ISSN 1611-1699 / eISSN 2029-4433

VILNIUS TECH
Vilnius Gediminas Technical University

JOURNAL of BUSINESS ECONOMICS & MANAGEMENT

2026
Volume 27
Issue 2
Pages 285–302
https://doi.org/10.3846/jbem.2026.25748

FINDING PREDICTORS OF CORRUPTION FROM EUROPEAN FIRM LEVEL SURVEY DATA: A RANDOM FOREST APPROACH

Valentina VUČKOVIĆ [1]✉, Marko DRUŽIĆ [2]

Department of Economic Theory Faculty of Economics and Business, University of Zagreb, Zagreb, Croatia

Article History:
- received 21 January 2025
- accepted 13 November 2025

Abstract. Corruption remains a significant constraint for firms in Europe, despite ongoing institutional reforms. The main goal of this paper is to obtain a list of firm-level variables that can serve as predictors of corruption perception using a machine learning approach. Drawing on agency and institutional theory, we analyse firm-level data from European firms from the World Bank Enterprise Survey (WBES). We employ a Random Forest classifier, which is well-suited for high-dimensional, categorical survey data, capturing non-linear relationships and interactions often missed by traditional models. The model achieves strong predictive performance (ROC AUC = 0.755; Accuracy = 79%). Results show that the most important prediction factors of corruption perception include firm age, size, ownership concentration, legal form, external financial audits, bribery experiences, sector, country group (EU vs. WB), innovation activity, and informal sector competition. The findings support the design of risk-based audits and encourage reforms to reduce informality through streamlined registration processes. The study contributes methodologically by applying machine learning to the field of political economy and expands theoretical insights into firm-level institutional barriers. It is one of the first research to apply Random Forest to firm-level corruption perception in both EU and Western Balkans.

Keywords: corruption, Europe, institutions, firms, machine learning, Random Forest.

JEL Classification: O17, O52, D73, C5.

✉Corresponding author. E-mail: ***vvuckovic@net.efzg.hr***

1. Introduction

Despite ongoing policy efforts, corruption remains a significant barrier to economic and institutional development across Europe. Although the European Union (EU) is regarded as one of the least corrupt regions globally, there is significant variation among its individual member states, as well as among candidate countries, particularly those from the Western Balkans (WB). In the EU, issues with corruption control emerged after the fall of the Santer Commission, triggered by corruption allegations, as well as with the accession of new member states (NMS) that underwent incomplete transitions and faced economic challenges (Mungiu-Pippidi, 2013). The latter is why we also included WB in the analysis, as they are in the process of joining the EU and cooperate with the EU through various policies and initiatives, which can influence regulatory, economic, and anti-corruption reforms, making them relevant for analysis. The 2008 economic crisis worsened the situation, with research

confirming a direct link between the crisis and an increased perception of corruption within the EU due to weakened governance structures and the increased strain on public resources (Gugiu & Gugiu, 2016; Dimant & Tosato, 2018; Cieślik & Goczek, 2018).

According to the latest data, the estimated cost of corruption in the EU ranges between 179 and 990 billion EUR annually, accounting for up to 6% of its GDP (European Commission [EC], 2024). Also, 35% of EU businesses consider corruption as a problem in doing business, and even a larger share of them, i.e. 59%, agree that bribery and the use of connections is often the easiest way to obtain certain public services (according to 2023 Eurobarometer survey: Business' attitudes towards corruption (EC, 2023)). Moreover, corruption diminishes the quality of public services and diverts public resources away from critical projects (see various Transparency International reports). As a result, it remains a persistent issue in public procurement and the allocation of EU funds, further weakening governance quality (Dávid-Barrett & Fazekas, 2020; Dimant & Tosato, 2018). Although it can be concluded that corruption greatly impacts various aspects of life, there is still a research gap from the perspective of exploring factors that could predict it. This paper addresses this gap by identifying the key firm-level predictors of corruption perception using machine learning (ML) techniques on a large-scale survey data.

Our main research question (RQ) is: *Which individual firms' characteristics and groups of similar characteristics are most predictive of corruption in European businesses?* We approach this question by applying Random Forest (RF) models to the World Bank Enterprise Survey (WBES) data for both EU members and WB countries. Such analysis enables us to identify patterns of corruption across firms that vary by size, sector, ownership, financing, and other characteristics. However, the main limitation of the RF model is that it identifies associations but does not establish causal relationships between firm characteristics and corruption exposure. While RF models do not provide conventional measures of statistical significance typically found in causal econometric approaches, they offer important complementary value by uncovering complex, non-linear relationships and interactions among predictors. As such, they enhance the empirical approach by identifying key variables that can inform the development of theoretically grounded hypotheses for further econometric causal analysis. Finally, including WB countries is particularly relevant, given their EU accession trajectories and engagement in governance and anti-corruption reforms, making them a significant group alongside EU members.

The contribution of the paper is threefold. First, we apply RF models, which are well-suited for analysing survey data because they can handle complex, nonlinear relationships and interactions between variables without requiring strict assumptions that standard econometric models often rely on. RF also works well with categorical and missing data, and is robust enough to overfit due to its ensemble nature. On the other hand, compared to other ML techniques, RF has the advantage of requiring less tuning and is interpretable through feature importance scores. These make RF an ideal choice for uncovering patterns in diverse and noisy survey data. Second, the paper contributes to the three strands of literature, including research on corruption and governance, business environment and institutions, on firm-level and the application of ML in political economy. The obtained results contribute to discussions on anti-corruption and governance reforms. Finally, this is one of the first studies to apply Random Forest to firm-level corruption perception in both the EU and the WBs.

The paper is organised as follows. After the introduction, Section 2 reviews literature on corruption and the use of ML in this domain. Section 3 presents methodology and data, while Section 4 brings the discussion of results. Finally, Section 5 concludes.

2. Theoretical background

2.1. Theoretical perspectives and empirical research on the causes and effects of corruption

Although being analysed across different disciplines (economics, political science, sociology), one of the most widely recognised definitions is the one from the World Bank, according to which corruption is the misuse of public office for private gain (World Bank, 1997). Next, the OECD expands this definition by explaining corruption as the abuse of both public and private positions for personal gain (Organisation for Economic Co-operation and Development [OECD], 2008), which is particularly relevant in cases where private firms also engage in corrupt practices, such as offering bribes to secure contracts or bypass regulations.

The research in this paper is grounded in institutional theory (North, 1990) and in agency theory and principal-agent model, in which corruption emerges due to the asymmetric information and weak accountability between agents (public officials) and citizens (principals) (see Klitgaard, 1988; Tanzi, 1998; Rose-Ackerman, 2017). While agency theory explains how corruption happens at the firm level, institutional theory offers the broader structural context. It shows how formal and informal rules can enable or restrict the individual and collective behaviours. The analysis of corruption in European firms through the lens of these theories could reveal interesting findings, as corruption is embedded in organisational structures, incentive systems, and broader institutional environments. However, the models often fail to explain systemic forms of corruption such as state capture and business capture. Precisely, when examining corruption within the business sector, Bartlett (2023) describes two phenomena. First, state capture, where connections exist between political and business elites that may resort to bribery and corruption to sway public policy in their favour. The second phenomenon is business capture, in which the state exerts control over the business sector through regulation (Bartlett, 2023). These are particularly relevant for post-transition countries of Central and Eastern Europe (CEE) and WB, where incomplete reforms have resulted in a hybrid system of formal institutions and informal influence (Grzymala-Busse, 2007; Mungiu-Pippidi, 2013).

Nonetheless, regardless of its level, there is a prevailing consensus in the literature that the impact of corruption is detrimental, affecting individuals, businesses, and society as a whole (see, e.g. Enste & Heldman, 2018; Lambsdorff, 2006; Burke & Cooper, 2009; Graeff & Svendse, 2013; Tanzi, 1998; Rose-Ackerman, 2017). Specifically, corruption exerts adverse effects through various channels. Some of these include undermining public trust and democratic institutions (Transparency International, 2021); distorting resource allocation and slowing economic growth (D'Agostino et al., 2016; Huang, 2016; Glaeser & Saks, 2006); increasing income inequality or reducing access to public services (Glaeser et al., 2004; Uslaner, 2017; Gupta et al., 2002; Egger & Winner, 2005; Khan, 2022); hindering financial results of firms (Fisman et al., 2024; Audretsch et al., 2022; Shleifer & Vishny, 1993; Djankov et al., 2002; Campos et al., 2010; Belitski et al., 2021).

The other stream of growing literature, which is at the centre of our analysis, focuses on various causes and determinants of corruption. These are typically attributed to increased competition and a weak institutional framework where the dynamics of economic and political competition creates incentives and opportunities for corruption, while weak accountability structures within institutions increases its probability (Warner, 2011). Various analyses demonstrate the importance of firm-level characteristics in explaining corruption, as they can act as conduits, amplifiers, or inhibitors of corrupt behaviour. Although the results on this are

somewhat mixed, they mainly focus on firm size, ownership structure, sector, innovation activities etc. (e.g. Goel et al., 2021; Nguyen, 2020; Svensson, 2003; Jaggi et al., 2021; Cieślik & Goczek, 2022). At the business level, the diversion of public spending due to corruption can indirectly affect entrepreneurs by increasing costs within the broader economic context in which they operate. Additionally, entrepreneurs who cannot afford to pay bribes to obtain permits may choose to remain in the informal sector due to corruption, which, in turn, increases the shadow economy, recognised as another a driver of corruption (Audretsch et al., 2022). Corruption typically hinders firms from securing financing, which in turn further restricts their access to funds and increases the likelihood of informal payments. García-Gómez et al. (2025) demonstrate that firms encounter fewer obstacles as countries enhance their transparency and institutional quality. Although research on the causes of corruption is expanding, there remains a limited understanding of the relation between firm-level characteristics and corruption. Also, most analyses depend on national data or linear models. This paper aims to fill that gap by employing RF technique to estimate predictors of corruption perception among EU and WB firms, providing a fresh analytical and regional perspective.

2.2. Potential of machine learning techniques in assessing corruption at the business level

Machine learning (ML) is becoming an increasingly popular tool for detecting and predicting corruption, offering a range of models that can process large datasets and identify patterns of fraudulent behaviour (see, e.g., Lima & Delen, 2020; Poltoratskaia & Fazekas, 2024; Fazekas et al., 2022; Köbis et al., 2022; Doria et al., 2022; Colonnelli et al., 2022). ML models can predict corruption in the private and public sectors by analysing firm-level data and sectoral characteristics. For instance, firms that frequently engage in public procurement or have close ties with government officials are at a higher risk of public sector corruption (Saha & Gounder, 2013). The models can effectively identify collusive interactions between private companies and public authorities, as well as other corruption indicators. Precisely, in large firms, where corrupt activities often involve bribe payments to government officials at irregular intervals and uncovering evidence of corruption can be challenging, incorporating ML into anti-corruption efforts shows great promise (Rusch, 2021).

There is already research that uses ML in this area, especially concerning public procurement corruption and municipality-level corruption. For example, using data from Brazil, Colonnelli et al. (2022) demonstrate that various ML models achieve strong performance in predicting corruption at the municipal level concerning public spending. They identify private sector activity, financial development, and human capital as the most significant predictors of corruption, while factors related to the public sector and politics are secondary. Next, Decarolis and Giorgiantonio (2022) examine the relationship between various public procurement characteristics and corruption risks in Italian municipalities. The authors evaluate the predictive power of different indicators through LASSO, Ridge regression, random forest, and OLS methods. They emphasise that competition among private companies plays a key role in lowering corruption risk, and that the ability to easily access information on tenders and submit bids systematically relates to corruption. On the other hand, they point out that warning signs often do not relate to corruption or might even indicate the opposite. For instance, this includes situations where special procedures are started due to "urgency" or the degree of visibility surrounding tender announcements (Decarolis & Giorgiantonio, 2022). Further, Lima and Delen (2020) analyse corruption using contemporary ML

techniques to discover the most important corruption perception predictors and find that the random forest is the most accurate prediction model. Their results show that the most influential variables in defining the corruption level include government integrity, property rights, judicial effectiveness, and education index.

An ML analysis can help policymakers and economists identify key firm-level and regional factors that explain corruption risks. The WBES database, for example, offers rich, firm-specific data, which, when combined with ML techniques, could identify patterns that may not be obvious through traditional statistical methods. WBES asks a large, representative sample of firms about their experiences of corruption or about their perceptions of corruption (Fazekas & Ferrali, 2023). The results obtained from such analyses can help in developing anti-corruption measures by highlighting the areas where firms are more exposed to corrupt practices. Furthermore, the analysis can contribute to predicting future corruption trends, which would support more effective resource allocation and enforcement strategies. By analysing firm characteristics, ML can uncover corruption depending on their interactions with public institutions and market competitors. This interaction between public and private sector corruption is essential for designing effective anti-corruption strategies as public sector corruption tends to erode trust in government and weaken institutional integrity, whereas private sector corruption may distort market competition and deter investment (Saha & Gounder, 2013; Lambsdorff, 2007).

3. Description of methodology

3.1. Dependent and independent variables

Since corruption is inherently difficult to measure, various indirect methods are used in practice. In this paper, we focus on the micro and firm levels and use the latest World Bank Enterprise Survey (WBES) data for European countries (EU member states and WB countries that aspire to become EU members). The WBES asks firms questions concerning corruption from different angles. First is the firms' participation rates in public procurement, which is estimated using the WBES data, where firms are asked whether they have secured or attempted to secure a government contract in the 12 months prior to the interview. According to the EC (2024), public procurement is among the activities most susceptible to corruption within government operations. This is primarily due to several factors, including the large volume and numerous transactions involving public funds, the intricate and opaque nature of the procurement process, the close relationships between public officials and businesses bidding for contracts, and the various stakeholders involved in the process (EC, 2024). Second, firms are also asked the question on the amount of bribes that firms like themselves pay to public officials to "get things done", which is a proxy of the self-reported incidence rate of bribery in public procurement (World Bank, 2023). Third, the WBES provides data on petty corruption, which refers to the corrupt practices that businesses encounter when seeking public services, licenses, and permits, including electricity and water connections, construction permits, import licenses, operating licenses, and interacting with tax officials through inspections or meetings. The measure of corruption in this case is captured by an incidence of petty corruption, represented as a binary variable, which is marked as 1 if a firm has faced a bribe payment or request related to any of the six transactions mentioned above (Amin & Soh, 2019).

Finally, there is also a question on the degree to which corruption is an obstacle to the

current operations of the firms (*j30f_ To what degree are each of the following an obstacle to the current operations of this establishment? – Corruption*). We opt for this aspect for our corruption measure, since the degree to which corruption is perceived as an obstacle to business can be beneficial for several reasons. First, it captures a broader view of how corruption affects businesses beyond specific cases of bribe payments to public officials to get things done (e.g. factors related to direct and indirect forms of corruption, such as regulatory delays and favouritism). In this way, we take into consideration what was regarded as the largest limitation of this dataset, and that is the criticism that information on bribes is incomplete due to measuring only interaction between firms and public officials (Gray et al., 2004). Fazekas and Ferrali (2023) also note that measures of perceptions of corruption are typically fine-grained. Additionally, by using the obstacle measure, one can gain a better understanding of how corruption affects firms of varying characteristics across sectors, which may not be fully captured by examining bribe payments alone. This aligns with a view on institutional and governance quality in different regions or industries, as the dimensions of corruption that present the largest obstacles to doing business are likely to vary across countries and firms within countries (Knack, 2006). Finally, firms may misreport bribes due to the sensitive nature of admitting such practices, and using a perception-based question could reduce social desirability bias (see, e.g. Jensen et al., 2010 for a detailed discussion). The question on paid bribes is included as an independent variable, as it is only one of the factors contributing to the perception of corruption as an obstacle to doing business. Table 1 presents the list and descriptions of all variables used in the analysis.

Table 1. List of variables (source: own compilation based on World Bank, n.d.)

Variable code	Variable description	Variable type
a1	Country	Categorical
b1	The firm's current legal status	Categorical
balkans	The firm is from the country in the WB region	Categorical
b3	Percentage of the firm's largest owner (ownership concentration proxy)	Numerical
b4	Female owners	Categorical
b5	Firms age	Numerical
b8	The firm has an internationally recognised quality certification	Categorical (1 if firm has an internationally recognized quality certification, 0 otherwise)
e6	The firm uses technology licensed from a foreign-owned company, excluding office software?	Categorical (1 if firm use technology licensed from a foreign-owned company, 0 otherwise)
e11	The firm competes against unregistered or informal establishments	Categorical (1 if firm compete against unregistered or informal establishments, 0 otherwise)
recent_ innovation	The firm has introduced new or improved products or services in the last three years	Categorical (1 if firm introduced new or improved products or services and /or introduced any new or improved process, 0 otherwise)

End of Table 1

Variable code	Variable description	Variable type
i1	The firm had to pay for security, for example, equipment, personnel, or professional security services, including internet security	Categorical (1 if firm pays for security, 0 otherwise)
k21	The firm had its annual financial statements checked and certified by an external auditor	Categorical (1 if firm had external audit, 0 otherwise)
j3	The firm was visited or inspected by tax officials or required to meet with them	Categorical (1 if firm was inspected, 0 otherwise)
j6a	The firm has secured or attempted to secure a government contract	Categorical (1 if firm secured or attempted to secure a government contract, 0 otherwise)
l1	Permanent, full-time workers at the end of last fiscal year	Numerical
a4a	Firm sector	Categorical

3.2. Methodological approach

While grounded in institutional and agency theory, our research differs methodologically by using a data-driven RF approach to uncover empirical patterns in firm-level behaviour. The RF algorithm, first proposed by Breiman (2001), has several advantages over more standard econometric tools (such as logit and other parametric methods). Chief among these is the ability to capture complex non-linear relationships between variables without imposing a specific functional form (Breiman, 2001). RF is also highly resistant to undue outlier influencing its predictions and results, making it a good choice for potentially noisy polling data. RF also has several advantages in the context of polling data compared to other machine learning techniques, such as SVM (*support vector machines*) and neural networks. Namely, RF handles missing and incomplete data better and can handle categorical variables without the need for one-hot encoding and similar pre-processing techniques. Additionally, due to its ensemble nature, it has a lower risk of overfitting than SVM or neural networks, as well as being more interpretable than both of them due to the existence of built-in feature importance metrics. These, among other benefits of RF are well documented in the economics and social sciences literature (e.g. Mullainathan & Spiess, 2017; Varian, 2014; Lima & Delen, 2020). Specifically, Lima and Delen (2020) find that the RF as an ensemble-type ML algorithm is the most accurate prediction classification model on their selected dataset, with SVM and neural network coming in second. Possible disadvantages of RF are potential interpretability issues, i.e. it is usually more difficult to find or determine causal connections than with logistic regression. Another common disadvantage is high computing cost compared with logistic regression and SVM, which, in our case, does not present a problem since our data can be analysed by a single machine without issues.

More technically, RF is a type of decision tree classifier that collects many "weak" learners (decision trees) into an ensemble that outperforms its precision constituents. For classification tasks, a decision tree recursively subsets an independent variable into disjoint regions, making predictions for each region based on the majority class. This forms a tree-like structure that, at each node, makes a split in the data based on an independent variable characteristic. A generic representation is given in Figure 1.

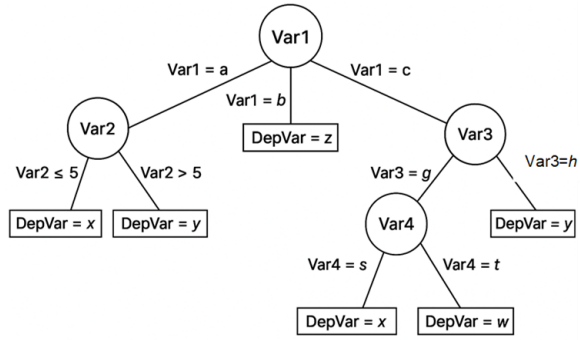

Figure 1. Generic representation of RF decision tree classifier (source: Murray & Scime, 2010)

Mathematically, for each node a split is chosen by selecting an independent variable X_j from the independent variable set $\{X_1, X_2, \ldots, X_d\}$ and a threshold for that variable t_j. At each split, the algorithm chooses a variable X_j and a threshold t_j that minimises Gini impurity, defined as:

$$G(p) = \sum_{i=1}^{C} p_i (1 - p_i),$$

where p_i is the proportion of classes i in the subset, and C is the number of classes. The innovation of the RF classifier is that at each node it does not use all of d available features, but instead chooses a random subset k (usually defined as $k = \sqrt{d}$) in order to reduce correlation between decision trees to improve prediction accuracy. Each decision tree T_b then outputs a class prediction $\hat{y}_b(X)$ where X is the input vector. The final RF prediction is then given as a majority vote from all decision trees B:

$$y_F(X) = \sum_{b=1}^{B} \arg\arg\max_c 1\{\hat{y}_b(X) = c\},$$

where 1 is the indicator function and c is the class label.

All data analysis was done in the R programming language, using the *tidyverse* and *tidymodels* families of packages, specifically implementing the Random Forest algorithm with the *ranger* package. This choice dictates some feature/variable engineering for the data to be usable. The raw data are coded as numbers, which R classifies as "numeric." However, the survey consisted mainly of categorical (multiple-choice) questions. All such variables were reclassified as "factors." If the respondent did not know or refused to answer, this was coded as a negative number, even for numerical values, creating variables that were both numeric and categorical. Our first step was therefore to turn all negative values into NA (Not Available). The question about the year of firm establishment was transformed into "firm age," representing the number of years a firm had been active. Several multiple-choice categorical variables were recoded into dummy variables, including the dependent corruption perception variable (1 if respondents perceive corruption in doing business, 0 otherwise), as well as legal status, sales_local, recent innovation, bribes, and Balkans. Finally, we omitted NA values from variables b1, b8, e6, e11, i1, k21, j3, and a4a (each with 1% missing) and applied k-nearest neighbours (kNN) imputation to the rest.

We ended up with 16081 observations of our variables. As a first step, observations were divided into a *train* (consisting of 12060 observations) and *test* (4021 observations) set.

Additionally, a process of cross-validation was used on the train set. Cross-validation is an ML technique to assess a model's performance and generalizability. It involves dividing the dataset into multiple subsets, or *folds* and then training the model on some of these subsets while testing it on the remaining ones. The most widely used k-fold cross-validation is where the dataset is split into k equally sized folds. The model is trained on k-1 folds and tested on the remaining folds. This process is repeated *k* times, with each fold used as the test set once. The performance metrics from all k iterations are then averaged to provide a more reliable estimate of how well the model will perform on unseen data. This helps to prevent issues like overfitting and ensures that the model is evaluated on different subsets of the data, leading to a more robust assessment.

We used a more advanced version of cross-validation, which resamples the data in fold creation, making each fold the size of the original train set and stratifying to keep the same ratio of positive to negative cases of the dependent variable as in the original data. Additionally, hyperparameter tuning is done concurrently, requiring each fold to be further divided into a train and test set. Considering these additional modifications, we can define our approach as *stratified nested k-fold cross-validation*. This is implemented by the *vfold_cv* function with default settings in the *rsample* package. We opted for 10 folds, each consisting of approximately 12050 observations in the fold train set, and 1340 in the fold test set.

3.3. Performance metrics

Our task was to predict whether a firm perceives the environment it operates in as corrupt (positive result), or not corrupt (negative result). There are four possible outcomes: we predict a positive result when in reality it is positive (this is called a true positive or TP), we predict a positive result when in reality it is negative (a false positive of FP), we predict a negative result when it is actually negative (a true negative or TN), or we predict a negative result when in reality it is positive (a false negative, or FN).

These results are often summarised by a confusion matrix, which gives a summarised view of the number of occurrences of the four possible results of a classification problem. A generalised version of a confusion matrix is given in Table 2.

Table 2. Generalised version of a confusion matrix (source: own compilation)

Reality/Prediction	Predict Positive	Predict Negative
Positive	TP	FN
Negative	FP	TN

The confusion matrix is the core from which other standard performance metrics are drawn, such as accuracy, which gives the total number of correct predictions defined as:

$$accuracy = \frac{TP + TN}{TP + TN + FN + FP}. \quad (1)$$

True positive rate (TPR), also known as Recall or Sensitivity, which gives the proportion of correctly predicted positive cases out of all positive cases:

$$TPR = \frac{TP}{TP + FN}. \quad (2)$$

Precision, which gives the ratio of the correct positive predictions out of all positive predictions, defined as:

$$precision = \frac{TP}{TP + FP}. \tag{3}$$

True negative rate (TNR, also known as specificity), which gives the ratio of correctly predicted negatives out of all actual negatives, defined as:

$$TNR = \frac{TN}{TN + FP}. \tag{4}$$

These standard metrics, when taken separately, can be misleading (for example, for rare events, you can have a very high accuracy simply by always predicting the negative result, but you would correspondingly have a very low TPR, etc.), but are often difficult to interpret when evaluated as a group. As a result, popular alternatives attempt to translate all the information a confusion matrix gives into a single "goodness of fit" number. One such combined metric is the F1 score, which gives the harmonic mean between precision and TPR:

$$F1 = 2 \times \frac{precision \times TPR}{precision + TPR}. \tag{5}$$

An additional problem in classification tasks is that the output of a classification algorithm is a probability of an event being positive or negative, which implies there must be a cutoff point (for instance, 50%) above which the event is classified as positive. However, different tasks require different cutoff points, resulting in vastly different confusion matrices and, consequently, different performance metrics. A very popular metric that takes this problem into account is ROC (Reciever Operating Characteristic) AUC (Area Under the Curve). The ROC plots the TPR vs the FPR, each point representing the model's ability to discriminate between positive and negative cases with a specific threshold. The AUC of the ROC then gives a single number of how well the model does, considering all possible thresholds. Mathematically it is simply an integral of TPR as a function of FPR:

$$\int_0^1 TPR\left(FPR\right) dFPR. \tag{6}$$

It is a very widely used metric in classification tasks due to its all-encompassing character and will be our default metric for choosing between models based on different hyperparameter settings.

There are several ways to measure variable importance in an RF setting. We opt for the classic *gini impurity* metric, a standard choice in classification tasks. Gini impurity is defined as:

$$Gini\ impurity = 1 - \sum_{i=1}^{k} p_i^2, \tag{7}$$

where k is the number of classes (in our case we are dealing with a binary classification task so k = 2), p_i is the proportion of class i in the node. Gini impurity takes on a value of 0 if there is a perfect split (i.e. a certain split is able to completely distinguish between positive and negative cases of the dependent variable), and conversely takes on a value of 0.5 when there is perfect uniformity in the post-split distribution of the dependent variable classes (i.e. maximum impurity). In effect, this measures how much a variable contributed to a model ability to distinguish between positive and negative cases in each split, and can therefore

be used to ascertain variable importance in a model. Since an RF is composed of hundreds/thousands of individual decision trees, each of which is composed of dozens of splits, a Gini impurity-based variable importance measure will usually add up all of the reductions in Gini impurity in each node and each tree for every independent variable in the model. This results in an importance ranking of every variable by total impurity reduction, which is a number that has meaning only relative to other variables, i.e., to obtain a ranking, its absolute value is of little consequence.

One of the advantages of RF, as opposed to more state-of-the-art models like xgboost, is that RF gives good values with default settings of hyperparameters; in other words, not much tuning and computation time is required. While there are many hyperparameters to choose from in RF, there are three main ones. These are the: *mtry* (number of random independent variables sampled at each node); *ntrees* (total number of trees created); *min_n* (minimum number of data points in a node to allow further splits). We take the default total number of trees (1000) as given and we are left with two hyperparameters for tuning: *mtry* and *min_n*. We opt for a grid depth of 11, which gives 11 different combinations of the two hyperparameters tested on our 10 folds. We plot our results in Figure 2. We focus on the *roc_auc* metric as our main criteria when evaluating hyperparameter performance as well as overall model performance. We can see that overall, a smaller number of randomly selected predictors (*mtry*) gives better results, while the model seems to do well with a wide range of minimal node sizes (*min_n*). We can observe (Figure 2) that after extensive cross-validation, the model predicts an accuracy of 0.79 (79%) and a ROC AUC of 0.755, with the selected hyperparameters of *mtry* = 3, and *min_n* = 9. We take this specification to finalise the model.

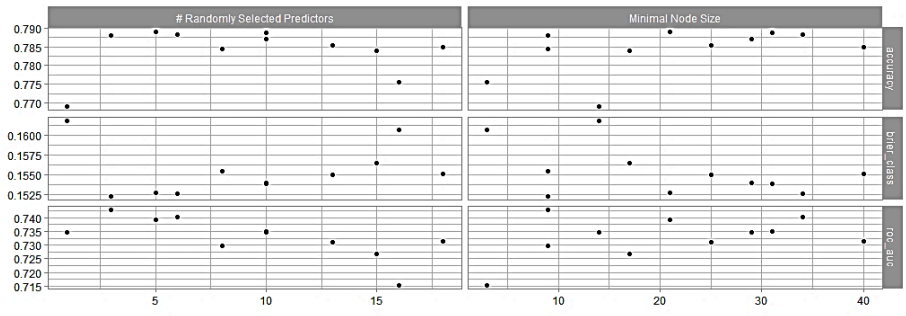

Figure 2. Accuracy and ROC AUC (source: own calculation)

4. Results and discussion

After selecting the best-performing hyperparameters, we used them again on the entire original *train* set to finalise our model. Then, we see how well our model performs on the original test set, i.e., unknown data. The results are very close to cross-validation expectations. The testing of our finalised model on the test set gives a ROC AUC of 0.755 and an accuracy of 0.79. Our final results are almost an exact match to our cross-validation expectations, which suggests that the model is well-calibrated and the risk of *overfitting* or *underfitting* is minimal.

We can also obtain variable importance metrics with the *vip* function from the *vip* package. Of the several ways to measure variable importance, we opt for a *gini impurity-based* technique, explained in Section 3. Our results are shown in Figure 3.

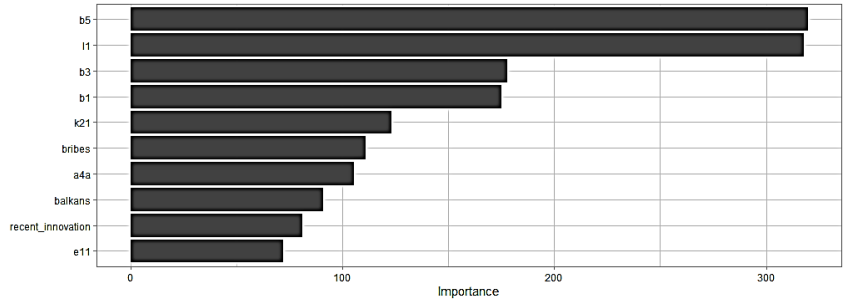

Figure 3. Obtained results on variable importance (source: own calculation)

The RF model output provides rankings of firm-level features by their importance in predicting corruption outcomes. While the model does not provide causal inference, it identifies the characteristics that most consistently differentiate firms experiencing corruption in terms of the previously defined corruption variable. Precisely, we obtained that the variables, i.e. factors that are best predictors for corruption are the following (in order of importance): firm age (*b5*), firm size (*l1*), percentage of firm owned by the largest owner (i.e. ownership concentration, *b3*), legal status of the firm (*b1*); financial statements checked and certified by external auditors (*k21*); experience with bribe requests in any of the dimensions of doing business (bribes); sector in which firm operates (*a4a*); countries non-members of EU (Balkans); innovation activities (*recent_innovation*); and pressures from informal sector competition (*e11*). These factors can be grouped into several categories. The first group is directly related to firm characteristics; the second is related to various processes (financial or regulatory) to which the firm is exposed; and the third is related to market competition stemming from the informal sector. Table 3 presents how results align with agency theory and institutional theory, and how they contribute to the existing research capturing the causality (through e.g. OLS, Instrumental variables, logistic regressions).

Table 3. Relevance of variables from the agency and institutional theory setting (source: own compilation)

Agency theory	Description	Previous research
Ownership concentration (*b3*)	A higher concentration may reduce corruption risk through tighter control, or increase it through monopolised decision-making.	Colonnelli and Prem (2022), Nguyen (2020)
External audit (*k21*)	External monitoring reduces information asymmetry and can act as a disciplining device.	Farooq and Shehata (2018), Cieślik and Goczek (2022)
Bribes (*bribes*)	Indicates agent-level behaviour under weak oversight and reflects direct corruption experience.	Gray et al. (2004), Fazekas and Ferrali (2023)
Innovation (*recent_innovation*)	It can reflect an internal strategy that increases interaction with regulators, raising agency risk if institutions are weak.	Belitski et al. (2021), Riaz et al. (2018)

End of Table 3

Agency theory	Description	Previous research
Legal status (*b1*)	Affects transparency, liability, and internal control structures, which are central to principal-agent dynamics.	Djankov et al. (2002), Audretsch et al. (2022)
Institutional theory		
Firm age (*b5*)	Older firms may navigate institutional inefficiencies better or become embedded in informal arrangements.	Colonnelli et al. (2022), Campos et al. (2010)
Firm size (*l1*)	Larger firms face more regulatory interaction.	Fisman et al. (2024), Djankov et al. (2002)
Country (b*alkans*)	Captures institutional maturity and anti-corruption enforcement differences.	Mungiu-Pippidi (2013), Dávid-Barrett and Fazekas (2020)
Informal competition (*e11*)	Reflects institutional failure to enforce market rules.	Audretsch et al. (2022), Dimant and Tosato (2018)
Sector (*a4a*)	Sectors vary in institutional exposure (e.g., construction and manufacturong) and corruption risk.	Decarolis and Giorgiantonio (2022), Fazekas et al. (2022)

Our findings complement the existing research on the significance and causality of various factors at the firm level for corruption that can be found in the literature and presented in Table 3. For example, Cieślik and Goczek (2022) demonstrate that the level of corruption is correlated with the time spent on regulations and inspections. They suggest that firms investing more time in administrative tasks tend to perceive higher corruption levels and are compelled to pay larger bribes. Additionally, while some initial studies suggested that larger firms were less likely to pay bribes or pay lower amounts, analyses that account for reverse causality by employing instrumental variables reveal a non-linear relationship, indicating that increased firm size actually leads to greater corruption and bureaucratic burdens (Nguyen, 2020). Moreover, robust controls and transparent reporting directly address the information asymmetry central to principal-agent theory. Farooq and Shehata (2018) demonstrate that firms with robust external auditing practices are significantly less likely to engage in corruption, underscoring the importance of effective oversight. In addition, the findings confirm the role of the informal sector (Jackson, 2023) and innovation activities (Riaz et al., 2018) as potential factors influencing corruption. Additionally, our findings highlight the importance of the *balkans* variable, as these countries are progressing towards EU membership.

In combination with previous research, our findings can inform policy in several ways, which can help to promote broader institutional reforms. First, they point to the need for institutional reforms to strengthen oversight in specific firm segments (e.g., medium-sized firms in vulnerable sectors) and ensure that audit systems and ownership structures promote transparency rather than enable discretion. Digitised reporting systems and blockchain technologies can lower the risk of corruption among high-risk firms by boosting transparency and reducing discretion. For instance, tax authorities and licensing bodies could introduce e-filing platforms with built-in flags for firms flagged by predictive models. Blockchain can also improve integrity in high-risk processes like procurement or financial disclosures by recording transactions and compliance data in tamper-proof ledgers, ensuring accountability

and traceability. Adam and Fazekas (2021) provide a detailed overview of different types of ICT-based anti-corruption interventions. Next, to address informal competition, governments must simplify registration processes and offer incentives for formalisation. Finally, obtained insights could also support EU accession evaluations by identifying structural corruption risks in candidate countries' business environments, particularly when aligned with other governance indicators.

5. Conclusions

The main goal of this paper was to assess the prediction capability of the various indicators at the firm level through the application of the random forest ML approach. Based on the analysis performed on a sample of firms from EU members and WB countries, and the results obtained, we classified the indicators by their relevance into several categories. The first group refers directly to the firm's characteristics, the second to different processes (financial or regulatory) to which the company is exposed, and the third to market competition from the informal sector. The results provide a contribution to understanding the key factors related to corruption in business and can help policymakers, regulatory bodies, and firms better understand and lessen corruption risks. In addition, the findings confirm the role of the informal sector and innovation activities as significant factors of corruption. Additionally, our findings highlight the importance of the *balkans* variable, as these countries are progressing towards EU membership. From an institutional perspective, firms in WB countries face higher predicted corruption risks, reaffirming that weak formal institutions and enforcement capacity are critical drivers of corruption. The EU must thus prioritise reforms addressing corruption, particularly given that past enlargement rounds, which included countries with high corruption and incomplete transitions, have resulted in increased corruption levels within the EU itself.

The findings confirm the main expectations from both agency theory and the institutional perspective. First, from the principal-agent model perspective, corruption arises from misalignment of incentives between principals, in our case, regulatory bodies, in terms of auditors, whose presence can reduce information asymmetries, but on the other hand also increase the possibilities for paying bribes. Next, the role of ownership concentration is relevant from both the principal-agent problem within (owners-managers) and outside of firms (owners-politicians). Also, the firm's size and age are relevant, as larger and older firms have more frequent interactions with agents and could have more resources to engage in or resist corruption (depending on institutional constraints). Thus, the results also confirm institutional theory expectations through variables related to WB countries, sectors, informal competition and innovation activities. They imply that corruption could be predicted by the variables that indirectly capture the quality of institutions. These are all insights useful for risk-based supervision.

However, the research has some limitations, primarily related to the data and measurement aspects of corruption. First, although the use of WBES data for corruption research is well-established, this study is one of the first to apply Random Forest to firm-level corruption perceptions in both the EU and the WB. Also, an important issue with survey-based corruption measures is whether respondents answer such potentially sensitive questions honestly. This is especially important for experiences with corruption, as misreported or underreported experiences pose the most significant threat to measuring corruption. However, the insights into existing literature showed that such reporting bias is negligible. Next, the feature

importance metric of an RF model orders variables based on their influence on the model's accuracy, but it does not provide direct evidence of causality. For example, some usual suspects of corruption, such as securing government contracts, did not emerge among the top predictors in our random forest model. While this may seem counterintuitive to corruption literature, it likely reflects data-specific factors such as limited variation or lower predictive relevance in this context. Nonetheless, this empirical exercise does give a good starting point for further investigations into concrete causal links between factors and outcomes.

Future research could expand the analysis by including more determinants of corruption, especially cultural and social factors that influence firm behaviour. Additionally, refining corruption variables to focus on areas like public procurement would allow for more targeted insights into how corruption distorts competition among firms dependent on government contracts. It would also be valuable to examine the long-term impact of corruption on firm sustainability, innovation capacity, and competitiveness, thereby linking integrity risks to broader economic outcomes. These would enable the design of more detailed models and offer a stronger empirical foundation for developing policies, targeted audits, and institutional reforms. Finally, the application of additional ML methods for the analysis of individual factors could provide a more comprehensive insight into corrupt practices at the level of firms, and enhance the explanatory power of models and capture complex, non-linear relationships between firm characteristics and corruption exposure.

Author contributions

VV was responsible for the research design and literature review. MD was responsible for ML data analysis in R. VV and MD were both responsible for data collection, data interpretation and discussion of results.

Disclosure statement

Authors declare that they do not have any competing financial, professional, or personal interests from other parties.

References

Adam, I., & Fazekas, M. (2021). Are emerging technologies helping win the fight against corruption? A review of the state of evidence. *Information Economics and Policy*, *57*, Article 100950. https://doi.org/10.1016/j.infoecopol.2021.100950

Amin, M., & Soh, Y. C. (2019). *Corruption and country size*: *evidence using firm-level survey data* (World Bank Policy Research Working Paper, 8864). SSRN. https://doi.org/10.1596/1813-9450-8864

Audretsch, D. B., Belitski, M., Chowdhury, F., & Desai, S. (2022). Necessity or opportunity? Government size, tax policy, corruption, and implications for entrepreneurship. *Small Business Economics*, *58*(4), 2025–2042. https://doi.org/10.1007/s11187-021-00497-2

Bartlett, W. (2023). The performance of politically connected firms in South East Europe: State capture or business capture? *Post-Communist Economies*, *35*(4), 351–367. https://doi.org/10.1080/14631377.2023.2188694

Belitski, M., Grigore, A. M., & Bratu, A. (2021). Political entrepreneurship: Entrepreneurship ecosystem perspective. *International Entrepreneurship and Management Journal*, *17*(4), 1973–2004. https://doi.org/10.1007/s11365-021-00750-w

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1257/jep.31.2.87

Burke, R. J., & Cooper, C. L. (Eds.). (2009). *Research companion to corruption in organizations*. Edward Elgar Publishing. https://doi.org/10.4337/9781849801928

Campos, N. F., Dimova, R. D., & Saleh, A. (2010). *Whither corruption? A quantitative survey of the literature on corruption and growth* (CEPR Discussion Paper No. DP8140). SSRN. https://doi.org/10.2139/ssrn.1716129

Cieślik, A., & Goczek, Ł. (2018). Corruption, privatisation and economic growth in post-communist countries. *Europe-Asia Studies*, *70*(8), 1303–1325. https://doi.org/10.1080/09668136.2018.1511771

Cieślik, A., & Goczek, Ł. (2022). Who suffers and how much from corruption? Evidence from firm-level data. *Eurasian Business Review*, *12*(3), 451–473. https://doi.org/10.1007/s40821-021-00185-x

Colonnelli, E., & Prem, M. (2022). Corruption and firms. *The Review of Economic Studies*, *89*(2), 695–732. https://doi.org/10.1093/restud/rdab040

Colonnelli, E., Gallego, J., & Prem, M. (2022). What predicts corruption?. In P. Buonanno, P. Vanin & J. Vargas (Eds.), *A modern guide to the economics of crime* (pp. 345–373). Edward Elgar Publishing. https://doi.org/10.4337/9781789909333.00020

d'Agostino, G., Dunne, J. P., & Pieroni, L. (2016). Government spending, corruption and economic growth. *World Development*, *84*, 190–205. https://doi.org/10.1016/j.worlddev.2016.03.011

Dávid-Barrett, E., & Fazekas, M. (2020). Grand corruption and government change: An analysis of partisan favoritism in public procurement. *European Journal on Criminal Policy and Research*, *26*(4), 411–430. https://doi.org/10.1007/s10610-019-09416-4

Decarolis, F., & Giorgiantonio, C. (2022). Corruption red flags in public procurement: New evidence from Italian calls for tenders. *EPJ Data Science*, *11*(1), Article 16. https://doi.org/10.1140/epjds/s13688-022-00325-x

Dimant, E., & Tosato, G. (2018). Causes and effects of corruption: What has past decade's empirical research taught us? A survey. *Journal of Economic Surveys*, *32*(2), 335–356. https://doi.org/10.1111/joes.12198

Djankov, S., La Porta, R., Lopez-de-Silanes, F. & Shleifer, A. (2002). The regulation of entry. *The Quarterly Journal of Economics*, *117*(1), 1–37. https://doi.org/10.1162/003355302753399436

Doria, L. M., Doria, F. F., Figueiredo, P., Sampaio, A., & Sampaio, R. R. (2022). A machine learning approach on the problem of corruption. *International Journal of Advanced Engineering Research and Science*, *9*(3), 277–282. https://dx.doi.org/10.22161/ijaers.93.33

Egger, P., & Winner, H. (2005). Evidence on corruption as an incentive for foreign direct investment. *European Journal of Political Economy*, *21*(4), 932–952. https://doi.org/10.1016/j.ejpoleco.2005.01.002

European Commission. (2023). *2023 Eurobarometer survey*: *Business' attitudes towards corruption*.

European Commission. (2024). *Directorate-General for Migration and Home Affairs*. *High-risk areas of corruption in the EU – A mapping and in-depth analysis*. Publications Office of the European Union. https://data.europa.eu/doi/10.2837/5907939

Enste, D. H., & Heldman, C. (2018). The consequences of corruption. In B. Warf (Ed.) *Handbook on the geographies of corruption* (pp. 106–119). Edward Elgar Publishing. https://doi.org/10.4337/9781786434753.00011

Farooq, O., & Shehata, N. F. (2018). Does external auditing combat corruption? Evidence from private firms. *Managerial Auditing Journal*, *33*(3), 267–287. https://doi.org/10.1108/MAJ-08-2017-1634

Fazekas, M., & Ferrali, R. (2023). *Advances in measuring corruption and agenda for the future* (Working Paper GTI-WP/2023:01). Government Transparency Institute. https://www.govtransparency.eu/wp-content/uploads/2023/03/Fazekas-Ferrali_Corr-measurement_article_WPformatted_final.pdf

Fazekas, M., Sberna, S., & Vannucci, A. (2022). The extra-legal governance of corruption: Tracing the organization of corruption in public procurement. *Governance*, *35*(4), 1139–1161. https://doi.org/10.1111/gove.12648

Fisman, R., Guriev, S., Ioramashvili, C., & Plekhanov, A. (2024). Corruption and firm growth: Evidence from around the world. *The Economic Journal*, *134*(660), 1494–1516. https://doi.org/10.1093/ej/uead100

García-Gómez, C. D., Bilyay-Erdogan, S., Demir, E., & Díez-Esteban, J. M. (2025). A new piece in the puzzle: Corruption and financial constraints – evidence from European firms. *Business Ethics, the Environment & Responsibility*, *35*(2), 693–716. https://doi.org/10.1111/beer.12815

Glaeser, E. L., & Saks, R. E. (2006). Corruption in America. *Journal of Public Economics*, *90*(6–7), 1053–1072. https://doi.org/10.1016/j.jpubeco.2005.08.007

Glaeser, E. L., La Porta, R., Lopez-de-Silanes, F., & Shleifer, A. (2004). Do institutions cause growth?. *Journal of Economic Growth*, *9*, 271–303. https://doi.org/10.1023/B:JOEG.0000038933.16398.ed

Goel, R. K., Mazhar, U., & Ram, R. (2021). *Size matters*: *corruption perceptions versus corruption experiences by firms* (CESifo Working Paper, No. 9221). SSRN. https://doi.org/10.2139/ssrn.3898319

Graeff, P., & Svendsen, G. T. (2013). Trust and corruption: The influence of positive and negative social capital on the economic development in the European Union. *Quality & Quantity*, *47*, 2829–2846. https://doi.org/10.1007/s11135-012-9693-4

Gray, C. W., Hellman, J. S., & Ryterman, R. (2004). *Anticorruption in transition 2*: *Corruption in enterprise-state interactions in Europe and Central Asia*, *1999–2002* (Vol. 2). The World Bank. https://openknowledge.worldbank.org/server/api/core/bitstreams/cfcbf274-136b-58e6-a587-048e18310f9f/content

Grzymala-Busse, A. (2007). *Rebuilding leviathan party competition and state exploitation in post-communist democracies*. Cambridge University Press. https://doi.org/10.1017/CBO9780511618819

Gugiu, M. R., & Gugiu, P. C. (2016). Economic crisis and corruption in the European Union. *Journal of Methods and Measurement in the Social Sciences*, *7*(1), 1–22. https://doi.org/10.2458/v7i1.19398

Gupta, S., Davoodi, H., & Alonso-Terme, R. (2002). Does corruption affect income inequality and poverty?. *Economics of Governance*, *3*, 23–45. https://doi.org/10.1007/s101010100039

Huang, C. J. (2016). Is corruption bad for economic growth? Evidence from Asia-Pacific countries. *The North American Journal of Economics and Finance*, *35*, 247–256. https://doi.org/10.1016/j.najef.2015.10.013

Jackson, E. (2023). Informality as a driving force for corruption in economy: A neoclassical simulation. *Economic Analysis Letters*, *2*, 60–65. https://doi.org/10.58567/eal02020008

Jaggi, B., Allini, A., Ginesti, G., & Macchioni, R. (2021). Determinants of corporate corruption disclosures: Evidence based on EU listed firms. *Meditari Accountancy Research*, *29*(1), 21–38. https://doi.org/10.1108/MEDAR-11-2019-0616

Jensen, N. M., Li, Q., & Rahman, A. (2010). Understanding corruption and firm responses in cross-national firm-level surveys. *Journal of International Business Studies*, *41*, 1481–1504. https://doi.org/10.1057/jibs.2010.8

Khan, S. (2022). Investigating the effect of income inequality on corruption: New evidence from 23 emerging countries. *Journal of the Knowledge Economy*, *13*(3), 2100–2126. https://doi.org/10.1007/s13132-021-00761-6

Klitgaard, R. E. (1988). *Controlling corruption*. University of California Press.

Knack, S. F. (2006). *Measuring corruption in Eastern Europe and Central Asia*: *A critique of the cross-country indicators* (World Bank Publications, Vol. 3968). https://doi.org/10.1596/1813-9450-3968

Köbis, N., Starke, C., & Rahwan, I. (2022). The promise and perils of using artificial intelligence to fight corruption. *Nature Machine Intelligence*, *4*(5), 418–424. https://doi.org/10.1038/s42256-022-00489-1

Lambsdorff, J. G. (2006). Causes and consequences of corruption: What do we know from a cross-section of countries. *International Handbook on the Economics of Corruption*, *1*, 3–51. https://doi.org/10.4337/9781847203106.00007

Lambsdorff, J. G. (2007). *The institutional economics of corruption and reform*: *Theory*, *evidence and policy*. Cambridge University Press. https://doi.org/10.1017/CBO9780511492617

Lima, M. S. M., & Delen, D. (2020). Predicting and explaining corruption across countries: A machine learning approach. *Government Information Quarterly*, *37*(1), Article 101407. https://doi.org/10.1016/j.giq.2019.101407

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106. https://doi.org/10.1257/jep.31.2.87

Mungiu-Pippidi, A. (2013). The good, the bad and the ugly: Controlling corruption in the European Union. *Advanced Policy Paper for Discussion in the European Parliament*, *9*, 61–63. https://www.againstcorruption.eu/wp-content/uploads/2013/03/ANTICORRP-Policy-Paper-on-Lessons-Learnt-1_protected1.pdf

Murray, G. & Scime, A. (2010). Microtargeting and electorate segmentation: Data mining the American National Election Studies. *Journal of Political Marketing*, *9*, 143–166. https://doi.org/10.1080/15377857.2010.497732

Nguyen, T. D. (2020). Does firm growth increase corruption? Evidence from an instrumental variable approach. *Small Business Economics*, *55*, 237–256. https://doi.org/10.1007/s11187-019-00160-x

Organisation for Economic Co-operation and Development. (2008). *Corruption*: *A glossary of international standards in criminal law*. OECD Publishing. www.oecd.org/daf/antiribery/41194428.pdf

Poltoratskaia, V., & Fazekas, M. (2024). Data analytics for anti-corruption in public procurement. In S. Williams, & J. Tillipman (Eds.), *Routledge handbook of public procurement corruption* (pp. 42–59). Routledge. https://doi.org/10.4324/9781003220374-6

Riaz, M. F., Cherkas, N., & Leitão, J. (2018). Corruption and innovation: Mixed evidences on bidirectional causality. *Journal of Applied Economic Sciences*, *13*(2(56)), 378–384.

Rose-Ackerman, S. (2017). When is corruption harmful?. In A. J. Heidenheimer & M. Johnston (Eds.), *Political corruption*: *Concepts and contexts* (pp. 353–372). Routledge. https://doi.org/10.4324/9781315126647-32

Rusch, J. (2021). *Is there a role for machine learning in anti-corruption risk and compliance?*. Faculty of Law Blogs, University of Oxford. https://blogs.law.ox.ac.uk/business-law-blog/blog/2021/06/there-role-machine-learning-anti-corruption-risk-and-compliance

Saha, S., & Gounder, R. (2013). Corruption and economic development nexus: Variations across income levels in a non-linear framework. *Economic Modelling*, *31*, 70–79. https://doi.org/10.1016/j.econmod.2012.11.012

Shleifer, A., & Vishny, R. W. (1993). Corruption. *The Quarterly Journal of Economics*, *108*(3), 599–617. https://doi.org/10.2307/2118402

Svensson, J. (2003). Who must pay bribes and how much? Evidence from a cross section of firms. *Quarterly Journal of Economics*, *118*(1), 207–230. https://doi.org/10.1162/00335530360535180

Tanzi, V. (1998). Corruption around the world: Causes, consequences, scope, and cures. *Staff Papers*, *45*(4), 559–594. https://doi.org/10.2307/3867585

Transparency International. (2021). *Addressing corruption as a driver of democratic decline*. Policy position. https://files.transparencycdn.org/images/2021-Addressing-corruption-as-driver-of-democratic-decline-Summit-for-Democracy-PositionPaper-EN.pdf

Uslaner, E. M. (2017). Political trust, corruption, and inequality. In S. Zmerli & T. W. Van der Meer (Eds.), *Handbook on political trust* (pp. 302–315). Edward Elgar Publishing. https://doi.org/10.4337/9781782545118.00030

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, *28*(2), 3–28. https://doi.org/10.1257/jep.28.2.3

Warner, C. (2011). *The best system money can buy*: *Corruption in the European Union*. Cornell University Press. https://doi.org/10.1093/bjc/azn061

World Bank. (1997). *Helping countries combat corruption*: *The role of the World Bank*. https://documents1.worldbank.org/curated/en/799831538245192753/pdf/Helping-Countries-Combat-Corruption-The-Role-of-the-World-Bank.pdf

World Bank. (2023). *Evidence on public procurement from firm-level surveys*: *Global statistics from the World Bank Enterprise Surveys and a novel public procurement survey module*. Equitable Growth, Finance & Institutions Insight. https://documents1.worldbank.org/curated/en/099122723173512944/pdf/P17503317d58130361a8391945d653eb092.pdf

World Bank. (n.d.). *World Bank enterprise surveys* [database]. https://www.enterprisesurveys.org/en/enterprisesurveys