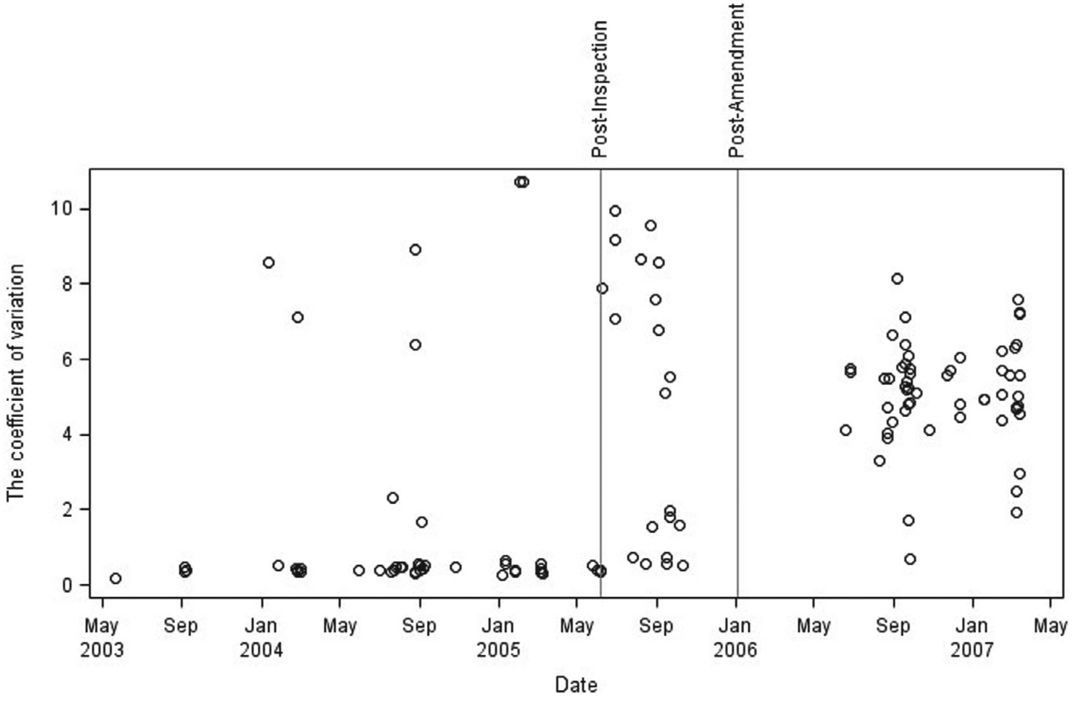

FIGURE 2 The coefficient of variation for the A+ contracts

TABLE 1 Descriptive statistics for the coefficient of variation in the A+ contracts

Periods	Mean	Std	Min	Lower Q.	Median	Upper Q.	Max	N
Pre-inspection period	1.52	2.84	0.17	0.36	0.42	0.54	10.73	48
Post-inspection period	4.00	3.77	0	0.55	1.90	7.73	9.95	24
Post-amendment period	5.58	3.84	0.7	4.63	5.22	5.79	32.15	57

Note: 'Mean', 'Std', 'Min', 'Lower Q.', 'Median', 'Upper Q.', 'Max' and 'N' denote the mean, standard deviation, minimum, lower quartile, median, upper quartile, maximum and number of observations, respectively.

TABLE 2 Statistical tests for the coefficient of variation in the A+ contracts

Periods	z-statistic	*p*-value MW	KS	*p*-value KS
Pre-inspection against Post-amendment	−6.68	0.00	4.27	0.00
Pre-inspection against Post-inspection	4.79	0.00	2.71	0.00
Post-inspection against Post-amendment	−0.10	0.92	1.53	0.02

Note: '*p*-value MW', 'KS' and '*p*-value KS' denote the *p*-value of the Mann–Whitney test, the Kolmogorov–Smirnov statistic and the *p*-value of the Kolmogorov–Smirnov test, respectively.

TABLE 3 Descriptive statistics for the normalized distance in the A+ contracts

	Mean	Std	Min	Lower Q.	Median	Upper Q.	Max	N
Pre-inspection period	4.72	2.31	0	3.52	4.42	6.75	10.4	48
Post-inspection period	2.34	2.19	0	1.00	2.26	2.88	10.14	20
Post-amendment period	1	1.94	0	0.03	0.28	1.02	10.76	57

Note: 'Mean', 'Std', 'Min', 'Lower Q.', 'Median', 'Upper Q.', 'Max' and 'N' denote the mean, standard deviation, minimum, lower quartile, median, upper quartile, maximum and number of observations, respectively.

the post-inspection period as well as the upper quartile in the pre-inspection period. We conclude that the statistical pattern of the bids significantly differs and steadily changes across periods as shown in Table 2. However, the changes appear particularly substantial in the post-amendment period and after March 2006 when the JFTC sanctioned the involved cartel participants for bid rigging.

We also more closely investigate why six tenders exhibit a high CV in the pre-inspection period and we find a reasonable explanation for three of them. For one tender, the JFTC found evidence that the cartel did not find an agreement and that cartel participants competed for obtaining the contract. In two other tenders, a firm not sanctioned by the JFTC participated and was therefore a potential outsider, which might explain why firms behaved in a more competitive way and submitted bids distributed similarly to the bids of the post-amendment period.

Turning to the normalized distance suggested by Huber and Imhof (2019), we observe a substantial change of symmetry in the distribution of bids. Table 3 recapitulates descriptive statistics for the normalized distance in each period and Table 4 presents the results for the MW and the KS tests. Figure 3 confirms the transitory phase of the post-inspection period. In each period, the normalized distance decreases. In the pre-inspection period, the mean and the median of the normalized distance amounts to 4.72 and 4.42, in the post-inspection period to 2.34 and 2.26 and in the post-amendment period to 1 and 0.28. Each period is statistically different from the other

TABLE 4 Statistical tests for the normalized distance in the A+ contracts

Periods	z-statistic	*p*-value MW	KS	*p*-value KS
Pre-inspection against Post-amendment	6.84	0.00	3.95	0.00
Pre-inspection against Post-inspection	−4.17	0.00	2.44	0.00
Post-inspection against Post-amendment	3.11	0.00	2.26	0.00

Note: '*p*-value MW', 'KS' and '*p*-value KS' denote the *p*-value of the Mann–Whitney test, the Kolmogorov–Smirnov statistic and the *p*-value of the Kolmogorov–Smirnov test, respectively.

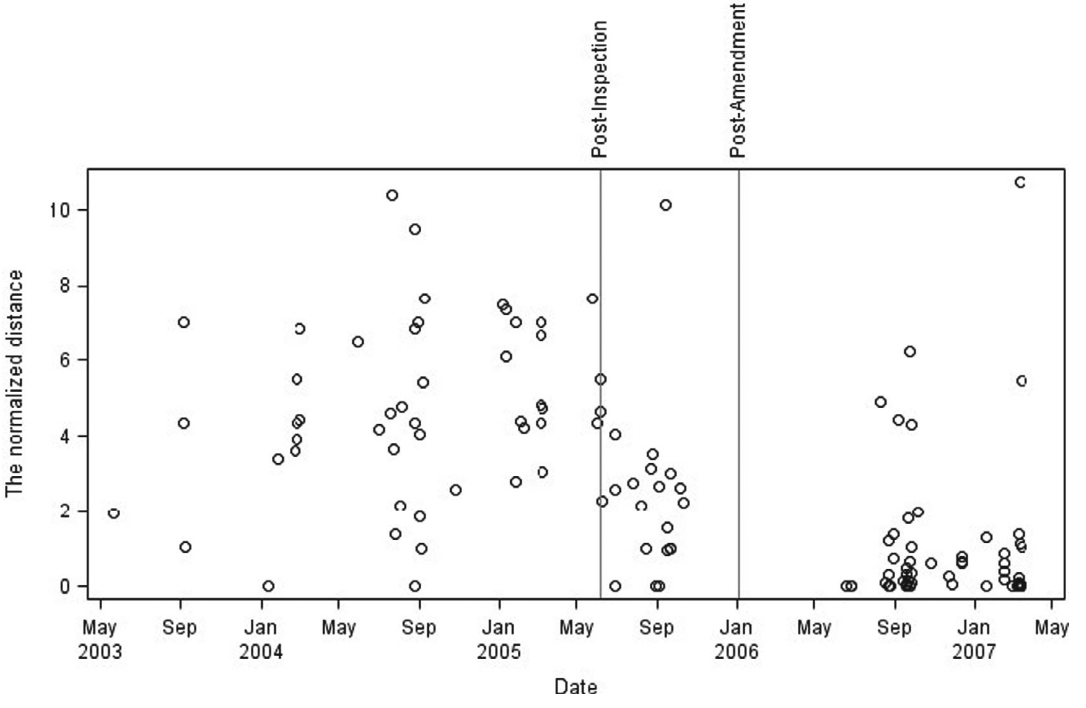

FIGURE 3 The normalized distance for the A+ contracts

periods as attested in Table 4. The higher values of the normalized distance in the pre-inspection period demonstrate that bid rigging in Okinawa produced asymmetry in the distribution of bids compared to the post-amendment period, for which three tenders out of four exhibit a value lower than 1.

3.3 | Models and machine learning

In our analysis based on machine learning, we investigate the performance of seven models that differ in terms of the number of predictors and are referred to as model 1 to model 7. Model 1 includes all screens discussed in Section 3.1 as well as two further characteristics as predictors, namely the number of bids and the contract value of a tender. Model 2 includes all screens, but no further characteristics like the number of bids or the contract value (which in contrast to the screens are scale-variant variables). Model 3 uses a subset of the screens, namely those related to the variance and the asymmetry of bids. Model 4 is based on the screens related to the variance and the uniformity of bids, model 5 on the screens related to the asymmetry and the uniformity of bids. Model 6 includes all the screens from model 2 that we demean by centring them within countries such that they have a zero mean. Finally, model 7 combines the demeaned screens from model 6 and the screens from model 2.

As in Huber and Imhof (2019), we apply a so-called ensemble method as machine learner, which is a weighted average of six algorithms: bagged decision trees, Bayesian additive regression trees, random forest, lasso regression, support vector machines and neural nets. In addition, we consider the performance of the random forests alone as a separate machine learner. For either method, we randomly divide the data into training and test samples, which contain 75% and 25% of the observations, respectively. Cross-validation in the training sample determines the optimal weight each of the six machine-learning algorithms obtains in the ensemble method.[7] To this end we apply the 'SuperLearner' package for 'R' by van der Laan et al. (2008) with default values for bagged tree, random forest, Bayesian additive regression tree, lasso regression, support vector machine and neural net algorithms in the 'ipred', 'partykit', 'bartMachine', 'glmnet', 'kernlab' and 'nnet' packages, respectively, see Peters et al. (2002), Hothorn and Zeileis (2015), Kapelner and Bleich (2016), Friedman et al. (2010), Karatzoglou et al. (2004) and Venables and Ripley (2002).

Once the best model has been selected by finding the optimal weight for each machine learner in the training data, it is applied in the test sample to predict the outcomes and evaluate the out-of-sample performance by comparing the predicted to the actual outcomes. More concisely, we predict the collusion probability in the test sample and classify a tender as collusive whenever that probability is larger than or equal to 0.5 and as competitive otherwise. Our performance measure is the correct classification rate, which is the share of observations in the test data for which the classification based on our prediction matches the actually observed outcomes of collusion or competition.[8] We repeat this procedure of randomly splitting the data into 75% training and 25% test data and evaluating the performance of the optimal model 100 times to calculate the average of the correct classification rate across the 100 data splits, which is our performance measure ultimately considered. For the random forests, the approach is analogous to that using the ensemble method, with the exception that cross-validation is not required in the training data as one need not weight alternative machine learners. As in Wallimann et al. (2020), the implementation is based on the 'randomForest' package for 'R' by Breiman and Cutler (2018), using 1000 trees to estimate the predictive models in the training data.

Random forests and bagged trees as discussed in Breiman (1996), Ho (1995) and Breiman (2001) are so-called decision tree methods. Decision trees are obtained by recursively splitting the data into subsamples in a way that minimizes the discrepancy between actual and predicted incidences of collusion or non-collusion within the subsamples according to some criterion like the Gini index. Both random forests and bagged trees consist of estimating such trees in a large

[7]An alternative approach to learning the optimal weights for the various algorithms by cross-validation is using an unweighted average of the algorithms' predictions. While an optimally weighted average can be expected to have a smaller bias than an unweighted one, the estimation of such weights in the data adds a further variance component to the prediction error of the ensemble method. For this reason, weighting does not necessarily dominate unweighted averaging, see for instance the discussion in Zhou (2012). The relative predictive performance of either approach may vary across different applications.

[8]Huber and Imhof (2019) also consider different probability thresholds than 0.5 to assess the trade-off in false positive and false negative predictions as well as the deterioration in the overall correct classification rate in Swiss data with a comparable number of collusive and competitive bids.

number of samples repeatedly drawn from the original training data and averaging over the tree (or splitting) structures across samples to obtain predictions of collusion. However, one difference is that while bagged trees consider all predictors as candidates for further data splitting at each step, random forests only use a random subset of the total of predictors to reduce the correlation of tree structures across samples and thus, the variance of the machine learner (related to overfitting in correlated samples). Bayesian additive regression trees (BART) as suggested in Chipman et al. (2010) are yet another tree-based method. However, rather than averaged over, trees are grown sequentially (similarly as in a related method called boosting) in order to obtain predictions based on the sequential tree structure. A so-called regularization prior penalizes and thus, avoids too many splits in the tree structure, which would entail an excessive variance due to overfitting.

Lasso regression, which has been proposed by Tibshirani (1996), consists of a penalized regression aiming at finding the best regression fit for modelling the association between the outcome and the predictors, however, under the constraint that the sum of absolute values of the coefficients (or weights) on the predictors does not exceed a certain threshold. Similarly to BART and many other machine learners, the aim of this penalization is to avoid overfitting by giving too much weight (i.e. assign too large coefficients) to relatively unimportant predictors. Support vector machines (SVMs), see Boser et al. (1992), make use nonlinear transformations of the predictors such that a separating (and linear) hyperplane can be fitted across the transformed predictors which classifies the outcomes into collusive and competitive tenders. Finally, neural nets as discussed in McCulloch and Pitts (1943) and Ripley (1996) aim at fitting a system of nonlinear regression functions that flexibly models the association of the predictors with collusion. Specifically, the predictors are used as inputs for nonlinear (for instance logit) intermediate functions, so called hidden nodes, which are themselves associated with the outcome of interest. Depending on the model complexity, hidden nodes may predict the outcome either directly or through other hidden nodes, such that several layers of hidden nodes allow modelling interactions between the functions and thus, predictors. The number of hidden nodes and layers gauges the flexibility of the model, with more parameters reducing the bias but increasing the variance.

4 | EMPIRICAL ANALYSIS

4.1 | Outcome definition and descriptive statistics in the Japanese data

Since we have knowledge about the suspended (i.e. sentenced) bidders by the JFTC, we consider a tender as collusive if a suspended bidder participated in it in the pre-inspection period, which mainly concerned contracts of rank A+ and to a lesser extent of ranks A, B and C. Only very few suspended bidders submitted bids for contracts of rank D and for this reason, we exclude any contracts of rank D from our analysis. We consider a tender as competitive if at least one suspended bidder participated in it in the post-amendment period, thus assuming that the presence of suspended bidders prevents the formation of new cartels in the short run. Since the descriptive price analysis in Section 2.2 and the analyses with the CV and the normalized distance in Section 3.1 suggest that the post-inspection period is a transition phase, we discard those tenders in order to avoid the risk of contaminating the sample by an inaccurate classification of competitive and collusive tenders.

Tables A1 and A2 in Appendix report descriptive statistics for all predictors in the competitive and collusive tenders. Table 5 provides the results of MW and KS tests for statistical

TABLE 5 Statistical tests for the predictors between collusive and competitive tenders

Screens	z-statistic	*p*-value MW	KSstat	*p*-value KS
NUMBID	14.66	0.00	6.84	0.00
VALUE	1.43	0.1530	1.13	0.1542
CV	13.94	0.00	7.90	0.00
KURTO	−1.10	0.2734	1.64	0.0093
SKEW	9.16	0.00	4.68	0.00
D	1.56	0.1189	1.65	0.0084
RD	−10.95	0.00	5.64	0.00
RDNOR	−8.92	0.00	4.68	0.00
ALTRD	−9.20	0.00	4.81	0.00
DIFFP	1.95	0.0507	2.43	0.00
SPD	14.27	0.00	7.95	0.00
KS	−13.97	0.00	7.90	0.00

Note: 'NUMBID', 'VALUE', 'CV', 'KURTO', 'SKEW', 'D', 'RD', 'RDNOR', 'ALTRD', 'DIFFP', 'SPD' and 'KS' denote the number of bids in a tender, the contract value, the coefficient of variation, the kurtosis, the skewness, the absolute difference between the two lowest bids in a tender, the relative distance, the normalized distance, the alternative distance, the percentage difference, the spread and the Kolmogorov–Smirnov statistic, respectively. 'Screens', 'z-statistic', '*p*-value MW' denote the screens tested, the z-statistic of the Mann–Whitney test and the *p*-value of the Mann–Whitney test, respectively. 'KSstat' and '*p*-value KS' denote the Kolmogorov–Smirnov statistic and the *p*-value of the Kolmogorov–Smirnov test, respectively.

differences in the predictors across competitive and collusive tenders. Regarding the variance screens, we find substantial differences between both periods for the CV and the spread (SPD) and the statistical tests reject the null hypothesis of no difference between the periods for both the CV and the SPD as reported in Table 5. The differences are less substantial for the KURTO but the KS test rejects the null hypothesis at any conventional level of significance. With regard to the KS statistic, both statistical tests reject the null hypothesis and we conclude that the distribution of bids is significantly less uniform in collusive than in competitive tenders.

Among the screens for the asymmetry of bids, we find important differences across periods with respect to the RD, the RDNOR and the ALTRD. As reported in Table 5, all statistical tests reject the null hypothesis. In addition, the differences in the non-normalized difference between the two lowest bids (D) are statistically significant across collusive and competitive periods, even though they are not that substantial in absolute terms. Concerning the percentage difference between the two lowest bids (DIFFP), we find important differences between the means and standard deviations whereas the medians are more similar. In this case, the statistical tests also yield rather low *p*-values.

We consider in model 1 two further tender-specific characteristics that are not among the scale-invariant screens outlined in Section 3.1. The first is the number of bids in a tender (NUMBID), for which we again find significant differences, explained by the OPG's decision to increase the number of invited bidders in the post-amendment period. The second characteristic is the contract value in million JPY (VALUE) for which we find no statistical significant difference. All in all, however, we find substantial and very often statistically significant differences in the screens across collusive and competitive periods. Similarly to Switzerland, the Okinawa

TABLE 6 Correct classification rates for the Japanese data

	Random forest			Ensemble method		
	All tenders	**Coll. tenders**	**Comp. tenders**	**All tenders**	**Coll. tenders**	**Comp. tenders**
Model 1	0.937	0.910	0.958	0.970	0.978	0.960
Model 2	0.889	0.853	0.919	0.890	0.901	0.876
Model 3	0.894	0.870	0.914	0.895	0.911	0.875
Model 4	0.889	0.862	0.911	0.885	0.894	0.875
Model 5	0.887	0.856	0.911	0.891	0.906	0.873
Model 6	0.890	0.919	0.855	0.902	0.908	0.896
Model 7	0.889	0.918	0.854	0.884	0.893	0.874

Note: 'All tenders'. 'Coll. tenders' and 'Comp. tenders' denote all tenders, the collusive tenders and the competitive tenders, respectively. Model 1 includes the screens from model 2 as well as the number of bids and the contract value of a tender. Model 2 includes the screens discussed in Section 3.1; Model 3 the screens for the variance and the asymmetry of bids; Model 4 the screens for the variance and the uniformity of bids; Model 5 the screens for the asymmetry and the uniformity of bids. Model 6 includes all the demeaned screens from model 2. Model 7 combines the screens from models 2 and 6.

bid-rigging cartels affected the distribution of bids by reducing the variance and increasing the asymmetry.

4.2 | Training and testing in the Japanese data

Table 6 reports the results for the out-of-sample performance of our machine learners when training and testing in the Japanese data. Model 1 yields a correct classification rate of 93.7% for the random forest and an even higher one of 97% for the ensemble method, which are the highest rates obtained among all models considered. Therefore, including the number of bids or the contract value as further predictors on top to the screens increases the accuracy by five to eight percentage points depending on the machine learner. Models 2 to 5 entail rather similar correct classification rates for the random forest of 88.9%, 89.4%, 88.9% and 88.7%, respectively. The same applies to the ensemble method, with rates of 89%, 89.5%, 88.5% and 89.1%, respectively, such that both machine learners have a very decent and similar predictive performance. Furthermore, models 6 and 7 including demeaned screens do not perform substantially better than models 2–5.

Compared to the application of machine learning in Swiss cartel data by Huber and Imhof (2019), the correct classification rate for the Okinawa bid-rigging cartel is four to six percentage points higher. This suggests that the screens originally investigated in the Swiss context can also perform well in other countries with a different institutional setting like Japan. Table A3 in the appendix presents the relative weights and the mean squared errors in out-of-sample prediction, associated with each machine learner entering the ensemble method for the various models. As expected, the weight obtained in the ensemble is negatively correlated with its prediction error.

It is worth noting that the performance of the machine learners is not exactly the same across the subgroups of truly collusive and truly competitive tenders, see Table 6 and the differences in correct classification rates can amount to several percentage points. Considering for instance model 1, whose results are also provided in Table 7 by means of a confusion matrix, the ensemble method classifies 97.8% of the truly collusive tenders as collusive (known as true positive rate or sensitivity) and 96% of the truly competitive tenders as competitive (known as true negative

TABLE 7 Confusion matrix for model 1

	Random forest		Ensemble method	
	Coll. tenders	**Comp. tenders**	**Coll. tenders**	**Comp. tenders**
Predicted collusion	0.910	0.042	0.978	0.040
Predicted competition	0.090	0.958	0.022	0.960

Note: 'Comp. Tenders', 'Comp. Tenders' denote the collusive tenders and the competitive tenders, respectively.

TABLE 8 Predictor importance for the Japanese data

Model 1		**Model 2**		**Model 3**	
Predictor	**Gini index**	**Predictor**	**Gini index**	**Predictor**	**Gini index**
SPD	31.2	SPD	40.3	SPD	42.9
CV	28.9	CV	35.3	CV	41.4
NUMBID	28.2	KS	34.7	RD	19.5
KS	28	RD	13.4	RDNOR	13.3

Note: 'Gini index' denotes the decrease in the Gini index when dropping the respective predictor (see column 'Predictor'). 'KS', 'CV', 'SPD', 'RD', 'RDNOR' and 'NUMBID' denote the Kolmogorov–Smirnov statistic, the coefficient of variation, the spread, the relative distance, the normalized distance and the number of bids in a tender, respectively.

rate or specificity). Put differently, the false positive rate of incorrectly classifying competitive tenders as collusive amounts to only 4%, while the false negative rate of incorrectly classifying collusive tenders as competitive amounts to 2.2%. For the random forest, the differences are generally somewhat larger, with true positive and true negative rates of 91% and 95.8%, respectively, when for example considering model 1. While the random forest generally yields a higher correct classification rate among competitive tenders (i.e. fewer false positive classifications of competitive tenders), the ensemble method performs better among collusive tenders (with fewer false negative classifications of collusive tenders as competition). Since our sample contains more collusive tenders (246) than competitive ones (192), one might have suspected the machine learners to be slightly better in classifying collusive tenders due to this imbalance, as it is the case for the ensemble method, but not for the random forest.

Considering the random forest as machine learner, Table 8 reports the mean decrease in the Gini index when dropping each of the four best predictors in the respective model. The Gini index thus provides a ranking of the most important predictors within each model, but does not allow for a comparison between models since the index depends on the number of predictors included in each model. In models 1, 2 and 3, we observe that the SPD and the CV are the best predictors in the random forest. If available in the respective model, the number of bids and the KS follow in terms of predictive power. When excluding the number of bids as in model 2, the RD comes in fourth place, however, with a substantial drop in the Gini index. When additionally excluding the KS in model 3, the RDNOR comes in fourth place behind the RD. However, both predictors (RDNOR and RD) entail a lower Gini index than the SPD and the CV. The analysis of predictor importance in the random forest suggests that screens for the variance or uniformity of bids are most capable of distinguishing competitive and collusive tenders in the Okinawa bid-rigging cartel.

In order to examine the robustness of the correct classification rate with respect to omitting important predictors, we drop the CV and the SPD, which are the best predictors in models 1, 2

TABLE 9 Correct classification rates for the Japanese data when dropping the two best predictors

	Random forest			Ensemble method		
	All tenders	**Coll. tenders**	**Comp. tenders**	**All tenders**	**Coll. tenders**	**Comp. tenders**
Model 1	0.884	0.910	0.857	0.882	0.896	0.866
Model 2	0.880	0.897	0.860	0.901	0.905	0.898
Model 3	0.842	0.872	0.805	0.835	0.881	0.777

Note: 'All Tenders'. 'Coll. tenders' and 'Comp. tenders' denote all tenders, the collusive tenders and the competitive tenders, respectively. Model 1 includes the screens from model 2 as well as the number of bids and the contract value of a tender. Model 2 includes the screens discussed in Section 3.1; Model 3 the screens for the variance and the asymmetry of bids.

and 3. As shown in Table 9, the correct classification rate using model 1 decreases by 8.8 and 5.3 percentage points when using the ensemble method and the random forest, respectively. Model 1 still attains a correct classification rate as high as that of models 2, 3, 4 and 5 when including all predictors such that its performance remains very decent. Furthermore, discarding the CV and the SPD barely affects the performance of model 2, such that the correct classification rates of models 1 and 2 are quite similar in Table 9. In contrast, model 3 now entails a lower classification rate of 84.2% and 83.5%, corresponding to a performance decreases of 5.2 to 6 percentage points when using the random forest and the ensemble method, respectively, due to the omission of the two most important predictors. Except for the KURTO, model 3 then exclusively relies on screens for the asymmetry of bids. This suggests that asymmetry-based screens perform somewhat worse than screens based on the variance or uniformity of bids for dismantling the Okinawa bid-rigging cartel.

4.3 | Training and testing in distinct country data

We consider the transnational transferability of trained models by using one country as training sample and the other as test sample. Table 10 reports the results when training the models in the Swiss data of Huber and Imhof (2019) and testing them in the Japanese data. For models 2 to 5, the correct classification rates vary from 81.7% to 87.3% for the ensemble method, which clearly dominate those of the random forest ranging from 58.2% to 62.2%. Concerning the optimal combination of algorithms in the ensemble method, Table A4 in the Appendix indicates that even though the random forest generally obtains a substantial weight across the models considered, it is always complemented by further algorithms like SVM, whose weight is important in all models. This greatly improves the performance compared to using the random forest only. However, we observe a significant imbalance in the predictive quality across truly collusive or truly competitive tenders for both the ensemble method and the random forest. For the ensemble method, true positive rates are 13–20 percentage points higher than true negative rates for models 2–4. We, however, find a more balanced prediction quality for model 5, which involves only screens for asymmetry and uniformity of the bids. The random forest predicts competitive tenders better than the collusive ones. In fact, the true negative rate of the random forest is even below 50%, such that the ensemble method appears substantially more attractive when training in one country to testing in the other.

The balanced classification rates for the ensemble method in model 5 point to potential issues with the variance screens, related to the fact that they are not normalized and could for this reason

TABLE 10 Correct classification rates when training in Swiss and testing in Japanese data

	Random forest			Ensemble method		
	All tenders	**Coll. tenders**	**Comp. tenders**	**All tenders**	**Coll. tenders**	**Comp. tenders**
Model 2	0.611	0.471	0.738	0.873	0.931	0.799
Model 3	0.622	0.471	0.762	0.819	0.910	0.703
Model 4	0.582	0.476	0.669	0.817	0.910	0.698
Model 5	0.597	0.405	0.789	0.866	0.858	0.870
Model 6	0.638	0.414	0.922	0.899	0.902	0.896
Model 7	0.751	0.745	0.751	0.895	0.910	0.875

Note: 'All Tenders'. 'Coll. tenders' and 'Comp. tenders' denote all tenders, the collusive tenders and the competitive tenders, respectively. Model 1 includes the screens from model 2 as well as the number of bids and the contract value of a tender. Model 2 includes the screens discussed in Section 3.1; Model 3 the screens for the variance and the asymmetry of bids; Model 4 the screens for the variance and the uniformity of bids; Model 5 the screens for the asymmetry and the uniformity of bids. Model 6 includes all the demeaned screens from model 2. Model 7 combines the screens from models 2 and 6.

have relatively similar distribution across collusive and competitive subsamples in different countries. Indeed, the CV has a mean and a median in collusive tenders amounting to 1.16 and 0.44 in Japan as well as 3.42 and 2.97 in Switzerland. The respective statistics under competition are 4.59 and 5.09 in Japan as well as 8.05 and 7.16 in Switzerland. Therefore, the CV in collusive Swiss tenders tends to be higher than the CV in collusive Japanese tenders and more comparable to the competitive Japanese tenders, which creates issues for discriminating between collusion and competition across countries. As shown in Figure 4, we observe a bimodal distribution of the CV in the competitive tenders in Okinawa, while the peak of its density in the collusive Swiss tenders is located in-between. For this reason, if we train in the Swiss data for testing in the Okinawa data, a predictive method based on the CV recognizes almost all collusive Okinawa tenders as collusive since their CVs are mostly either equal or below the CVs of the collusive Swiss tenders. However, it unsatisfactorily recognizes competitive tenders in Okinawa since their CVs fall into the range of collusive tenders in Switzerland. We find similar issues for the spread (SPD) and the kurtosis statistic (KURTO), see the respective graphs provided in the Appendix (See Figures A1 to A8).

Therefore, the institutional difference between countries might affect the values of the screens and their support. In order to attenuate such issues due to institutional differences, we in a next step centre every screen entering model 2 within countries by subtracting the mean of the respective screen. For the ensemble method, we now obtain higher correct classification rates than before, amounting to 89.9% and 89.5% for models 6 and 7, respectively, without important imbalances across truly collusive and truly competitive tenders. Therefore, roughly nine tenders out of ten are correctly predicted when training in the Swiss data to test in the Japanese data and when including demeaned screens. In contrast to the ensemble learner, the performance of the random forest does not substantially improve under model 6 when compared to model 2. However, combining the demeaned screens of model 6 with the original ones of model 2 boosts the correct classification rate of the random forest to 75.1%, with little differences across truly collusive and competitive tenders. Yet, the performance is substantially worse than that of the ensemble learner.

Also when training models in the Japanese data to test the out-of-sample performance in the Swiss data, we find that the ensemble method outperforms the random forest by at least 20 percentage points for models 2–5 and attains a correct prediction rate varying from 83.7% to 86.1%

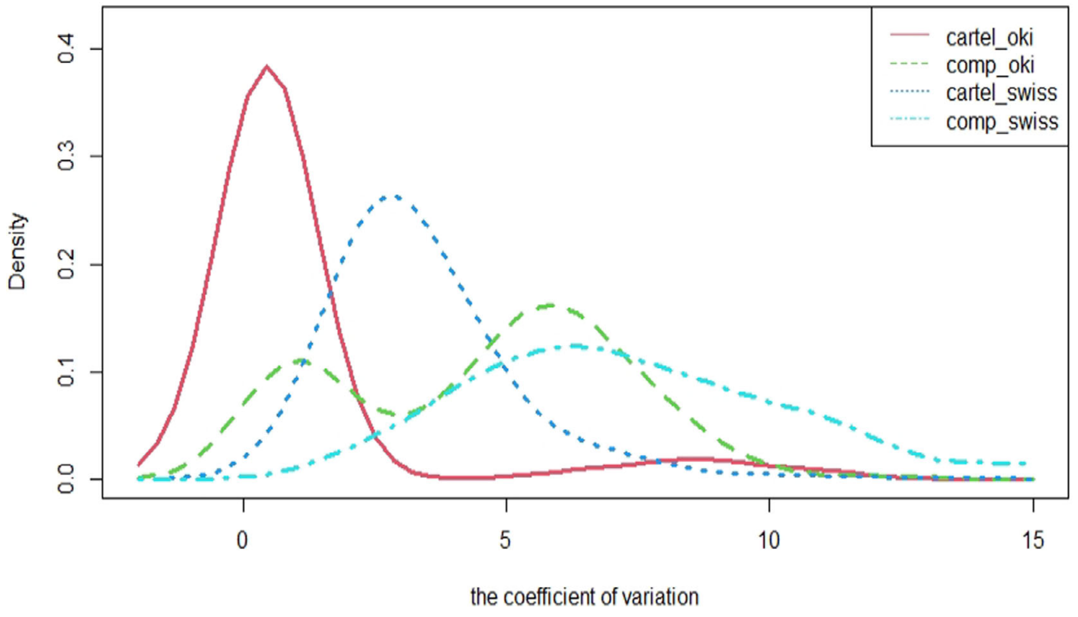

FIGURE 4 Density of the coefficient of variation by country [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 11 Correct classification rates when training in the Japanese and testing in the Swiss data

	Random forest			Ensemble method		
	All tenders	**Coll. tenders**	**Comp. tenders**	**All tenders**	**Coll. tenders**	**Comp. tenders**
Model 2	0.596	0.422	0.762	0.856	0.839	0.874
Model 3	0.608	0.422	0.787	0.861	0.829	0.895
Model 4	0.569	0.427	0.696	0.852	0.833	0.872
Model 5	0.582	0.348	0.821	0.837	0.786	0.891
Model 6	0.798	0.736	0.863	0.854	0.853	0.856
Model 7	0.724	0.508	0.951	0.851	0.856	0.846

Note: 'All Tenders'. 'Coll. tenders' and 'Comp. tenders' denote all tenders, the collusive tenders and the competitive tenders, respectively. Model 1 includes the screens from model 2 as well as the number of bids and the contract value of a tender. Model 2 includes the screens discussed in Section 3.1; Model 3 the screens for the variance and the asymmetry of bids; Model 4 the screens for the variance and the uniformity of bids; Model 5 the screens for the asymmetry and the uniformity of bids. Model 6 includes all the demeaned screens from model 2. Model 7 combines the screens from models 2 and 6.

as displayed in Table 11. The ensemble method predicts competitive tenders slightly better than collusive ones under models 2–4, while the imbalance increases in model 5. Table A5 in the Appendix provides the weights and mean squared errors of the various algorithms and shows that at least three machine learners enter the ensemble in any of the models. When considering the random forest alone, its performance is again unacceptably low with a correct prediction rate of only around 60% and strong imbalance across collusive and competitive tenders. This result can most likely be explained by the previously mentioned issues with variance screens. While the CV, for instance, permits recognizing most of the Swiss competitive tenders as competitive, it is not suitable for recognizing Swiss collusive tenders, whose values are relatively similar to those for Japanese competitive tenders that serve as training data.

When introducing the demeaned screens in models 6 and 7, the overall correct classification rate of the ensemble method is not improved when compared to model 2, however, the performance is now more balanced across collusive and competitive tenders than before. Regarding the random forest, the use of demeaned screens in model 6 boosts the correct prediction rate to 79.8%, but important imbalances across collusive and competitive tenders remain. In model 7, the correct prediction rate even falls to 72.4% while the imbalance increases. Yet, the random forest performs substantially worse than the ensemble method for any model considered.

In conclusion, when training in one country to test in the other, the ensemble method again correctly classifies more than eight out of ten tenders and clearly dominates the random forest, whose performance falls to unacceptably low levels. The better predictive power of the ensemble method is likely driven by its increased flexibility in approximating the statistical association between the screens and collusion. This flexibility is due to considering and combining several algorithms that are based on distinct statistical modelling approaches, where the combination (like a weighted average) can entail a better approximation than each individual algorithm per se, as for instance discussed in Dietterich (2000). However, an issue also arising for the ensemble method are the non-negligible imbalances in true positive and true negative rates, which occur to an even stronger extent with the random forest. This is related to the use of non-normalized predictors, which is problematic if for instance the levels of certain screens like the CV among Swiss collusive tenders fall into the range of values among Japanese competitive tenders.

Such a phenomenon indicates that institutional factors limit the transferability of a detection method from one country to the other. In our case, bidders have information about the cost estimates of engineers or the reserve price set by the procurement agency in the Okinawa tendering procedure, implying that cartel participants can rely on price information when rigging the tenders. Most of them bid just below the cost estimates or the reserve price when they collude, which truncates the distribution of bids and reduces its support as well as its variance. In contrast, Swiss bidders do not have access to cost estimates when trying to rig a contract such that the distribution of bids is not truncated. Unsurprisingly, the variance of collusive bids is therefore higher in Switzerland.

Such institutional differences suggest that the absolute level of such screens is not an appropriate statistic to be used for training in one country in order to make predictions in the other. Instead, it makes sense to consider relative measures of screens (like deviations from the mean) since bid rigging reduces the variance of the bids in both countries. In line with many other machine learning approaches relying on normalized predictors, we therefore demean the screens within countries. Indeed, the use of normalized screens somewhat increases the correct prediction rate of the ensemble method and reduces imbalances in the true positive and true negative rates. It also significantly improves the performance of the random forest in some cases, which, however, always remains below that of the ensemble method.[9]

[9]Even though we do not report these results in the paper, we consider standardized screens obtained by demeaning and dividing by the standard deviation of the respective screen within countries. Compared to demeaning alone, standardizing entails rather similar correct classification rates for the ensemble method when training in one and testing in the other country, while the performance of the random forest is generally somewhat reduced. Furthermore, for both algorithms, the imbalance in true positive and true negative rates tends to go up when using standardized screens.

4.4 The effect of bid rigging on screens

As a complement to our machine learning approach, we subsequently analyse how bid rigging affects particular screens in the Japanese and Swiss data in order to judge whether some screens are associated with cartel formation in a similar way in both countries. This would imply that insights drawn from such screens in one country could be transferred to another country for further screening activities. To this end, we estimate the following equation by linear regression in our mixed data set consisting of all in all 1022 Japanese and Swiss observations:

$$Screen = \beta_0 + \beta_1 Okinawa + \beta_2 Cartel + \beta_3 CarOki + \beta_4 Bids + \beta_5 NormalizedValue + \varepsilon. \quad (10)$$

Screen indicates a specific screen used as dependent variable, *Okinawa* and *Cartel* are dummy variables for observations from Okinawa and the incidence of a cartel, respectively, *CarOki* is an interaction term of *Okinawa* and *Cartel*, *Bids* is the number of bids in a tender and *NormalizedValue* is the normalized contract value within the observation's respective country. β_0 is the intercept, β_1 to β_5 denote the slope coefficients and ε is the error term.

Table 12 reports the coefficient estimates of β_2 and β_3 in (10), which yield the conditional effects of cartel formation on the respective screen in Switzerland, as well as the difference in the effect of cartel formation between Japan and Switzerland, respectively. Also the standard error (SE), the level of statistical significance (***, ** and *, for being significant at the 1%, 5% and 10% level, respectively) and the adjusted R^2 (Adj. R^2) are provided. All the coefficients on cartel are statistically significant at (least at) the 5%, suggesting that the effect of bid-rigging on any screen is non-zero in the Swiss data even when controlling for the number of bids and the normalized contract value.

Furthermore, some screens like the skewness (SKEW) or the ALTRD appear to be similarly affected in the Japanese data as in the Swiss sample, as suggested by the relatively modest and statistically insignificant coefficient on the interaction term (CarOki). Also for the spread (SPD), albeit statistically significant, the difference in the effect is rather small across countries. Even for the CV, one may argue that relative to the magnitude of baseline coefficient of −4.62 in the Swiss data, the effect is not too different in the Japanese data (−4.62 + 1.07 = −3.55). The results

TABLE 12 Effect of bid rigging on screens by countries

Screen	CV	KURTO	SKEW	SPD	RD	RDNOR	ALTRD	DIFFP	KS
Cartel	−4.62***	1.39***	−0.89***	−1.41***	1.95***	1.18***	3.32***	−0.91**	20.54***
SE	(0.28)	(0.17)	(0.08)	(0.01)	(0.18)	(0.08)	(0.59)	(0.31)	(1.26)
CarOki	1.07*	−1.39***	0.17	0.04**	−1.41*	0.70**	0.19	−0.88	163.61***
SE	(0.42)	(0.37)	(0.15)	(0.01)	(0.68)	(0.23)	(1.06)	(0.51)	(10.62)
Adj. R^2	0.42	0.1	0.22	0.35	0.07	0.24	0.07	0.13	0.60

Note: 'CV', 'KURTO', 'SKEW', 'D', 'RD', 'RDNOR', 'ALTRD', 'DIFFP', 'SPD' and 'KS' denote the coefficient of variation, the kurtosis, the skewness, the absolute difference between the two lowest bids in a tender, the relative distance, the normalized distance, the alternative distance, the percentage difference, the spread and the Kolmogorov–Smirnov statistic, respectively. The screens are linearly regressed on a constant, the number of bids, the normalized contract value within a country and dummies for Okinawa (base category is Switzerland), the incidence of a cartel (base category is competition) and an interaction between cartel and Okinawa. 'Cartel' and 'CarOki' provide the coefficients on the cartel dummy and the interaction, respectively. 'SE' and 'Adj. R^2' denote the standard error and the adjusted R^2, respectively. ***, ** and * denotes statistical significance at the 1%, 5% and 10% level, respectively.

therefore suggest that even if the levels of screens may differ between Japan and Switzerland, bid-rigging affects a subset of screens in a comparable way across countries, which may help for developing more generalizable screening methods. In contrast, some other screens like for instance the kurtosis (KURTO) change quite dramatically as a function of cartel formation from one country to another.

4.5 | Ex ante analysis: an example how to screen markets

In previous sections, we used information on collusive and competitive tenders for building models using random forests and the ensemble methods in the training data in order to evaluate their predictive performance in the test data. This obviously requires ex post information about whether tenders were collusive or competitive. In what follows, we perform a prediction in unseen data where the presence or absence of cartels is not clear, in order to see whether our method points to systematic collusion in that data.

We therefore perform an ex ante or screening analysis as in Imhof et al. (2018), just like a competition agency would do by applying a trained classification model to newly obtained data on tenders in order to assess the likelihood of cartel formation. Screening includes a more comprehensive set of analyses than merely using a predictive model. The latter constitutes only the first step, namely that of flagging conspicuous tenders, for instance in the analyses of Imhof et al. (2018). In a next step, a competition agency must conduct further investigations to substantiate the suspicions of collusion, for example by revealing a collusive logic among a specific set of contracts, bidders or regions that suggests that the conspicuous tenders do not occur randomly in the data. In the following, we outline how such a comprehensive investigation may proceed based on our data from Japan.

For training predictive models, we consider both the random forest and the ensemble method based on model 2, that is the specification using all screens, but not the scale-dependent characteristics NUMBID and VALUE. To maximize the sample size for training, we now use all the Japanese data with information on collusion and competition as training data, that is all observations either used in the training or test samples to obtain the results of Section 4.2. We then apply the trained models to the remaining observations from Japan that is all tenders of the post-inspection period as well as the pre-inspection and post-amendment periods without suspended bidders in the Okinawa data (and also including contracts of rank D), where the presence of collusion is ambiguous. Among these 950 tenders, we classify those as conspicuous (i.e. likely collusive) whose predicted probability of bid rigging is equal to or higher than 0.5 according to both the random forest and the ensemble method. 9.7% of the tenders are classified as collusive by one machine learner but not the other one, in which case a tender is classified as competitive. Therefore, the random forest and the ensemble method agree in 90.3% of the tenders.

Table 13 reports the contracts classified as conspicuous and competitive by type and period. For the pre-inspection period, 85%–93% of the contracts of ranks A, B, C and D are classified as conspicuous. In the post-inspection period, the share of tenders classified as competitive rises to 65% and 32% for contracts of ranks A+ and A, respectively, while it barely changes for the contracts of ranks B, C and D. In the post-amendment period, most of the tenders are classified as competitive, even though there is heterogeneity across contract types. 93%, 67% and 61% of the tenders among contracts of ranks A, B and D, respectively, are classified as competitive, but only 42% of those of rank C. This suggests that not all market participants have adapted their behaviour after the change in the OGP procurement system following the amendment of Japanese

TABLE 13 Tenders classified as conspicuous and competitive per contract type in the post-amendment period

		Pre-inspection		Post-inspection		Post-amendment	
	Tenders	**Freq.**	**Perc.**	**Freq.**	**Perc.**	**Freq.**	**Perc.**
Contract A+	Coll.	2	100%	6	35.3%	0	0%
	Comp.	0	0%	11	64.7%	6	100%
	All	2		17		6	
Contract A	Coll.	37	90.2%	45	68.2%	2	7.4%
	Comp.	4	9.8%	21	31.8%	25	92.6%
	All	41		66		27	
Contract B	Coll.	94	93.1%	62	82.7%	24	32.9%
	Comp.	7	6.9%	13	17.3%	49	67.1%
	All	101		75		73	
Contract C	Coll.	140	90.9%	85	92.4%	59	58.4%
	Comp.	14	9.1%	7	7.6%	42	41.6%
	All	154		92		101	
Contract D	Coll.	90	90.9%	34	85%	22	39.3%
	Comp.	9	9.1%	6	15%	34	60.7%
	All	99		40		56	

Note: 'All'. 'Coll.' and 'Comp.' denote all tenders, the collusive tenders and the competitive tenders, respectively.

competitive laws and the decision of the JFTC against the involved cartel participants. Among the bidders for contracts of rank C are apparently companies worth being scrutinized by an in-depth investigation concerning the possible incidence of bid rigging.

In the following analysis, we consider only contracts of rank B, C and D since we find a substantial share of contracts classified as conspicuous by the algorithms in the post-amendment period. We first search for geographical patterns that might point to a specific logic behind our results (see also Abrantes-Metz et al., 2006; Imhof et al., 2018, for geographical analyses in screening markets). Finding a more important share of conspicuous tenders in one or some regions could suggest the existence of an ongoing cartel that has not collapsed after the changes in the OPG procurement rules and the sanctions of the JFTC. For this reason, we calculate for each bidder the share of conspicuous contracts in which the bidder participates during the post-amendment period. Furthermore, we separately create for each rank B, C and D a map of Okinawa, depicting the location of each bidder, its bidding frequency and the previously computed share of conspicuous contracts. Each circle on the map corresponds to a bidder and the larger the circle is, the more frequently participates the respective bidder in tenders in the post-amendment period. Moreover, the colour of the circle indicates the share of conspicuous contracts for each bidder.

Figure 5 provides the map for bidders of rank B, which suggest that bidders participating in conspicuous tenders are relatively more frequently situated in the North. In contrast, bidders located in the South mostly participate in tenders that are classified as competitive by our algorithms and often also tend to bid less frequently (as indicated by the size of the circles). However,

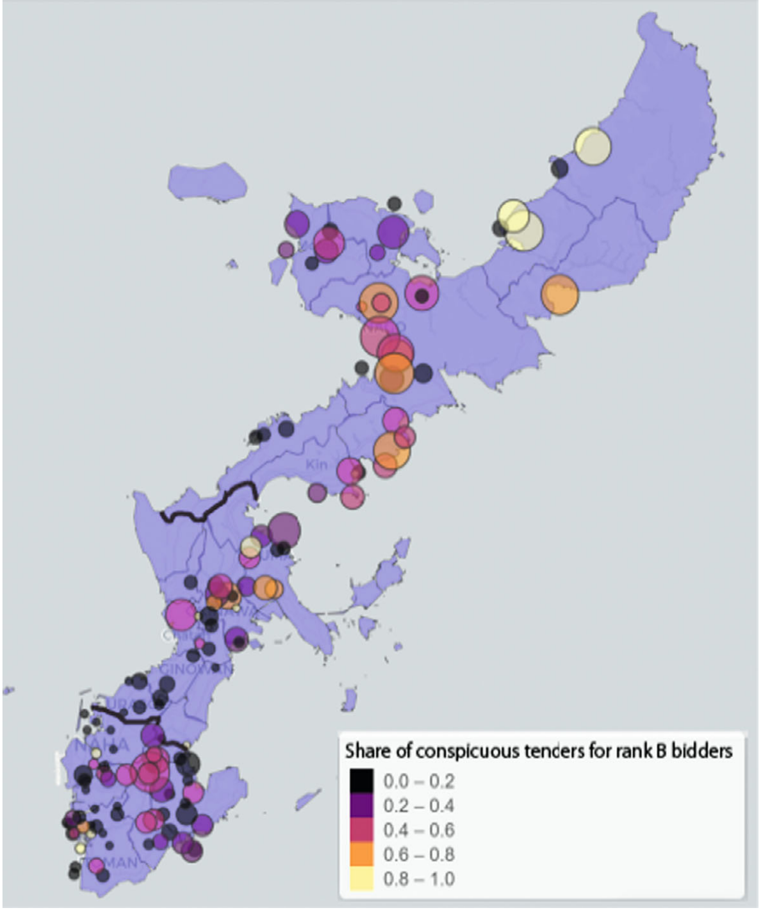

FIGURE 5 Share of conspicuous tenders for rank B bidders in Okinawa Main Island [Colour figure can be viewed at wileyonlinelibrary.com]

frequent bidders in the South and the centre also exhibit a tendency of participating in conspicuous tenders, although this phenomenon is weaker than in the North. Turning to the analysis of the bidders of rank C provided in Figure 6, we find an even stronger difference between the North and the South than for rank B. The share of conspicuous tenders is above 40% for most of the rank C bidders in the North, while the South mostly consists less frequent bidders which rarely participate in conspicuous tenders according to the classification of our algorithms. The analysis of the rank D bidders in Figure 7 provides a qualitatively similar pattern, with more bidders with a high share of conspicuous tenders in the North than in the South. We also find that less frequent bidders tend to exhibit a smaller share of conspicuous tenders than frequent bidders do.

To sum up, the geographical pattern indicates that the North is more prone to collusion than the South and the centre. Moreover, frequent bidders in procurement markets appear more likely to collude, which is in line with some practical experiences and theoretical works (see Aoyagi, 2003; Motta, 2004; Harrington, 2006).[10]

[10]Motta (2004) notes on page 145: *Regular orders facilitate collusion. [...] The high frequency of the orders also helps collusion because it allows for a timely punishment. If orders arrive only with large time intervals between them, one has a higher incentive to deviate because the punishment will be started only much later in the future, and will accordingly be discounted.*

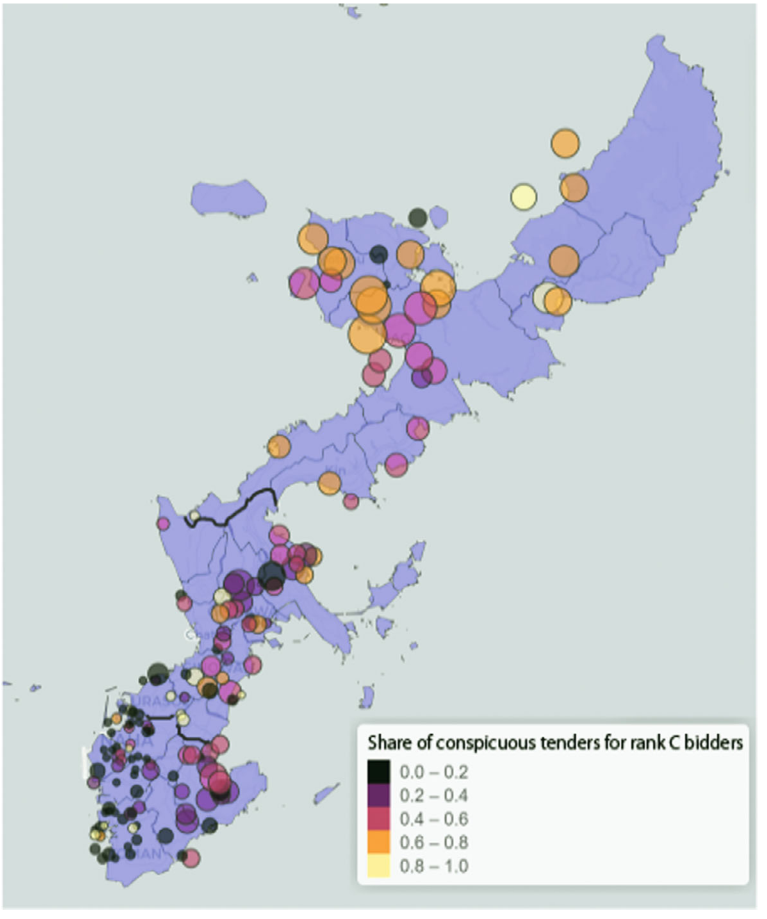

FIGURE 6 Share of conspicuous tenders for rank C bidders in Okinawa Main Island [Colour figure can be viewed at wileyonlinelibrary.com]

Following up on the observed geographical patterns in the post-amendment period, we in a next step investigate the evolution of the CV separately within the northern, central and southern regions over time jointly for contracts of ranks B, C and D. For the South and the centre, we observe in Figures 8 and 9, respectively, that all tenders without exception exhibit low values for the CV in the pre-inspection period and also mostly in the post-inspection period. In the post-amendment period, however, the CVs of many tender in the South and the centre are substantially higher. This substantial change in the dispersion of the CVs indicates that firms have importantly changed their bidding behaviour in the South and in the centre of Okinawa Main Island. The new procurement rules of the OGP, the change in competition laws in Japan and the sanctions of the JFTC against bidders of rank A+ involved in bid-rigging conspiracies have very likely contributed to this change. However, when taking a look at the evolution of the CVs in the North as provided in Figure 10, we find a different pattern. The CVs are relatively low for most tenders across all periods apart from relatively few exceptions and we do not observe striking changes in terms of the dispersion of CVs across periods. Therefore, firms in the North seem to bid in a similar manner in the post-amendment period as in the pre-inspection period, very much in contrast to bidders in other regions.

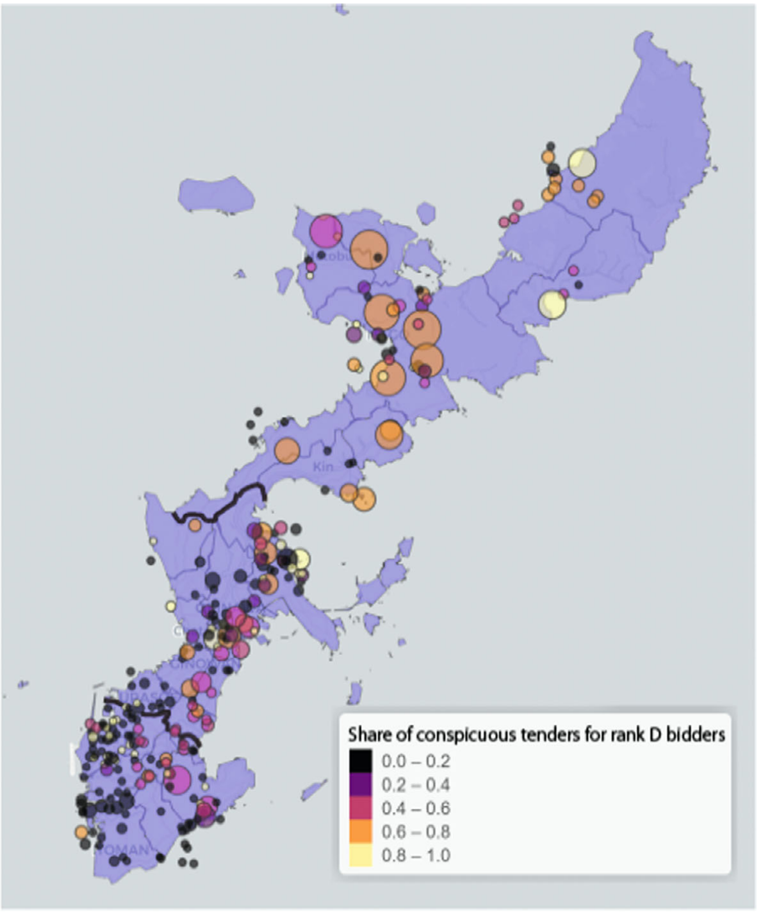

FIGURE 7 Share of conspicuous tenders for rank D bidders in Okinawa Main Island [Colour figure can be viewed at wileyonlinelibrary.com]

In a last step, we refine the analysis by normalizing all bids submitted in a tender by the reserve price. This implies that the bids are divided by the reserve price of a tender, a variable we did not make use of in the machine learning approach. In the presence of a cartel, prices tend to go up, which means that we should find higher normalized bids in the North compared to the centre or the South if the North is more prone to collusion. Moreover, if bid rigging does not decrease in the North over time, we should find little differences between the normalized bids of the pre-inspection and the post-amendment periods. We therefore apply MW and KS tests to investigate whether normalized bids statistically significantly differ across these periods. We distinguish between two categories of bidders when testing, namely less frequent bidders participating in one to five tenders in the post-amendment period and frequent bidders submitting more than five bids in the post-amendment period. This distinction is motivated by the fact that less frequent bidders tend to be relatively seldom involved in conspicuous tenders as depicted in Figures 8–10 for contracts B, C and D jointly.

The upper panel of Table 14 presents the results of the tests exclusively for rank B contracts. The MW and KS tests clearly reject the null hypothesis for all groups of bidders at any conventional level of statistical significance with the exception of frequent bidders in the North. For the latter group, the null hypothesis of equal normalized bids across periods is not rejected at the 1%

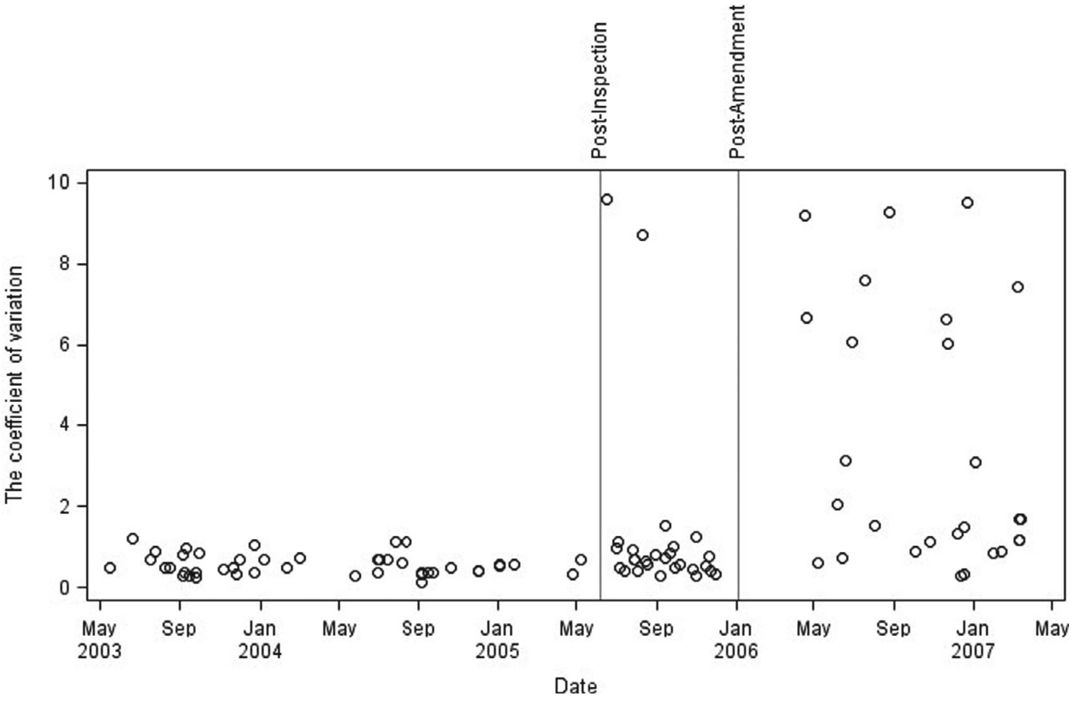

FIGURE 8 Coefficient of variation for contracts of ranks B, C and D in the South

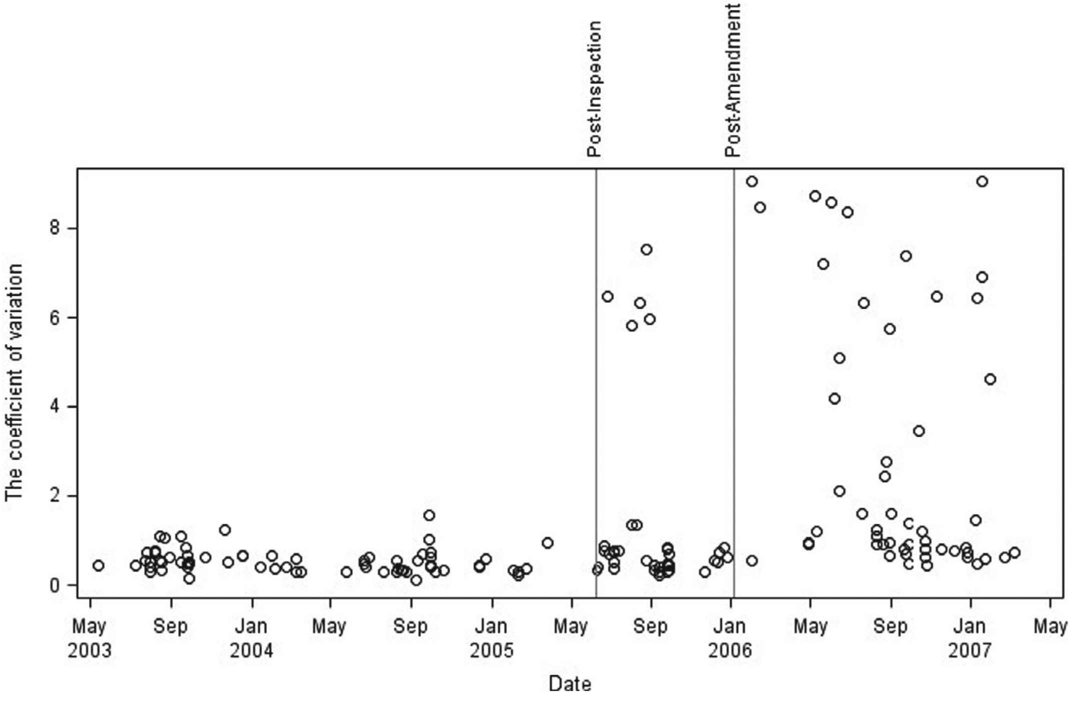

FIGURE 9 Coefficient of variation for contracts of ranks B, C and D in the centre

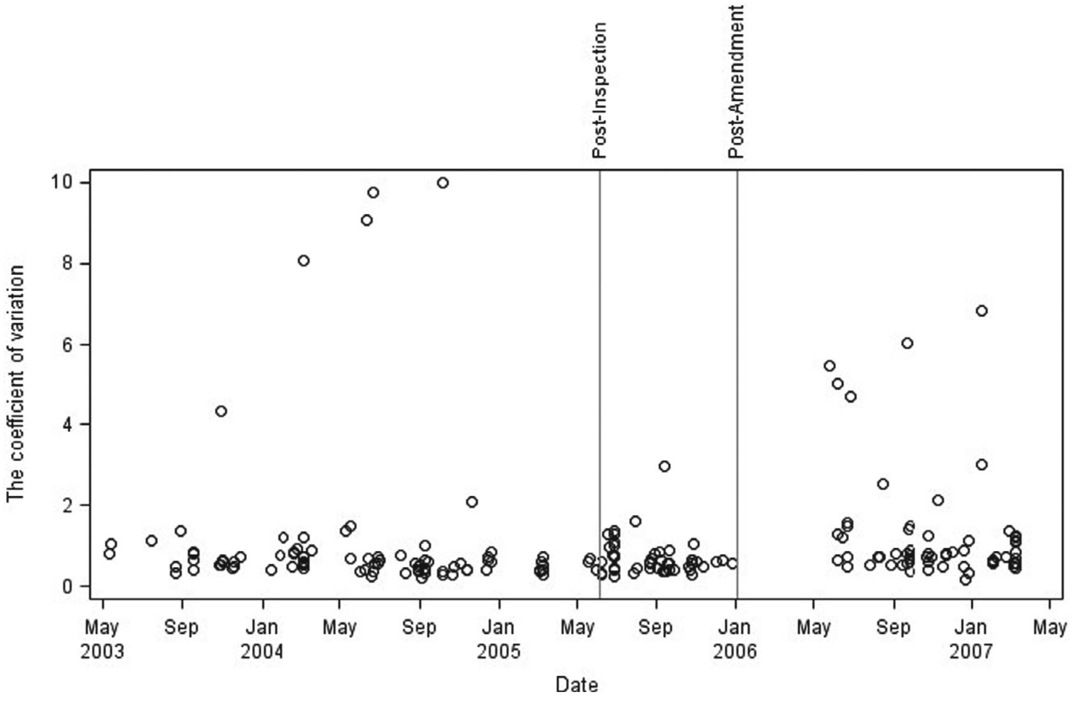

FIGURE 10 Coefficient of variation for contracts of ranks B, C and D in the North

level, as the *p*-value of either test amounts to 2%. Looking at changes in the average levels of normalized prices between the pre-inspection and the post-amendment periods across the various groups, we find them to be less important for frequent bidders, where the changes vary between 0.4 and 2.6 percentage points, than for less frequent bidders (4.9–6.2 percentage points).

For contracts C as presented in the middle panel of Table 14, the tests do not reject the null hypothesis for less frequent bidders in the north at any conventional significance level, while they do so for frequent bidders in the north. However, the differences in average normalized bids across periods are very small for both less frequent and frequent bidders in the North, namely 0.3 and 0.5 percentage points, respectively. The standard deviation does not change much either among bidders in the North pointing to a comparably stable bidding pattern across periods. In the south and the centre, however, we observe statistically significant and economically more meaningful changes of 1.6–3.2 percentage points in the average normalized bids. In general, the differences in contracts of rank C are smaller than for those of rank B, which is in line with the fact that more tenders of rank C are classified as conspicuous by our algorithms than those of rank B, see the results in Table 13. A potential explanation is that rank B bidders are more likely to participate in tenders in which rank A and A+ bidders are present who have been sanctioned by the JFTC and are for this reason reluctant to engage in collusive behaviour. For contracts of rank D as considered in the lower panel of Table 14, the tests do not reject the null hypothesis at the 5% level for (less and more frequent) bidders in the North and frequent bidders in the centre, but for bidders in the South and less frequent bidders in the centre. The statistically significant differences vary from 2% to 4.1%, which is smaller than for contracts of rank B.

Summing up, we find a specific geographical pattern for conspicuous tenders suggesting that bidders in the North are more prone to collusion. We also observe that frequent bidders are more likely to participate in conspicuous tenders, which fits theoretical findings. Moreover, the

TABLE 14 Normalized bids (scaled by a tender's reserve price) across periods

	Freq.	Statistical tests MW	*p*-value MW	KS	*p*-value KS	Post-amendment period Mean	Std	N	Pre-inspection period Mean	Std	N	Diff.
Bidders in tenders of rank B												
South	high	5.44	0	2.95	0	0.968	0.04	68	0.993	0.004	21	2.6
	low	7.53	0	4.06	0	0.946	0.07	76	0.993	0.01	68	5.0
Middle	high	7.02	0	3.73	0	0.971	0.04	97	0.995	0.01	42	2.5
	low	−8.91	0	4.82	0	0.936	0.07	70	0.994	0.01	82	6.2
North	high	2.27	0.02	1.49	0.02	0.984	0.03	244	0.988	0.02	210	0.4
	low	−3.45	0	3.06	0	0.932	0.08	87	0.986	0.03	144	5.8
Bidders in tenders of rank C												
South	high	7.16	0	3.31	0	0.966	0.05	75	0.991	0.01	73	2.6
	low	−7.27	0	3.88	0	0.961	0.06	105	0.992	0.01	231	3.2
Middle	high	10.22	0	4.98	0	0.971	0.04	235	0.992	0.02	212	2.2
	low	−6.05	0	3.31	0	0.972	0.04	78	0.988	0.03	155	1.6
North	high	3.43	0.001	1.89	0	0.984	0.03	360	0.989	0.02	316	0.5
	low	−0.5	0.62	0.66	0.77	0.987	0.01	12	0.99	0.01	25	0.3
Bidders in tenders of rank D												
South	high	2.74	0.01	1.62	0.01	0.976	0.01	17	0.996	0.01	5	2.0
	low	−4.45	0	2.55	0	0.964	0.05	80	0.991	0.01	94	2.8
Middle	high	1.83	0.07	1.2	0.11	0.98	0.03	97	0.99	0.01	42	1.0
	low	−4.33	0	3.12	0	0.952	0.07	122	0.991	0.02	146	4.1
North	high	0.57	0.57	0.81	0.11	0.985	0.01	98	0.986	0.01	17	0.1
	low	0.73	0.47	0.84	0.48	0.986	0.02	93	0.987	0.01	105	0.1

Note: 'Freq.' indicates if a bidder bids with low or high frequency. 'z-statistic' and '*p*-value MW' denote the z-statistic of the Mann–Whitney test and the *p*-value of the Mann–Whitney test, respectively. 'KS' and '*p*-value KS' denote the Kolmogorov–Smirnov statistic and the *p*-value of the Kolmogorov–Smirnov test, respectively. 'Mean', 'Std' and 'N' denote the mean, standard deviation and number of observations, respectively, in the post-amendment and pre-inspection periods. 'Diff.' indicates the average difference in normalized bids across the post-amendment and pre-inspection periods measured in percentage points.

evolution of the CV by regions over time suggests that the bidding behaviour of firms has changed importantly in the South and in the centre of Okinawa while remaining more stable in the North, even after the sanctions of the JFTC. Finally, we analyse the normalized bids scaled by the reserve price between the pre-inspection and the post-amendment periods and find that their distribution to be comparably stable for bidders in the North. Even when statistical tests reject the null hypothesis of equal distributions, the differences in the mean and standard deviation of normalized bids in the North are relatively small when compared to the South and the centre. All these findings indicate that collusive behaviour might have continued in the North and/or among frequent bidders even after the OPG changed its procurement rules and the JFTC sanctioned bidders of rank A+ involved in bid rigging.

5 | CONCLUSION

In this paper, we investigated the transnational transferability of machine learning methods based on statistical screens of bid distributions in tenders for detecting bid-rigging cartels in the construction sector. First, we illustrated that machine learning approaches originally considered in Swiss data perform very well in Japanese data, when using the latter to both train and test predictive models for classifying tenders as collusive or competitive. Depending on the model (i.e. the number of screens used as predictors), the correct out-of-sample classification rates were between 88% to 97% when using a random forest or an ensemble method in data from the so-called Okinawa cartel. This is an encouraging result with respect to the usefulness of such methods for tracking the statistical patterns produced by bid rigging also in other countries.

When training the machine learners based on data of only one country for testing in the other country, the out-of-sample performance deteriorated for the random forest, while it remained substantially higher for the ensemble method, which is based on an optimally weighted combination of various algorithms. Furthermore, we observed a non-negligible imbalance in the correct classification rates across collusive and competitive tenders in particular for the random forest, but to a lesser extent also for the ensemble method. This result was driven by some statistical screens having values that were relatively similar across collusive and competitive tenders in different countries.

For example, the CV (the standard deviation over the mean of the bids in a tender) in collusive Swiss tenders was quite comparable to that of competitive Japanese tenders. Therefore, when training in the Japanese data for testing in the Swiss data, the CV was not helpful for recognizing Swiss collusive tenders. Our results thus suggest that a country's institutional context like the legal environment and the tendering process matters for the distribution of bids and for the statistical screens. One solution to mitigate such issues due to institutional differences across countries consists of normalizing, e.g. demeaning the screens within countries. Indeed, we found demeaning to importantly reduce the imbalance in the accuracy across collusive and competitive tenders in the case of the ensemble method, which generally clearly outperforms the random forest, even though also the performance of latter improved considerably in some scenarios due to demeaning. This is in line with the more general finding in the machine learning literature that the normalization of predictors may boost the prediction quality. Furthermore, we investigated if bid rigging similarly affected particular screens in Japan and Switzerland, which was the case for a subset of screens. Relying on such screens that are similarly affected across countries may be useful for developing screening strategies that work well in an international context.

Finally, we applied the models trained in the Japanese data with information on the incidence of collusion to predict bid rigging in other Japanese data without information on the presence of cartels that have been collected at a later point in time. In this screening exercise, which may serve as guiding example for competition agencies, we judged a tender to be conspicuous if both the random forest and the ensemble method classified the tender as collusive. In a second step, we verified the existence of specific geographical patterns of conspicuous tenders. We find that firms that are situated in the North of Okinawa and/or bid comparably frequently are more prone to participate in tenders classified as collusive by our algorithms. This finding is corroborated with the evolution of the CV in tenders by regions over time, suggesting that bidding in the South and the centre of Okinawa has changed toward a more competitive behaviour in the post-amendment period, i.e. after previous cartel participants have been sanctioned, while this appears to be much less the case in the North. As a further check, we analysed by regions the normalized bids scaled by the reserve price and found them to remain rather stable across periods for firms situated in the North and in particular those bidding relatively frequently, in contrast to other regions. All these findings suggest that collusive behaviour might have endured in the North, especially among frequent bidders, while decreasing in the centre and the South after the JFTC sanctioned cartel participants.

DISCLAIMER

All views contained in this paper are solely those of the authors and cannot be attributed to the Swiss Competition Commission, its Secretariat, the University of Fribourg, Unidistance (Switzerland) or Shiga University.

ORCID

Martin Huber https://orcid.org/0000-0002-8590-9402
David Imhof https://orcid.org/0000-0002-2315-6628

REFERENCES

Abrantes-Metz, R.M., Froeb, L.M., Geweke, J.F. & Taylor, C.T. (2006) A variance screen for collusion. *International Journal of Industrial Organization*, 24, 467–486.

Abrantes-Metz, R.M., Kraten, M., Metz, A.D. & Seow, G. (2012) Libor manipulation. *Journal of Banking and Finance*, 36, 136–150.

Aoyagi, M. (2003) Bid rotation and collusion in repeated auctions. *Journal of Economic Theory*, 112, 79–105.

Aryal, G. & Gabrielli, M.F. (2013) Testing for collusion in asymmetric first-price auctions. *International Journal of Industrial Organization*, 31, 26–35.

Asker, J. (2010) A study of the internal organization of a bidding cartel. *American Economic Review*, 100, 724–762.

Athey, S., Bagwell, K. & Sanchirico, C. (2004) Collusion and price rigid. *Review of Economic Studies*, 71, 317–349.

Bajari, P. & Ye, L. (2003) Deciding between competition and collusion. *The Review of Economics and Statistics*, 85, 971–989.

Baldwin, L.H., Marshall, R.C. & Richard, J.-F. (1997) Bidder collusion at forest service timber sales. *Journal of Political Economy*, 105, 657–699.

Banerji, A. & Meenakshi, J.V. (2004) Buyer collusion and efficiency of governement intervention in wheat markets in northern India: an asymmetric structural auctions analysis. *American Journal of Agricultural Economics*, 86, 236–253.

Bergman, M.A., Lundberg, J., Lundberg, S. & Stake, J.Y. (2019) Interactions across firms and bid rigging. *Review of Industrial Organization*, 56(1), 107–130.

Bolotova, Y., Connor, J.M. & Miller, D.J. (2008) The impact of collusion on price behavior: empirical results from two recent cases. *International Journal of Industrial Organization*, 26, 1290–1307.

Boser, B.E., Guyon, I.M. & Vapnik, V.N. (1992) A training algorithm for optimal margin classifiers. In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, Association for Computing Machinery, NY, USA: New York. pp. 144–152, Available from: https://doi.org.10.1145/130385.130401

Breiman, L. (1996) Bagging predictors. *Machine Learning*, 24(2), 123–140.

Breiman, L. (2001) Random forests. *Machine Learning*, 45(1), 5–32.

Breiman, L. & Cutler, A. (2018) Random forests for classification and regression. Available from: http://cran.r-project.org

Chassang, S., Kawai, K., Nakabayashi, J. & Ortner, J. (2020) Data driven antitrust: Theory and application to missing bids. *Working Paper*.

Chipman, H.A., George, E.I. & McCulloch, R.E. (2010) BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4, 266–298.

Chotibhongs, R. & Arditi, D. (2012a) Analysis of collusive bidding behavior. *Construction Management and Economics*, 30, 221–231.

Chotibhongs, R. & Arditi, D. (2012b) Detection of collusive behavior. *Journal of Construction Engineering and Management*, 138, 1251–1258.

Conley, T.G. & Decarolis, F. (2016) Detecting bidders groups in collusive auctions. *American Economic Journal: Microeconomics*, 8(2), 1–38.

Dietterich, T.G. (2000) Ensemble methods in machine learning. In: *Multiple Classifier Systems*, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1–15

Esposito, F. & Ferrero, M. (2006) Variance screens for detecting collusion: An application to two cartel cases in Italy. *Working Paper*.

Estrada, E. & Vasquez, S. (2013) Bid rigging in public procurement of generic drugs in Mexico. *Competition Policy International*.

Feinstein, J.S., Block, M.K. & Nold, F.C. (1985) Asymmetric information and collusive behavior in auction markets. *The American Economic Review*, 75, 441–460.

Friedman, J., Hastie, T. & Tibshirani, R. (2010) Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33, 1–22.

Froeb, L.M., Sibley, D.S., Doane, M.J. & Pinto, B.P. (2014) Screening for collusion as a problem of inference. *The Oxford Handbook of International Antitrust Economics*, 2, 523–553.

Harrington, J.E. (2006) How do carels operate? *Foundations and Trends in Microeconomics*, 2(1), 1–105.

Harrington, J.E. (2007) Behavioral screening and the detection of cartels. In: Ehlermann, C. & Atanasiu, I. (Eds.) *European competition law annual 2006: enforcement of prohibition of cartels*, Oxford: Hart.

Harrington, J.E. & Chen, J. (2006) Cartel pricing dynamics with cost variability and endogenous buyer detection. *International Journal of Industrial Organization*, 24, 1185–1212.

Ho, T.K. (1995) Random decision forests. *Proceedings of the 3rd International Conference on Document Analysis and Recognition Montreal*, pp. 278–282.

Hothorn, T. & Zeileis, A. (2015) Partykit: a modular toolkit for recursive partytioning in R. *Journal of Machine Learning Research*, 16, 3905–3909.

Huber, M. & Imhof, D. (2019) Machine learning with screens for detecting bid-rigging cartels. *International Journal of Industrial Organization*, 65, 277–301.

Hueschelrath, K. & Veith, T. (2014) Cartel detection in procurement markets. *Managerial and Decision Economics*, 35(6), 404–422.

Imhof, D. (2017) Econometric tests to detect bid-rigging cartels: Does it work? *Working Papers SES 483, Faculty of Economics and Social Sciences, University of Fribourg (Switzerland)*.

Imhof, D. (2019) Detecting bid-rigging cartels with descriptive statistics. *Journal of Competition Law and Economics*, 15(4), 427–467.

Imhof, D., Karagoek, Y. & Rutz, S. (2018) Screening for bid rigging, does it work? *Journal of Competition Law and Economics*, 14(2), 235–261.

Ishii, R. (2009) : Collusion in repeated procurement auction: a study of paving market in Japan. *International Journal of Industrial Organization*, 27, 137–144.

Ishii, R. (2014) Bid roundness under collusion in Japanese procurement auctions. *Review of Industrial Organization*, 44, 241–254.

Jakobsson, M. (2007) Bid rigging in Swedish procurement auctions. *Working Paper*.

Jimenez, J.L. & Perdiguero, J. (2012) Does rigidity of price hide collusion?. *Review of Industrial Organization*, 41, 223–248.

Kapelner, A. & Bleich, J. (2016) bartMachine: machine learning with Bayesian additive regression trees. *Journal of Statistical Software*, 70, 1–40.

Karatzoglou, A., Smola, A., Hornik, K. & Zeileis, A. (2004) kernlab - an S4 package for kernel methods in R. *Journal of Statistical Software*, 11, 1–20.

van der Laan, M., Polley, E.C. & Hubbard, A.E. (2008) Super learner. *Statistical Applications of Genetics and Molecular Biology*, 6(1), 1–21.

Lee, I. K. & Hahn, K. (2002) Bid-rigging in auctions for Korean public-works contracts and potential damage. *Review of Industrial Organization*, 21, 73–88.

Marshall, R.C. & Marx, L.M. (2007) Bidder collusion. *Journal of Economic Theory*, 133, 374–402.

McCulloch, W. & Pitts, W. (1943) A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5(4), 115–133.

Mena-Labarthe, C. (2012) Mexican experience in screens for bid rigging. *Antitrust Chronical*, 3, 1–8.

Motta, M. (2004) *Competition policy: theory and practice*. Cambridge: Cambridge University Press.

OECD. (2014) Roundtable on ex officio cartel investigations and the use of screens to detect cartels.
Pesendorfer, M. (2000) A study of collusion in first-price auction. *The Review of Economic Studies*, 67, 381–411.
Peters, A., Hothorn, T. & Lausen, B. (2002) ipred: improved predictors. *R News*, 2, 33–36.
Porter, R.H. & Zona, J.D. (1993) Detection of bid rigging in procurement auctions. *The Journal of Political Economy*, 101, 518–538.
Porter, R.H. & Zona, J.D. (1999) Ohio school milk markets: an analysis of bidding. *RAND Journal of Economics*, 30, 263–288.
Ragazzo, C.E.J. (2012) Screens in the gas retail market: the Brazilian experience. *Antitrust Chronical*, 3, 1–8.
Ripley, B.D. (1996) *Pattern recognition and neural networks*. Cambridge: Cambridge university press.
Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58, 267–288.
Venables, W.N. & Ripley, B.D. (2002) *Modern applied statistics with S*. 4th edn. Berlin: Springer.
Wallimann, H., Imhof, D. & Huber, M. (2020) Machine learning approach for flagging incomplete bidrigging cartels, *Working papers SES 513, Faculty of Economics and Social Sciences of Fribourg (Switzerland)*.
Zhou, Z.-H. (2012) *Ensemble methods: foundations and algorithms*. Boca Raton, FL, USA: Chapman & Hall/CRC.

How to cite this article: Huber, M., Imhof, D. & Ishii, R. (2022) Transnational machine learning with screens for flagging bid-rigging cartels. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 185(3), 1074–1114. Available from: https://doi.org/10.1111/rssa.12811

APPENDIX

Appendix Figures A1–A8 and Tables A1–A5

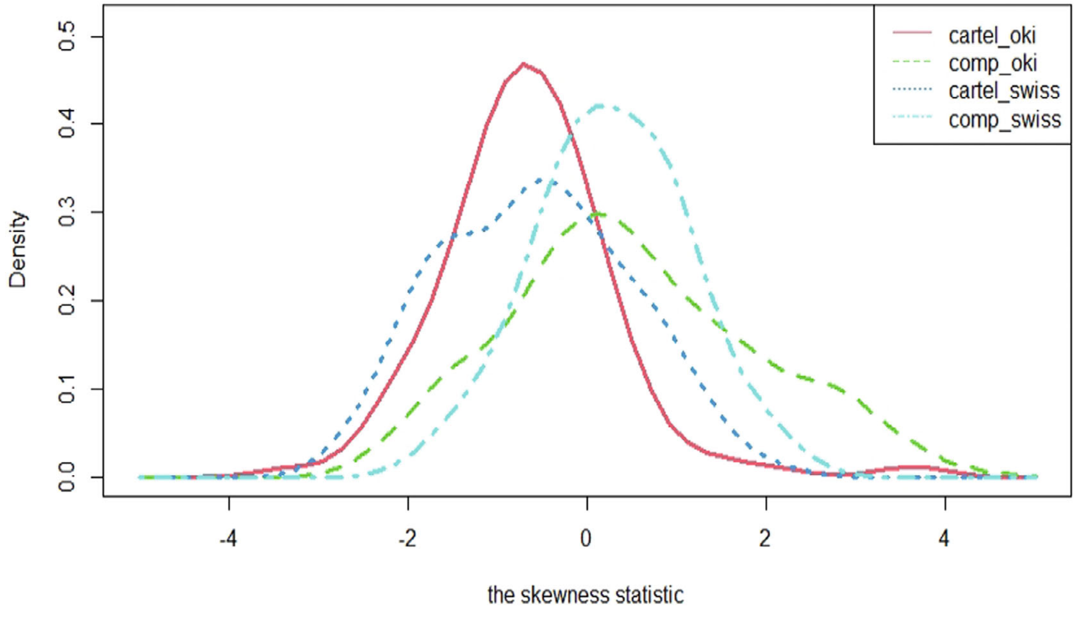

FIGURE A1 Density of the skewness by country [Colour figure can be viewed at wileyonlinelibrary.com]

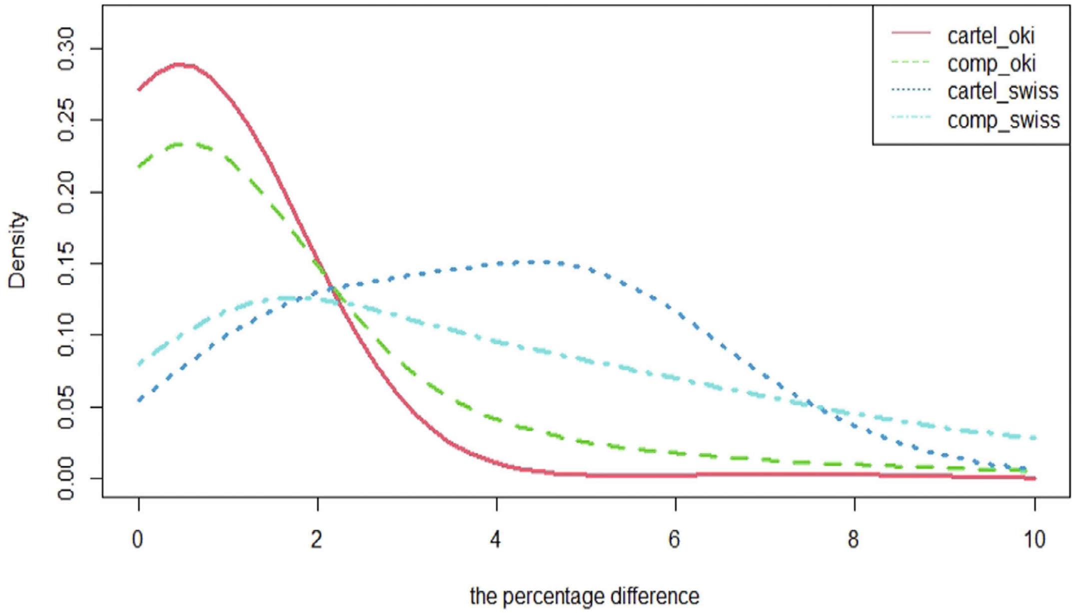

FIGURE A2 Density of the percentage difference by country [Colour figure can be viewed at wileyonlinelibrary.com]

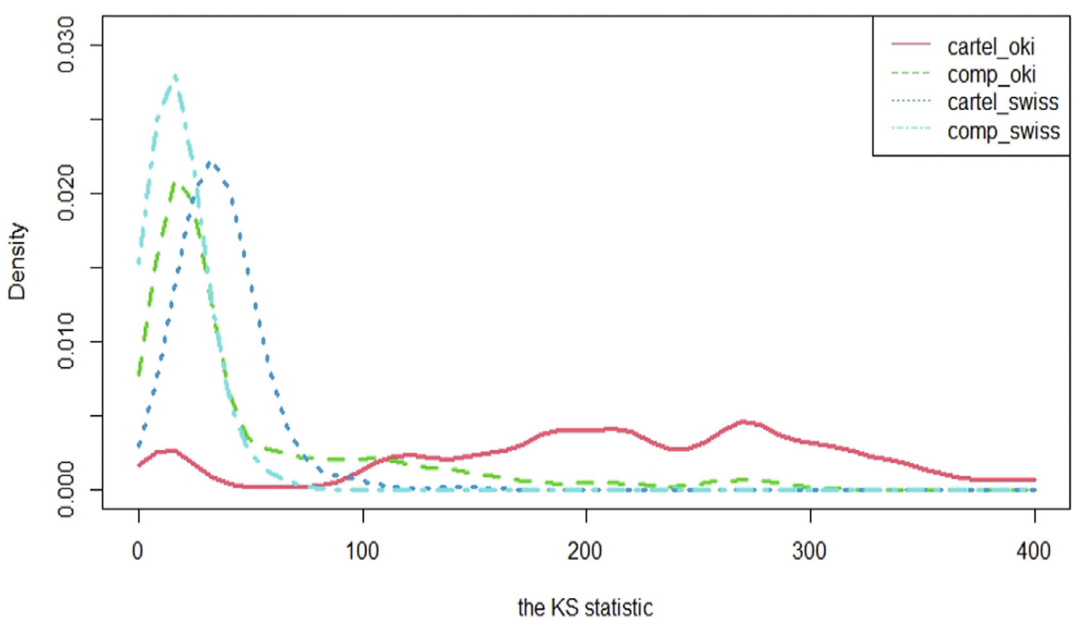

FIGURE A3 Density of the KS statistic by country [Colour figure can be viewed at wileyonlinelibrary.com]

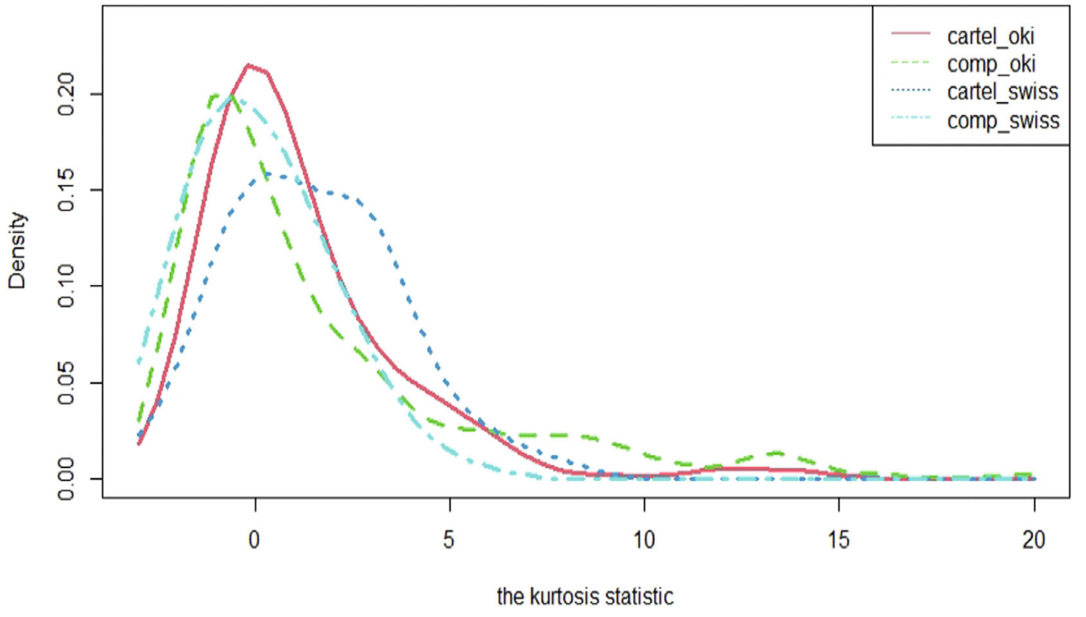

FIGURE A4 Density of the kurtosis statistic by country [Colour figure can be viewed at wileyonlinelibrary.com]

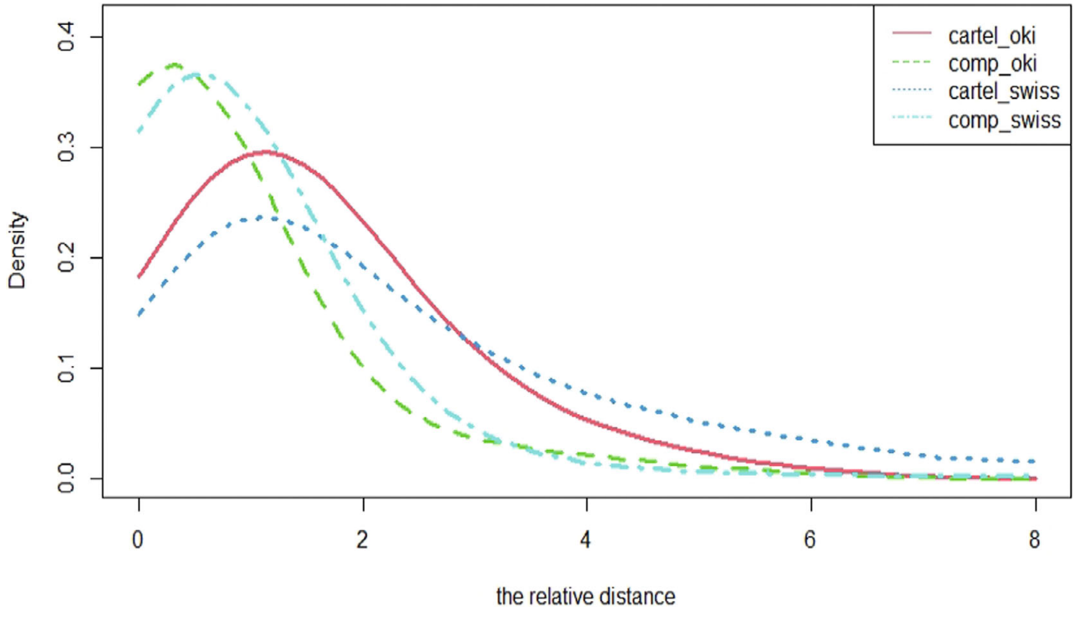

FIGURE A5 Density of the relative distance by country [Colour figure can be viewed at wileyonlinelibrary.com]

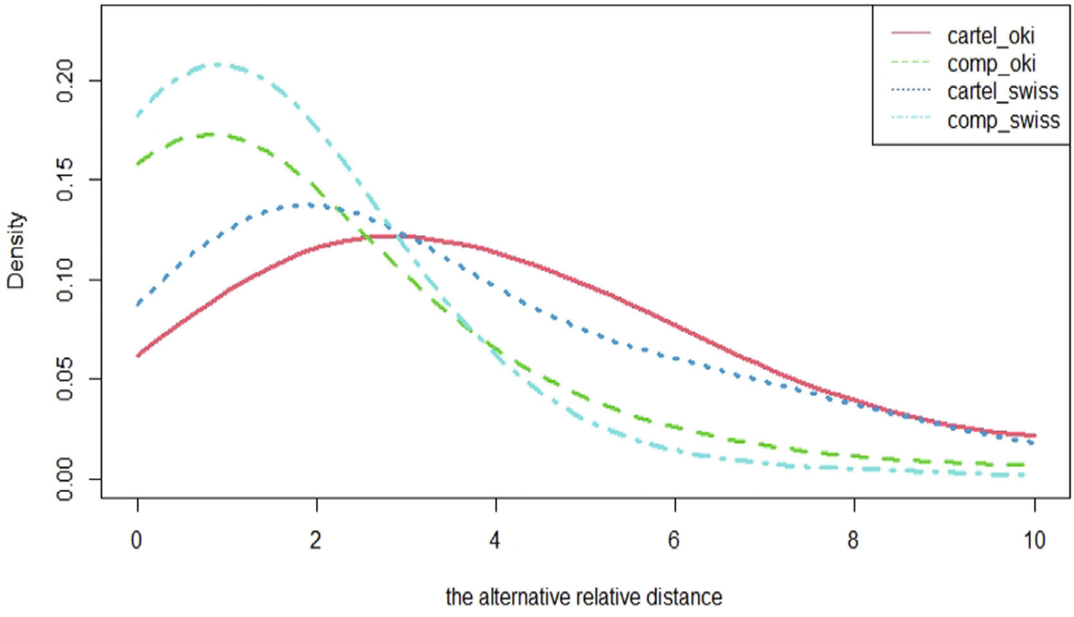

FIGURE A6 Density of the alternative distance by country [Colour figure can be viewed at wileyonlinelibrary.com]

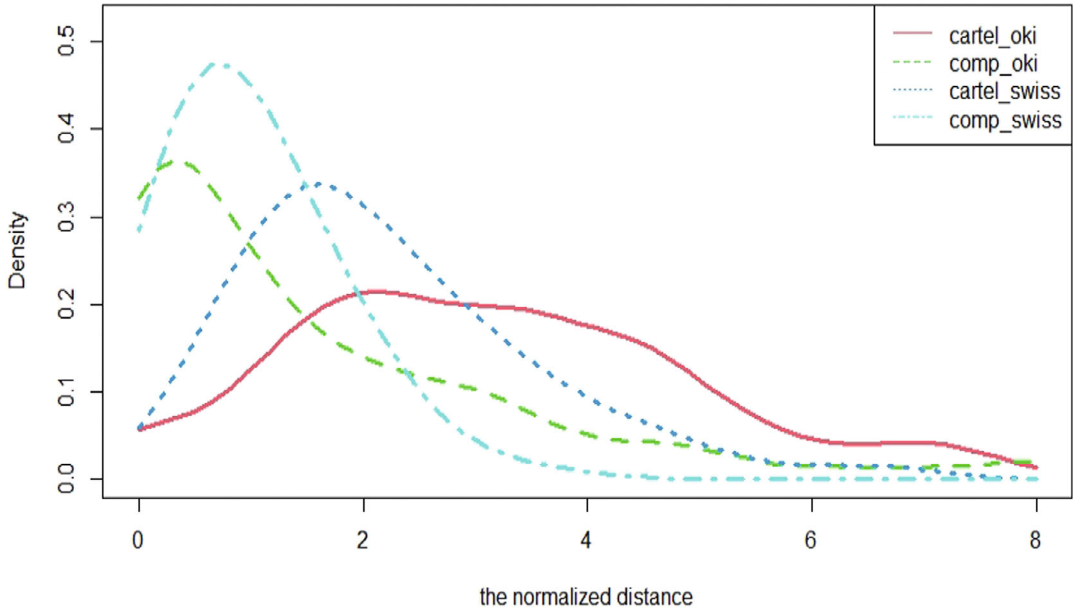

FIGURE A7 Density of the normalized distance by country [Colour figure can be viewed at wileyonlinelibrary.com]

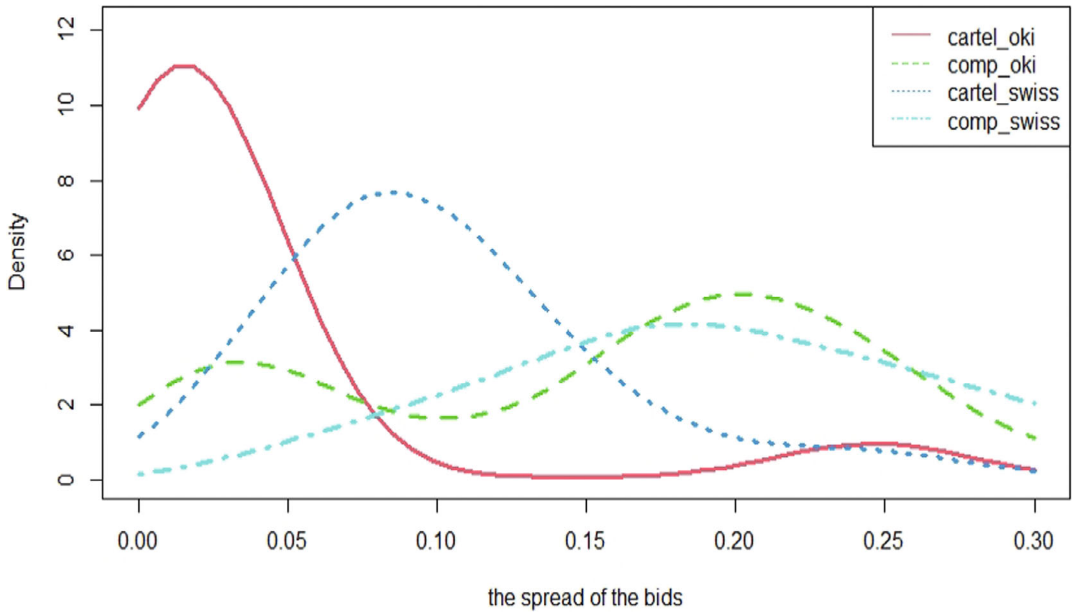

FIGURE A8 Density of the spread by country [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE A1 Descriptive statistics for predictors in collusive tenders

	Mean	Std	Min	Lower Q.	Median	Upper Q.	Max	Obs
NUMBID	11.37	1.83	5	10	10	13	16	246
VALUE	76.55	68.88	9.3	23.30	45.82	129.00	280.56	246
CV	1.16	2.35	0.10	0.33	0.44	0.60	11.24	246
KURTO	1.16	2.63	−2.25	−0.50	0.47	2.14	14	246
SKEW	−0.65	0.97	−3.48	−1.18	−0.69	−0.19	3.74	246
D	0.96	3.82	0.001	0.10	0.17	0.5	35.36	246
RD	2.18	5.24	0.0001	0.73	1.21	2.09	70.63	245
RDNOR	3.21	1.91	0.001	1.82	3	4.33	10.4	246
ALTRD	6.06	12.42	0.001	2	3.93	6	171.6	245
DIFFP	0.93	2.84	0.002	0.26	0.41	0.66	23.37	246
SPD	0.04	0.06	0.003	0.01	0.02	0.02	0.28	246
KS	243.46	143.46	9.37	167.44	227.39	300.85	1051.36	246

Note: ‘Mean’, ‘Std’, ‘Min’, ‘Lower Q.’, ‘Median’, ‘Upper Q.’, ‘Max’ and ‘N’ denote the mean, standard deviation, minimum, lower quartile, median, upper quartile, maximum and number of observations, respectively. ‘NUMBID’, ‘VALUE’, ‘CV’, ‘KURTO’, ‘SKEW’, ‘D’, ‘RD’, ‘RDNOR’, ‘ALTRD’, ‘DIFFP’, ‘SPD’ and ‘KS’ denote the number of bids in a tender, the contract value, the coefficient of variation, the kurtosis, the skewness, the absolute difference between the two lowest bids in a tender, the relative distance, the normalized distance, the alternative distance, the percentage difference, the spread and the Kolmogorov–Smirnov statistic, respectively.

TABLE A2 Descriptive statistics for predictors in competitive tenders

	Mean	Std	Min	Lower Q.	Median	Upper Q.	Max	Obs
NUMBID	16.19	3.34	5	14	16	19	23	192
VALUE	83.90	68.50	9.99	26.66	59.70	133.48	244.85	192
CV	4.59	2.67	0.35	1.70	5.09	6.48	12.53	192
KURTO	1.69	4.02	−1.99	−1.04	0.19	2.96	19.65	192
SKEW	0.50	1.41	−2.46	−0.49	0.33	1.44	4.36	192
D	1.30	4.62	0.003	0.07	0.22	1.003	56.34	192
RD	0.77	2.60	0.0004	0.04	0.22	0.72	33.92	192
RDNOR	1.73	2.11	0.002	0.15	0.89	2.56	10.76	192
ALTRD	2.62	5.08	0.002	0.14	0.87	2.88	48	192
DIFFP	2.06	5.62	0.002	0.17	0.52	1.51	53.54	192
SPD	0.17	0.12	0.01	0.06	0.18	0.23	1.11	192
KS	49.67	60.19	9.16	16.32	20.92	60.21	290.33	192

Note: 'Mean', 'Std', 'Min', 'Lower Q.', 'Median', 'Upper Q.', 'Max' and 'N' denote the mean, standard deviation, minimum, lower quartile, median, upper quartile, maximum and number of observations, respectively. 'NUMBID', 'VALUE', 'CV', 'KURTO', 'SKEW', 'D', 'RD', 'RDNOR', 'ALTRD', 'DIFFP', 'SPD' and 'KS' denote the number of bids in a tender, the contract value, the coefficient of variation, the kurtosis, the skewness, the absolute difference between the two lowest bids in a tender, the relative distance, the normalized distance, the alternative distance, the percentage difference, the spread and the Kolmogorov–Smirnov statistic, respectively.

TABLE A3 Weights and mean squared errors of methods in the ensemble (Japanese data)

		Random forest	Bagging	Lasso	BART	SVM	Neural net
Model 1	Weights	0	0	0.03	0.763	0.207	0
	MSE	0.06	0.058	0.053	0.03	0.044	0.247
Model 2	Weights	0.01	0.215	0.264	0.262	0.099	0.15
	MSE	0.092	0.092	0.093	0.089	0.096	0.136
Model 3	Weights	0	0.22	0.071	0.272	0.1	0.338
	MSE	0.095	0.092	0.126	0.089	0.1	0.103
Model 4	Weights	0.034	0.199	0.369	0.226	0.104	0.069
	MSE	0.091	0.092	0.091	0.089	0.095	0.143
Model 5	Weights	0.001	0.304	0.142	0.27	0.06	0.223
	MSE	0.1	0.1	0.101	0.093	0.1	0.128
Model 6	Weights	0.005	0.172	0.169	0.205	0.048	0.401
	MSE	0.092	0.092	0.093	0.89	0.096	0.092
Model 7	Weights	0.045	0.224	0.299	0.275	0.145	0.012
	MSE	0.091	0.092	0.093	0.089	0.096	0.149

Note: 'Weights' and 'MSE' denote the relative weights and mean squared errors, respectively, of the various machine learners that enter the ensemble method.

TABLE A4 Weights and mean squared errors (training in the Swiss and testing in the Japanese data)

		Random forest	Bagging	Lasso	BART	SVM	Neural net
Model 2	Weights	0.525	0	0	0.09	0.312	0.073
	MSE	0.284	0.063	0.07	0.116	0.24	0.126
Model 3	Weights	0.532	0	0	0.148	0.321	0
	MSE	0.136	0.139	0.156	0.136	0.14	0.171
Model 4	Weights	0.476	0	0	0.198	0.325	0
	MSE	0.135	0.139	0.155	0.136	0.14	0.16
Model 5	Weights	0.399	0	0	0	0.296	0.306
	MSE	0.123	0.129	0.143	0.126	0.124	0.133
Model 6	Weights	0.539	0	0	0	0.191	0.271
	MSE	0.109	0.115	0.137	0.111	0.112	0.12
Model 7	Weights	0.065	0	0	0.62	0.267	0.048
	MSE	0.109	0.114	0.125	0.108	0.112	0.151

Note: ‘Weights’ and ‘MSE’ denote the relative weights and mean squared errors, respectively, of the various machine learners that enter the ensemble method.

TABLE A5 Weights and mean squared errors (training in the Japanese and testing in the Swiss data)

		Random forest	Bagging	Lasso	BART	SVM	Neural net
Model 2	Weights	0.46	0	0	0.256	0.284	0
	MSE	0.122	0.127	0.142	0.123	0.127	0.164
Model 3	Weights	0.525	0	0	0.111	0.106	0.258
	MSE	0.122	0.126	0.144	0.123	0.129	0.14
Model 4	Weights	0.509	0	0	0.168	0.324	0
	MSE	0.136	0.14	0.156	0.136	0.14	0.163
Model 5	Weights	0.34	0	0	0	0.365	0.246
	MSE	0.123	0.129	0.143	0.126	0.124	0.14
Model 6	Weights	0.512	0	0	0	0.307	0.181
	MSE	0.11	0.115	0.136	0.112	0.111	0.125
Model 7	Weights	0.327	0.007	0	0.487	0.179	0
	MSE	0.108	0.113	0.124	0.107	0.112	0.155

Note: ‘Weights’ and ‘MSE’ denote the relative weights and mean squared errors, respectively, of the various machine learners that enter the ensemble method.