UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Department
of Economics

WORKING PAPERS

A Machine Learning Approach to Analyze and Support Anti-Corruption Policy

Elliott Ash, Sergio Galletta, Tommaso Giommoni

December 2021 - WP N. 8 Year 2021

Working papers – Department of Economics
n. 08

A Machine Learning Approach to Analyze and Support Anti-Corruption Policy

Elliott Ash, Sergio Galletta, Tommaso Giommoni

Università degli Studi di Bergamo
2021

A Machine Learning Approach to Analyze and Support Anti-Corruption Policy / Elliott Ash, Sergio Galletta, Tommaso Giommoni – Bergamo: Università degli Studi di Bergamo, 2021.
Working papers of Department of Economics, n. 8
ISSN: 2974-5586
DOI: 10.13122/WPEconomics_8

Progetto grafico: Servizi Editoriali – Università degli Studi di Bergamo
Università degli Studi di Bergamo
via Salvecchio, 19
24129 Bergamo
Cod. Fiscale 80004350163
P. IVA 01612800167

https://aisberg.unibg.it/handle/10446/201129

A Machine Learning Approach to Analyze and Support Anti-Corruption Policy[⋆]

Elliott Ash[1], Sergio Galletta[2], Tommaso Giommoni[1]

[1]*ETH Zürich*
[2]*University of Bergamo*

Abstract

Can machine learning support better governance? In the context of Brazilian municipalities, 2001-2012, we have access to detailed accounts of local budgets and audit data on the associated fiscal corruption. Using the budget variables as predictors, we train a tree-based gradient-boosted classifier to predict the presence of corruption in held-out test data. The trained model, when applied to new data, provides a prediction-based measure of corruption that can be used for new empirical analysis or to support policy responses. We validate the empirical usefulness of this measure by replicating and extending some previous empirical evidence on corruption issues in Brazil. We then explore how the predictions can be used to support policies toward corruption. Our policy simulations show that, relative to the status quo policy of random audits, a targeted policy guided by the machine predictions could detect almost twice as many corrupt municipalities for the same audit rate. Similar gains can be achieved for a politically neutral targeting policy that equalizes audit rates across political parties.

JEL Classification: C53, D73, H83, K42.
Keywords: algorithmic decision-making, corruption policy, local public finance.

[⋆] We benefited from comments by participants at seminars at the Collegio Carlo Alberto Turin, University of St. Gallen, ETH Zürich, University of Zürich, the "Advances in Economics Winter Symposium" (Bergamo), Bonn Law and Economics Workshop, CESifo Area Conference on Public Economics and the Online Workshop in Computational Analysis of Law. We are grateful to Francisco Cavalcanti for sharing essential data for conducting the analysis. David Cai, Matteo Pinna, and Angelica Serrano provided excellent research assistance. Thanks to Daniel Bjorkegren, Aniket Kesari, Himabindu Lakkaraju, Michael Livermore, Julian Nyarko, Alexis Perakis, and Martin Stancsics for comments on previous drafts.
This version: December 2021 (First version: April 2020).

Email addresses: ashe@ethz.ch (Elliott Ash), sergio.galletta@unibg.it (Sergio Galletta), giommoni@ethz.ch (Tommaso Giommoni)

1. Introduction

A large body of anecdotal and empirical evidence speaks to the deep and negative impacts of corruption. According to recent United Nations statistics, for example, international corruption costs the global economy over 3.6 trillion USD annually. On a more micro level, social scientists have demonstrated that foul play by government actors does real harm to the average citizen. These harms lead to responses in politics and political participation (Ferraz and Finan, 2008; Chong et al., 2015), undermine trust toward institutions (Morris and Klesner, 2010), and have additional side effects on the economy (Lagaras et al., 2017).

Accordingly, researchers continue to seek a scientific understanding of corruption. Broadly speaking, the previous research has identified two important factors. First, electoral incentives play a crucial role in discouraging misbehavior by officials (Ferraz and Finan, 2008; Winters and Weitz-Shapiro, 2013; Poblete-Cazenave, 2021). Second, an effective judicial system to prosecute offenders and enforce the law may be necessary to deter corrupt actions (Becker, 1968; Djankov et al., 2003; Vannutelli, 2021). Despite this impressive progress in understanding the causes and consequences of corruption, a major impediment to further research is the relative lack of data. Corrupt actors have strong incentives to conceal their actions, and therefore measurements of corruption traditionally come from costly government auditing programs.

The difficulties facing corruption research also apply to anti-corruption policy efforts. Even with accountable politicians and well-functioning courts, anti-corruption policies are still often frustrated by the costs of detecting corruption in the first place. Hence, although several countries have introduced monitoring programs to detect wrongdoing, these are typically limited to a relatively small subset of public offices (e.g., Olken, 2007; Ferraz and Finan, 2008; Bobonis et al., 2016). Overall, producing more data on corruption has high social value in terms of social science research, policy experimentation, and strengthening enforcement.

This paper aims to address the problem of undetected corruption using tools from machine learning. The core of our idea is to exploit the fact that corruption, by its nature, is related to how politicians and public officials manage public resources (Mauro, 1998). Thus our analysis focuses on local public finances in Brazilian municipalities. We start with a ground-truth measure of detected corruption, identified and quantified by professional government auditors (Ferraz and Finan, 2008; Brollo et al., 2013). We link this corruption outcome with a rich historical account of local public budgets (with information on 797 fiscal categories).

We use machine learning to predict corruption from the features of the budget accounts. We implement a gradient boosted classifier consisting of an ensemble of decision trees, typically used to identify patterns in high-dimensional datasets. Using only municipal budget characteristics, the classifier can detect the existence and predict the intensity of corruption with high accuracy in held-out (unseen) data. In the best model, we get an accuracy of 72% and an AUC of 0.77, far better than guessing the modal category or prediction using linear models.[1] We show that the model accurately ranks municipalities by the probability of corruption, reproducing the distribution of corruption in held-out data. In a dataset of municipalities that were audited twice, the model can predict within-municipality changes in corruption over time.

The trained prediction model is a complex forest of dozens of decision trees, containing over 10,000 decision nodes that capture decisive non-linearities and interactions between features. Because pairwise correlations between budget features and corruption risk do not give useful insights, we use a feature importance score that identifies which budget factors the model most attends to when making predictions (Hastie et al., 2009). This ranking of features provides some insight into how the model works, especially because we can compare it to the relative frequency that the features are mentioned in the published audit reports. Applying text analysis to a 55-million-word corpus of digitized reports, we show that the corruption-related features identified by the gradient boosted machine are also mentioned significantly more often in those reports. This validation provides evidence that the model is capturing corruption-related activities, rather than just correlated proxies.

With some confidence that the model predictions measure corruption, we extrapolate to the unlabeled budgets and form a synthetic measure of corruption for all municipalities and years. To demonstrate the empirical applicability of the method, we use the predicted corruption measure to replicate previous causal results on local corruption in Brazil. First, we replicate the result from Brollo et al. (2013) that a revenue windfall, based on population thresholds, increases corruption. In particular, we can show this

[1] As a reference, our classifier's performance is similar to that of other papers in economics using machine learning to analyze and support decision-making (Kleinberg et al., 2018; Mullainathan and Obermeyer, 2019)/

result in an untouched sample of municipalities that were never audited by the Brazilian authorities. Normalized coefficient magnitudes using the model predictions are comparable with the estimates obtained by Brollo et al. (2013) using the auditor-produced corruption label as the outcome.

As a second empirical application, we extend the analysis from Avis et al. (2018) and analyze the causal effect of auditing on corruption. Because we have a measure of corruption by year, we can implement an event study analysis. We show that audits reduce corruption in fiscal accounts over the subsequent years, with an average drop of around 2.7% in the probability of malfeasance. Moreover, the effect is especially large for audits that did find corruption, with an average decline of around 18%, about half of the pre-audit mean of 39.5%. In comparison, there is no effect on our measure for audits that did not find corruption.

Beyond their direct interest, these empirical results provide some additional validation about what our machine learning model is capturing. Because predicted corruption responds the same way to treatments (revenue shocks and audits) as true corruption, that suggests that model predictions are driven by budgeting choices that local officials have some control over. If the model were picking up (mostly) non-budget factors, then we would not see a comparable response to the causal treatments. Thus, we obtain additional confidence that the model is capturing corruption-related activities, rather than just correlated proxies. While not strictly necessary for the limited task of helping to detect corruption in the existing data, this validation is reassuring that the model could generalize to new settings and time periods.

In the last part of the paper, we investigate the potential of an audit policy guided by our model's predictions on corruption risk. We show that, compared to the status quo policy of random audits, a targeted approach based on predicted corruption would be significantly more efficient in the policy goals of detecting and deterring corruption. According to our policy simulations, a targeted approach would detect about 80 percent more corrupt municipalities relative to the random lottery (for the same number of implemented audits). Similarly, by targeting the municipalities at highest risk for corruption, the audit agency could obtain the same number of corruption detections as the random-lottery system but with 45 percent fewer audits, with a corresponding reduction in administrative costs. From a deterrence perspective, it is notable that the annual audit probability, conditional on being corrupt, almost doubles under targeted audits, compared to random audits.

Finally, we consider the implementation issue that algorithmic targeting could change the audit rates across political parties, potentially leading to perceptions of bias in a politically sensitive environment. Using the party affiliations of municipal mayors, we show that this concern is relevant, as there is substantial variation across the five main parties in targeting incidence relative to random audits. To address this potential barrier to implementation, we draw on recent developments in algorithmic fairness (Barocas et al., 2019; Rambachan et al., 2020) and adjust our audit targeting policy to equalize audit rates across parties. We show that such a politically neutral targeting policy can achieve similar gains in policy effectiveness (detecting more corruption).

Our findings are related to several literatures in economics. First, our paper contributes to the literature studying the relationship between corruption and public finance. Many studies emphasize the connection between governmental transfers and public corruption: Brollo et al. (2013) focus on the Brazilian setting, while De Angelis et al. (2020) study the impact of European funds on rent-seeking activity. Another set of papers analyze the extent to which corruption originates from public spending (Hessami, 2014; Cheol and Mikesell, 2018), and there is evidence that policies that constrain public expenditure may reduce corruption (Daniele and Giommoni, 2020). Further, other works attend to the link between public procurement and rent-seeking (Conley and Decarolis, 2016; Coviello and Gagliarducci, 2017; Decarolis and Giorgiantonio, 2020).[2] Our results build on this work with a broader view on the entire budget, rather than focusing on single elements. Our method provides a proof of concept that these issues can be analyzed using machine predictions on corruption, in addition to measurements produced by human audits.

In addition, we add to the existing evidence on the efficacy of auditing programs on corruption in developing countries. Olken (2007) set up an RCT with villages in Indonesia and find that the introduction of the auditing scheme decreased corruption. Bobonis et al. (2016), studying municipalities from Puerto Rico, show that audits effectively reduce corruption and rent-seeking activities by enhancing electoral accountability in the short run, but these effects do not last. In the Brazilian context, Zamboni and Litschig (2018) show that increasing the probability of being audited could reduce corruption, while Avis et al. (2018) find that the implementation of an audit in a specific city reduces

[2] Notably, Mexico recently introduced a corruption risk index in order to tackle corruption in public procurement (link).

future corruption levels in that city. Our event study analysis confirms the latter results, and we are the first to show the dynamics of this effect. Moreover, we find that the effect is particularly strong in cities where corruption is actually detected.[3]

Methodologically, our study adds to the emerging literature in economics applying machine learning techniques to overcome limitations of standard datasets (Athey, 2018). The most established technique in empirical work is to use unsupervised learning to analyze high-dimensional data. For example, Hansen et al. (2018) use Latent Dirichlet Allocation (an unsupervised machine learning algorithm) to measure topics and diversity of discussion in Central Bank committee meeting transcripts. Bandiera et al. (2020) use a similar method to detect CEO behavioral types from their work activity records. Like these papers, we use machine learning to extract relevant dimensions from high-dimensional data. However, we use supervised learning (rather than unsupervised learning) to construct these measurements. This approach is related to several papers in political economy that have used supervised learning to extract measures of partisanship from text, to show (for example) changes in polarization over time or to analyze media influence (Gentzkow and Shapiro, 2010; Ash et al., 2017; Gentzkow et al., 2019; Widmer et al., 2020).

At the intersection of machine learning and development economics, several papers have applied machine learning methods to detect corruption. The closest paper is Colonnelli et al. (2019), who also predict the results of corruption audits in Brazilian municipalities but focusing on non-budget variables (private sector activity, financial development, and human capital measures). Besides our focus on fiscal factors, the main difference in our paper is to use the measure of corruption for an empirical analysis and policy simulation analysis.[4]

Our use of machine learning to guide auditing is most relevant to the literature on AI-powered policy design (Kleinberg et al., 2015; Athey, 2018; Knaus et al., 2018; Athey and Wager, 2021). For example, Kleinberg et al. (2018) show how an algorithm can support the decisions of judges on pre-trial bail release, by identifying which offenders should be

[3] Our study also contributes to the body of work on corruption and politics in Brazil. For instance, Ferraz and Finan (2008) show that the disclosure of scandals reduces vote shares for the incumbent. Cavalcanti et al. (2018) emphasize that exposing corrupted incumbents affects the quality of candidates selected by their party to run in the following election.

[4] In addition, López-Iturriaga and Sanz (2018) predict the presence of a corruption case each year in 52 Spanish provinces. More at the micro level, Gallego et al. (2018) predict corruption investigations associated with a sample of 2 million public contracts in Colombia.

denied bail. Correspondingly, we show that machine learning can support government efforts to identify municipalities with suspicious public budgets, where further investigation is warranted. The growing corpus of work in this vein has used machine learning to detect higher-quality teachers (Rockoff et al., 2011), support physician decision-making (Kleinberg et al., 2015; Mullainathan and Obermeyer, 2019), identify restaurants for targeted health inspections (Kang et al., 2013; Glaeser et al., 2016), allocate tax rebates and tax audits (Andini et al., 2018; Battiston et al., 2020), identify crime hotspots (Mohler et al., 2015), assign refugees to their economically optimal locations (Bansak et al., 2018), demarcate areas of the Amazon for protection against deforestation (Assunção et al., 2019), or identify individuals who are most responsive to marketing nudges (Hitsch and Misra, 2018; Knittel and Stolper, 2019). Besides the new setting (corruption policy), we expand on this work in several methodological directions. First, we use model explanation to validate how the model makes its predictions. Second, we validate the empirical relevance of our machine predictions by showing that they respond appropriately as outcomes in causal regressions. Third, we adopt methods from algorithmic fairness (e.g., Rambachan et al., 2020; Kasy and Abebe, 2020) to address potential political biases in the targeted audits.

The paper is organized as follows. In Section 2 we present the institutional setting and the data. Section 3 describes the prediction procedures and model performance results. Section 4 shows how the model predictions can be used in empirical analyses of corruption. Section 5 reports a set of policy simulations for guided audits supported by machine learning. Section 6 concludes.

2. Institutional Background and Data Sources

2.1. Local Government and Budgets

Brazil has a decentralized governance structure composed of 26 states and 5563 municipalities. At the municipal level, the central political authorities are the mayor (*prefeito*) and the city council (*Câmara de Vereadores*), which are directly elected by citizens every four years. Starting from the 1980s, local governments have enjoyed substantial autonomy in public budgeting decisions. They have primary responsibility for the provision of health and education services and municipal transportation and infrastructure. For the most part, these services are funded by upper-level jurisdictions via intergovernmental transfers. Yet, the mayor has autonomy in setting the tax rate for important local taxes, e.g., taxes on buildings and lands (*Imposto sobre a Propriedade Predial e Territorial Urbana* - IPTU), as well as sales taxes on services (*Imposto sobre Serviços*).

We collected the annual budget of all Brazilian municipalities for 2001 through 2012. Building on the previous local public finance literature, we gather detailed information about the categories of expenditure, revenue, active positions (assets), and passive positions (liabilities). These data are publicly available in the Finance Ministry's online database.[5] We downloaded the datasets for each year and cleaned the variables to make them comparable across years.

In the period of our analysis, the budgets were composed of a large number of different categories for each of the four macro-categories. In total, we have 797 accounting variables from the original data source. The expenditure section has the most components, while the passive section has the fewest. Over time, there is an increasing level of detail about the use and sources of local governments' revenue as the budget adapts to changes in legislation (see Appendix Table A1).

2.2. Anti-corruption policy in Brazil

In 2003, the Brazilian government introduced new policies to reduce corruption. In particular, the policymakers behind this agenda were concerned about the misuse of federally transferred funds by local authorities. Thus, a cornerstone of the reform was a system of random audits, in which municipalities are randomly selected to have their fiscal accounts audited for corruption.[6]

The government invested significant planning and resources in these inspections. In particular, random assignment of audits was implemented to ensure fairness in their allocation. In a given audit round, of which there are around four per year, between 50 and 60 municipalities are chosen. Separate lotteries are run for each state (meaning some states are getting slightly more lotteries per municipality than others), and cities with more than 500,000 inhabitants are excluded.

The audits are implemented by officials from *Controladoria Geral da União* (CGU), an independent federal public agency. Every selected municipality is visited by 10 to 15 auditors. Their inspections focus on a list of randomly selected items provided by the

[5] https://www.tesourotransparente.gov.br

[6] Starting in 2016 (after our period of analysis), the policymakers began selecting some municipalities for audit based on a risk score, using previous audit data.

CGU from the sample of federal transfers the municipality received in the previous 3-4 years. They usually spend a couple of weeks in municipal offices collecting information to identify potential mismanagement in the use of public funds. The auditors summarize the presence of irregularities in reports made available to the public within a few months of the inspection.

These audit reports provide detailed information that can be used to create measures of municipal-level corruption (Ferraz and Finan, 2008; Brollo et al., 2013; Zamboni and Litschig, 2018). We use the corruption measures provided by Brollo et al. (2013). These data include several measures for all 1,481 municipalities audited in the first 29 lotteries of the anti-corruption program (i.e., audits from 2003 to 2009). Focusing on a particular mayor's term of office, they compute the share of corrupted resources (i.e., the ratio between the total amount of funds involved in the detected violation and the total amount audited). Our analysis focuses on a binary variable identifying the presence of what the authors call *narrow corruption*, which is restricted to severe irregularities such as illegal procurement, fraud, favoritism, and over-invoicing.[7] On this definition, 42% of audited municipalities at their first audit are found to be corrupt.

For robustness, we have access to an alternative set of corruption variables from Avis et al. (2018). That measure is constructed using a slightly different approach to coding the audit report documents. It is available for a different (but mostly overlapping) set of audits. We find that the two measures are highly correlated (Appendix Figure B3). In Appendix B, we will provide supplementary analysis using this alternative measure of corruption.

2.3. Linked Dataset

We join the corruption outcome with the local budget factors based on the years of the budget that were examined by the auditors. The resulting dataset is at the municipality-year level. We then add data on local demographics, intergovernmental transfers, and political party control. Specifically, we add demographics from the 2000 Brazilian Census, including *mean income*, *share of population employed*, *sector of occupation* (agriculture, industry, commerce, transportation, services and public administration), *share with college education*, *poverty rate*, and *Gini Coefficient of income*. Federal-to-municipal

[7] In addition, they define a measure of *broad corruption*, which also includes inconsistencies that could be linked to government mismanagement, but not intentional misuse. This concept of corruption is less useful because it is so widespread: 76% of audited municipalities have broad corruption.

Table 1: Summary Statistics

Variable	Mean	Std. Dev.	Min	Max	N
True corruption (term)					
Main Labels from Brollo et al. (2013)	0.424	0.494	0.000	1.000	2087
Alternative Labels from Avis et al. (2018)	0.238	0.426	0.000	1.000	1604
Budget categories (year)					
Total assets	1184.5	2272.3	-19.5	300364.4	64933
Financial assets	216.1	473.2	-3153.9	40128.9	64933
Cash	3.5	35.0	-1607.5	5017.8	64933
Financial liabilities	136.1	272.7	-3023.7	25282.2	64933
Taxes	8.9	15.9	0.0	781.5	64933
Revenues from municipal properties	21.4	127.5	-29.2	27471.4	64933
Total expenditures	1315.4	1402.5	7.1	179411.6	64933
Capital expenditure	181.2	250.2	0.0	21258.5	64933
Current expenditures	1134.1	1206.1	0.0	159532.0	64933
Budget surplus/deficit	41.0	3339.2	-3743.5	650900.8	64933
Municipal characteristics					
Mean income	593.0	319.8	29.8	3062.5	64933
Agriculture (% employed)	16.9	10.1	0.0	72.3	64933
Industry (% employed)	4.2	4.2	0.0	37.5	64933
Commerce (% employed)	7.5	3.6	0.3	27.8	64933
Transport (% employed)	1.2	0.7	0.0	5.9	64933
Service (% employed)	6.8	2.7	0.3	19.3	64933
Public administration (% employed)	2.1	1.2	0.1	16.1	64933
Employed population	38.4	8.5	9.7	79.8	64933
Graduated people	1.2	1.3	0.0	16.5	64933
Poor population	10.0	8.1	0.3	54.4	64933
Gini coefficient	0.6	0.1	0.3	0.9	64933
Audit reports mentions					
Number of mentions	1284.8	11014.3	0.0	190610.0	709

Notes: Main Labels from Brollo et al. (2013) captures the binary variable measuring the presence of corruption according to Brollo et al. (2013) (*narrow corruption* variable). *Alternative Labels from Avis et al. (2018)* captures the binary variable measuring the presence of corruption according to Avis et al. (2018). All budget variables are expressed in per-capita terms. The municipal characteristics are drawn from the 2000 Brazilian census. *Mean income* captures the average income of the working population, the variables *Agriculture*, *Industry*, *Commerce*, *Transport*, *Service* and *Public administration* capture the population employed in a specific sector. *Employed population* measures the fraction of employed population, *Graduated people* is expressed in percentage points and *Poor population* is the fraction of poor population. *Number of mentions* is the number of times each budget item is mentioned in the text of the audits reports.

revenue transfers data come from the Brazilian National Treasury (*Tesouro Nacional*), while population data come from the Brazilian Institute of Geography and Statistics (IBGE). Finally, we collected information about the mayor party affiliations in the 2000, 2004, and 2008 elections. Summary statistics on these variables are reported in Table 1.

3. Predicting Corruption from Budget Data

Our goal is to take the information in the municipal budget and learn a prediction function to provide a probability that a given municipality is fiscally corrupt. To that end, this section outlines how we build our dataset and machine learning model to form those predictions. We evaluate and interpret the predictive model, and then apply it to all municipalities in Brazil for use in the subsequent analysis.

3.1. Corruption Prediction Dataset

Our data consists of budget predictors and corruption outcomes. For the budget features X, the only pre-processing step taken is imputing missing values with the mean value for the associated variable.[8] The resulting matrix X of budget factors has 797 columns, corresponding to the budget fields, and rows corresponding to each municipality and year.

The corruption label $Y \in \{0, 1\}$ equals one for years of examined budgets where the audits found corruption, and equals zero for years in which the audits did not find corruption. For the model training, any municipality-years that were not audited are excluded because we do not have any labels. For the set of municipalities that were randomly audited twice, the second audit is excluded during model training because the samples are not comparable. However, we will use those twice-audited municipalities later on to help with validation.

3.2. Machine Learning Approach

We face a binary classification task. We want to learn a conditional expectation function $Y(X)$ that provides a predicted probability that a municipality is corrupt based on the publicly observed budget features. Classical statistical models, like probit or logit, do not extrapolate well to new datasets because they tend to over-fit the training sample

[8] We got similar results when experimenting with additional pre-processing steps, including adding missing indicator variables, standardizing variables, transforming variables as per capita, or imputing missing variables in different ways.

(e.g., Hastie et al., 2009). The contribution of machine learning tools, now becoming widespread in economics (e.g., Belloni et al., 2014; Mullainathan and Spiess, 2017; Athey, 2018), is to address the over-fitting problem and provide robust out-of-sample prediction with high-dimensional datasets.

Researchers and policymakers now have access to a variety of machine learning tools for solving binary classification tasks. For example, one of the baseline models that we will use below is penalized logistic regression. This model is very similar to the binary logit, which learns a set of linear coefficients on X, sums them, and puts them through a sigmoid transformation to obtain a probability for Y between zero and one. What is new is a penalty term, which adds an additional cost to the training objective that penalizes larger coefficients. The penalty addresses overfitting and helps the model predict better in held-out test set data. During the training process the strength of the penalty is calibrated in a process called cross-validation, where the training data is split up and the out-of-sample performance of different penalties is evaluated. Then the best model is taken to the unseen test set for a clean evaluation and for any downstream tasks.

A state-of-the-art model for binary classification using high-dimensional tabular datasets is gradient boosted trees (Friedman, 2001; Hastie et al., 2009). Gradient boosting models consist of an ensemble of decision trees that "vote" on the predicted outcome. Each decision tree iteratively selects informative variables (e.g., property taxes), splits on a value of that variable (e.g., $x > 100$) to better predict the outcome, branches off for additional splitting, and so on, until reaching a terminal node and an associated prediction ($\hat{Y} = 0$ or $\hat{Y} = 1$). With gradient boosting, additional layers of trees are gradually added during the training process to fit residuals and fix errors in the initial layers. This iterative growth approach tends to perform better than other ensemble methods, such as random forests, which grow trees in parallel.

More specifically, we train a gradient boosted classifier using the implementation from the python package XGBoost (Chen and Guestrin, 2016). Feurer et al. (2018) systematically compared XGBoost to many other classifiers, including a sophisticated automated ML system, and found that XGBoost consistently performed best on our type of machine learning task. We used cross-validation grid search to tune hyperparameters, which include the learning rate, L1 and L2 regularization penalties on the learned parameter weights, the max depth of the constituent decision trees, and an additional regularization constraint specifying a minimum threshold for the size of the decision tree terminal nodes. Appendix Table A2 shows the selected values for these hyperparameters across each of five different training folds.

In the next subsection on model performance, we compare XGBoost to a number of baselines. First, as the weakest baseline, we guess the modal category (not corrupt). Second, we train ordinary least squares (OLS), or non-penalized linear regression, dropping multi-collinear predictors. Third, LASSO, perhaps the most familiar machine learning model to economists (e.g Belloni et al., 2014), is a linear regression model but adds an L1 penalty that penalizes larger coefficients and outputs a sparse model. For both OLS and LASSO, the predicted probabilities for Y might be below zero or above one, but a decision threshold of 0.5 is used for assigning a predicted label. Finally, as the strongest alternative baseline, we use penalized logistic regression, a linear classifier with a sigmoid transformation and elastic net penalty (that is, both a LASSO (L1) and a ridge (L2) penalty). For LASSO and Logistic, the penalty is selected by cross-validation grid search in the training set. All three of these linear baselines are implemented using the python package scikit-learn (Pedregosa et al., 2011).

We train and evaluate models using nested cross-validation, which works as follows: First, we randomly split the sample of audited municipalities into five different sets. Next, we train five separate models using each time four different subsets (80% of the sample) and take the tuned models to get performance metrics in the test set (the remaining 20 % of the sample, which also rotates). Each time, we tune the hyperparameters in the training set using five-fold cross-validation. In each fold, early stopping is used (with patience of ten training epochs) to stop training when the model begins to over-fit the training set.

The nested approach provides five sets of predictions for each model. These model outputs can be averaged to produce a single predicted probability of corruption for a given data point. In the model evaluation section, we report the mean and standard error of performance metrics across these five models. In the policy simulations, we will use the multiple predictions to assess the importance of sampling variability in the predictions when making decisions based on them.

The XGBoost model training algorithm produces complex forests of decision trees. As reported in Appendix Table A2, the fitted forests consist of between 46 and 97 trees, with an average of 71 trees. Each constituent tree can take up to 10 sequential variable splits before a terminal node, leading to fully grown ensembles containing 12,108 nodes on average (171 variable splits per tree). The numerous variable splits learned during the training process demonstrate the model's capacity to detect extremely subtle budget

Table 2: Out-of-Sample Metrics for Predicting Corruption

	Guessing (1)	OLS (2)	Lasso (3)	Logistic (4)	XGBoost (5)
Accuracy	0.580	0.476	0.474	0.560	0.723
		(0.022)	(0.022)	(0.022)	(0.012)
AUC-ROC		0.487	0.507	0.568	0.777
		(0.016)	(0.012)	(0.016)	(0.013)
F1	0.000	0.480	0.538	0.545	0.632
		(0.031)	(0.050)	(0.054)	(0.018)

Notes: Column (1) reports the performance metrics (by row) from naively guessing the modal category ("not corrupt"). Columns (2) to (5) report the mean and standard error (in parentheses) for the indicated test-set performance metrics (by row) across the five model runs, produced using separate training-set folds. Columns indicate the machine learning model used.

patterns related to corruption.

3.3. Model Performance

We evaluate our set of models by their scores on a set of standard classification metrics in the held-out test data. These metrics, reported by row in Table 2 Panel A, describe how well a model trained on the budget accounts can replicate the auditing agency's judgments about fiscal corruption. First, the most straightforward metric is accuracy, which gives the proportion of test-set observations for which the machine-predicted label matches the true label. A naive guessing model that chooses the modal category (not corrupt), shown as a weak baseline in Column 1, would obtain accuracy = 0.58. Second, we report AUC-ROC (area under the receiver operator characteristic curve), another standard metric in binary classification. AUC-ROC, which takes values between 0.5 (random guessing) and 1.0 (perfect accuracy), can be interpreted as the probability that a randomly sampled corrupt municipality is ranked more highly by predicted probability of corruption than a randomly sampled non-corrupt municipality. Because AUC-ROC requires a ranking of predicted probabilities, it is undefined for naive guessing (Column 1). Third, we report F1 for the corrupt class, defined as the harmonic mean of precision (proportion true corrupt within the set predicted corrupt) and recall (proportion predicted corrupt within the set true corrupt). F1, ranging from 0.0 (guessing the modal category) and 1.0 (perfect accuracy for the corrupt class), penalizes both false positives and false negatives.

Columns 2 through 4 of Table 2 show the predictive performance for a set of additional baselines: OLS (non-penalized linear regression), Lasso (L1-penalized linear regression), and logistic regression (with L2 penalty), respectively. Each table cell reports the average test-set performance metric across the five cross-validated model runs, with the standard error of the mean in parentheses. OLS and Lasso (Columns 2 and 3) significantly improve on the zero F1 from guessing (Column 1) at the expense of lower test-set accuracy. Logistic regression (Column 3) is significantly better than the linear models on all metrics, yet still similar in overall accuracy to guessing.

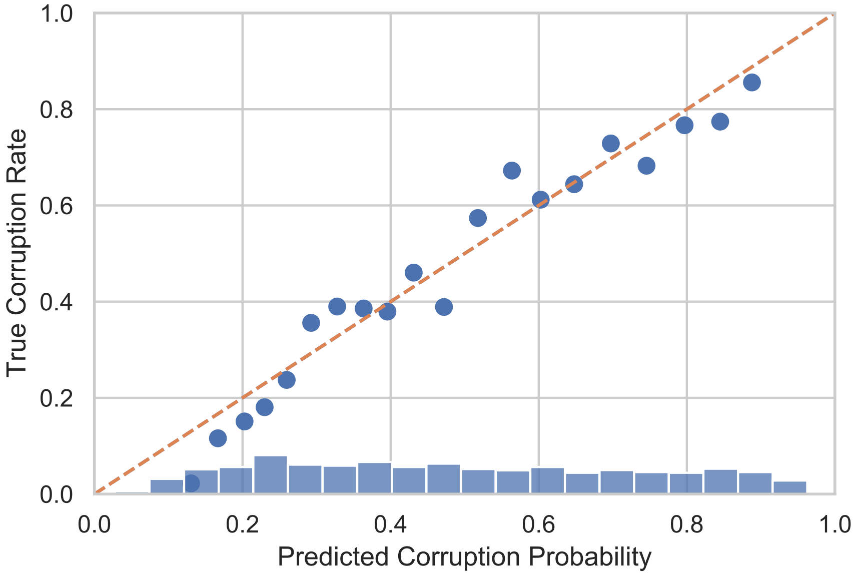
With XGBoost (Column 5), meanwhile, we report a substantial lift in all metrics over all baselines. The average test set accuracy for the predictions across five nested folds is 0.723, with the minimum accuracy being 0.692 and the maximum 0.755 across folds. For AUC-ROC, the average is 0.777 (with min = 0.743 and max = 0.809), while for F1, the average is 0.632 (min = 0.584, max = 0.675).[9] The boost in performance suggests a nonlinear, interactional relationship among the predictors that the tree ensemble is better able to learn.

Beyond the binary classification metrics, another relevant metric for model performance is the calibration of the predicted class probabilities. That is, do the predicted corruption probabilities faithfully rank the municipalities by risk and reproduce the correct distribution of corruption rates? Figure 1 shows that the model is well-calibrated: In each of 20 bins in the held-out test set, the predicted corruption probability in that bin corresponds closely to the true corruption rate of test-set observations in that bin. Appendix Figure A1 provides additional plots, showing good calibration (a) across all five models trained on different training folds, (b) for alternative train/test sampling, and (c) for an alternative measure of corruption.

Appendices A and B report additional evaluations of the prediction task. First, to help visualize the distribution of predictions, Appendix Table A3 shows the confusion matrices for the test-set predictions. For XGBoost, we can see good precision and recall across categories. The confusion matrices for OLS, LASSO, and logistic regression show that the linear models tend to produce many false positives (not-corrupt municipalities

[9] The model performance is similar to that in other recent papers using machine learning for economic analysis of human decisions. For example, the gradient boosted ensemble in Kleinberg et al. (2018) trained to predict criminal recidivism obtains an AUC-ROC of 0.707. The model in Mullainathan and Obermeyer (2019) trained to predict a cardiac medical intervention obtains AUC-ROC = 0.731.

Figure 1: Predicted Probabilities in Test Set are Well-Calibrated to True Corruption Rates



Notes: Calibration plot showing true corruption rates (blue marks on the vertical axis), binned by predicted corruption probability (horizontal axis), in held-out test set. Blue histogram shows the density of the predicted corruption probability. Dashed 45-degree line (in orange) demarcates perfect calibration.

are often labeled as corrupt), yet fewer false negatives. Note that in our performance evaluation so far, we have treated false positives and false negatives symmetrically in terms of their policy costs. But from the perspective of an auditing cost, it is not clear that they are substitutable. In practice, if false positives are valued different from false negatives, the XGBoost model can be calibrated to account for that.

Second, we focus on the municipalities that have been audited twice and see if our prediction model can reproduce within-municipality changes in corruption over time. To that end, we regress the change in true corruption against the change in predicted corruption, adjusting for audit year fixed effects and demographic characteristics. Appendix Figure A2 shows that there is a significant positive relationship in this regression. This within-municipality validation is important for the usefulness of our measure in empirical tasks, where one would like to be able to examine changes in corruption over time.

Third, in Appendix Table B4 we report performance metrics with an alternative sampling approach and with an alternative corruption outcome. In Columns 1-3, we apply random splits between training and test set by municipality, instead of by municipality-year, which allows us to compare the model performance using budget factors, fixed demographic factors, or both. The alternative sampling approach obtains comparable performance to our baseline model and shows that a model trained using just demographic information is less accurate than a model using budget information. Finally, in Columns 4-7, we show the model performance in predicting the alternative corruption variable from Avis et al. (2018), obtaining even higher accuracy than the baseline corruption variable (AUC-ROC = 0.903, s.e. = 0.009 for XGBoost).

3.4. Interpreting the Predictions

Gradient boosted machines, like all ensembles and other sophisticated machine learning algorithms, are black boxes. At the end of model training, we have a dense forest of decision trees. With 797 variables being input into those trees, and thousands of splitting nodes within the forest (Appendix Table A2), it is difficult to tell how the model is making its predictions. In this subsection we use model explanation methods to better understand how the model works.

The applied machine learning literature has discovered an advantage of gradient boosted machines that compensates for their basic lack of interpretability (e.g., Hastie et al., 2009; Molnar, 2020). One can rank the input variables by their *feature importance*, computed as the number of times the model "uses" that variable in the sense that one of the constituent decision trees splits on it. Moreover, the important features can be seen as *pivotal* in the sense that they are the most useful variables for predicting the outcome, even among clusters of highly collinear predictors.

Here we use the feature importance ranking to get some insight into how our corruption detection model makes its predictions. After model training, we have feature importance scores for each of the five cross-validated models. We average the scores across folds and then rank the most important features. From the feature importance scores, we learn immediately that our dataset contains many noise predictors. Out of the 797 variables input to the ensemble, 351 are ignored by the ensemble across all five folds.

Still, that leaves 446 variables that the tree ensemble finds useful for predicting corruption, implying a high-dimensional learned function. Appendix Table C5 lists the set of most important predictors, ranked by the feature importance score. Some of the top-ranked factors have been mentioned in the previous literature as being related to corruption, including: (3) expenditures on transportation (Hessami, 2014), (5) municipal debt (Liu et al., 2017), (7) real estate and construction (Kyriacou et al., 2015), and (12) public health services (Machoski and de Araujo, 2020).

For these and the other important variables, one could think of (perhaps many) ways that they could contribute to corruption. Many of these stories would be unsatisfying, however, and they would likely be incorrect because the feature importance ranking does not identify direct links between a budget factor and corrupt behavior. The important features could be either positively or negatively correlated with the predicted corruption. Moreover, they could be important through a non-linear relation, or through interactions with other variables.

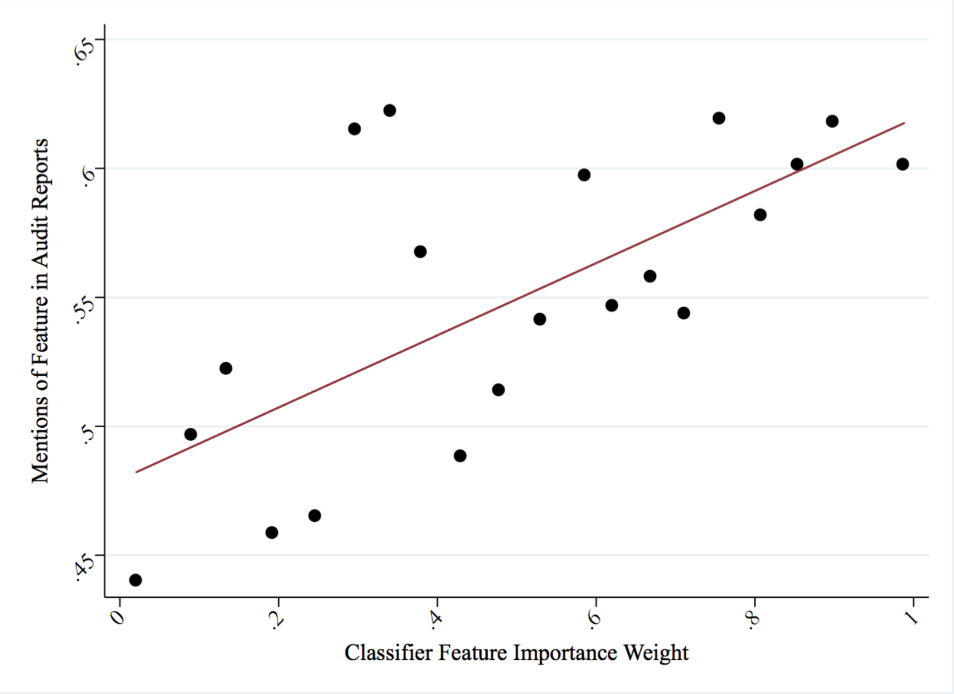
Appendix C.1 provides some supporting analysis to illustrate the non-monotonicities learned by the tree ensemble. Starting with the full dataset, we perturb each feature individually and record the predicted change in corruption risk. Looking at the distribution of the changes, we can infer a highly non-linear, interacted functional form. Of the 446 informative variables, only 39 variables have a monotonic relationship with corruption risk. In the set of important features (according to the feature importance score), none of them have a monotonic relationship with corruption risk. Instead, perturbing budget features could have either a positive or negative relation to predicted corruption depending on the status quo values of the variables.

Given these pivotal non-linearities, a further qualitative discussion of the highest-ranked features would obtain limited insights. As a more comprehensive alternative, we ask whether the corruption-related features identified by the gradient boosted machine are also identified as corruption-related by the Brazilian auditing agency. To do that, we look for mentions of these items in the best available place – the text of the published audit reports.

As described in Appendix C.2, we obtained the text of over two thousand reports for the time period of our analysis (containing over 55 million words) and counted the mentions of each budget item in the corpus of reports. Appendix Table C6 shows the budget components with the highest number of mentions in the reports. We produced a dataset at the feature variable level, containing the percentile rank in the model feature importance score and the percentile rank in the audit-report mention count.

To compare these values, Figure 2 plots the audit mention percentiles against the feature importance percentiles. We see a strong positive relationship that is statistically significant in a univariate regression ($p = .001$). Our classifier, trained on the budget accounts with just corruption labels, identifies as important the same budget features that

Figure 2: Model Feature Importance and Mentions in Audit Report Texts



Notes: Binscatter diagram for percentile that a budget feature appears in the municipal audit reports (vertical axis) against binned percentiles of feature importance weights for each feature (horizontal axis). Pearson's correlation is 0.21. The regression coefficient is 0.139 with $p = .001$ (robust standard errors).

tend to be mentioned in the audit report documents.[10] These validation results support the view that our measure captures activities that are indeed related to corruption.

3.5. Measuring Corruption in Non-Audited Municipalities

An essential contribution of our approach is to measure corruption for all Brazilian municipalities and all years from 2001 to 2012. Using the trained models, we form five predicted corruption probabilities for all observations based on the budget data. In Figure 3 we provide a visualization of the difference between the sample of only audited municipalities (Panel a) and the sample of municipalities that we can analyze when using our predicted measure of corruption (Panel b). The map illustrates quite clearly the additional information produced by the machine learning method. With the machine predictions, we can then analyze corruption in municipalities (and years) regardless of whether they have been audited.

4. Empirical Applications

This section replicates and extends existing evidence from the literature on corruption in Brazil. This exercise has two purposes. On the one hand, it provides checks on the internal validity of our synthetic measure of corruption – that is, we can check whether it responds to causal treatments the same way as auditor-measured corruption. On the other hand, we extend previous results by taking advantage of the larger sample of municipalities and the time variation of our corruption measure.

4.1. Revenue Shocks and Corruption

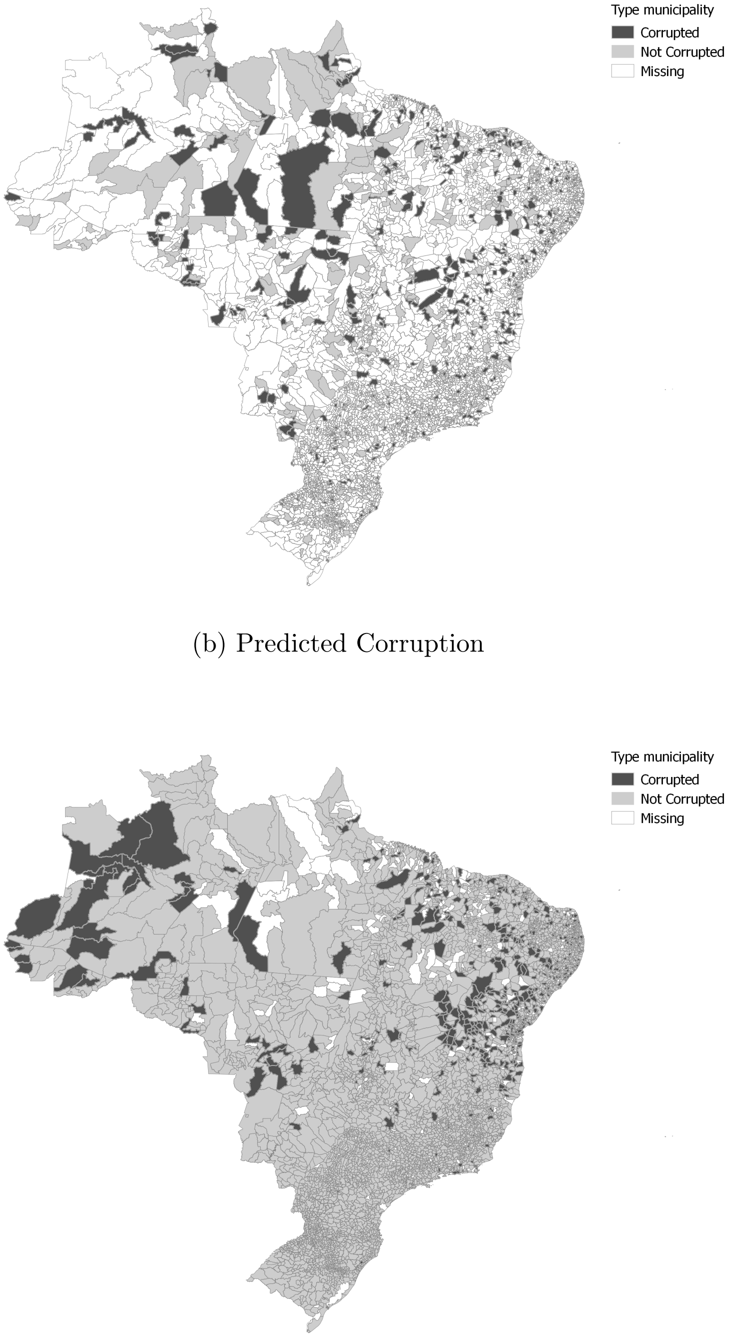
As a first analysis, we use the new synthetic measure of corruption to analyze the effect of revenue shocks on corruption, replicating and extending the findings by Brollo et al. (2013). This paper studies whether a windfall of public revenues can lead to an increase in rent-seeking by the public administration (as measured by a subsequent surge in corruption). They estimate the impact of federal transfers on the occurrence of corruption as detected by the random audits.

Brazilian municipalities receive transfers from the states and from the federal government. Federal transfers are the largest single source of municipal revenues (around 40%

[10] Notably, we obtain similar results if we consider only those model features that tend to have a positive effect in the perturbation analysis discussed in Appendix C.1.

Figure 3: The Geography of (Predicted) Corruption

(a) Actual Corruption



(b) Predicted Corruption

Notes: Actual (Panel a) and predicted (Panel b) corruption by municipality, using budgets from 2004. A municipality is predicted to be corrupted if mean prediction is >0.5.

of the total budget). The amount transferred through this *FPM* program (*Fundo de Participação dos Municipios*) depends on exogenous population thresholds, where municipalities in the same state and in a given population bracket receive the same amount of resources.[11] Due to imperfect compliance, the statutorily prescribed transfers do not perfectly determine the amounts actually transferred.[12] Thus, Brollo et al. (2013) use a fuzzy regression discontinuity design methodology, instrumenting actual transfers (τ_i) with prescribed transfers ($\hat{\tau}_i$).

Formally, we have the first stage

$$\tau_i = g(P_i) + \alpha_\tau \hat{\tau}_i + \delta_t + \gamma_p + u_i \tag{1}$$

and reduced form

$$y_i = g(P_i) + \alpha_y \hat{\tau}_i + \delta_t + \gamma_p + \eta_i \tag{2}$$

where y_i is corruption, $g(\cdot)$ is a high order polynomial in P_i (the population of city i), δ_t contains time fixed effects, γ_p contains state fixed effects, and u_i and η_i are the error terms. The coefficients α_τ and α_y capture the effects of prescribed transfers on actual transfers and (predicted) corruption, respectively. For the two-stage-least squares analysis, we estimate the second stage

$$y_i = g(P_i) + \beta_y \tau_i + \delta_t + \gamma_p + \epsilon_i \tag{3}$$

where prescribed transfers $\hat{\tau}_i$ are used as an instrument for actual transfers τ_i and all other terms are defined as above. The coefficient β_y captures the causal effect of actual transfers on (predicted) corruption. For inference, standard errors are clustered by municipality.[13]

[11] Appendix Table D7 shows these coefficients and the corresponding population brackets: Following Brollo et al. (2013), we focus on the initial seven brackets and restrict the sample to cities with a population below 50,940. This sample represents about 90 percent of Brazilian municipalities. Furthermore, we follow the approach of Brollo et al. (2013) and restrict the sample, for the sake of symmetry, to municipalities from 3,396 below the first threshold to 6,792 above the seventh threshold. More precisely, the amount of revenues received by municipality i in state k follows the allocation mechanism: $FPM_i^k = \frac{FPM_k \lambda_i}{\sum_{i \in k} \lambda_i}$, where FPM_k is the total amount allocated in state k and λ_i is the municipality-specific coefficient, as shown in Table D7.

[12] This imperfect compliance is due to many factors (*e.g.,* municipalities splitting, manipulation in population figures). See Brollo et al. (2013).

[13] See Brollo et al. (2013) for a detailed discussion and testing of the econometric assumptions in this setting. Note that our measure of corruption is predicted from a machine learning model, and thus is measured with error. Because we use it as a dependent variable, that is likely to make standard errors larger without biasing the coefficient.

Table 3: Effect of Revenue Shocks on (Predicted) Corruption

	Audited cities (1)	All cities (2)	Non-audited cities (3)
Panel A. First Stage			
Theoretical transfers	0.6805*** (0.0205)	0.6909*** (0.0233)	0.6996*** (0.0230)
Panel B. Reduced Form			
Theoretical transfers	0.0040*** (0.0009)	0.0041*** (0.0003)	0.0040*** (0.0003)
Panel C. 2SLS			
Actual transfers	0.0058*** (0.0013)	0.0059*** (0.0005)	0.0057*** (0.0005)
N. Observations	1115	5808	4693

Notes: Effects of FPM transfers on (predicted) corruption measures. Panel A reports the estimates of the first-stage analysis, the dependent variable is *actual transfers*. Panel B reports the estimates of reduced form analysis, the dependent variable is *predicted corruption*. Panel C reports the estimates of the 2sls estimates, the dependent variable is *predicted corruption* and *actual transfers* is instrumented with *theoretical transfers*. Column headings indicate the sample of municipalities included. All regressions controls for a third-order polynomial in normalized population size, term dummies, and macro-region dummies. Robust standard errors clustered at the municipal level are in parentheses: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Our data cover the two mayoral terms, January 2001–December 2004 and January 2005–December 2008. While Brollo et al. (2013) focus only on municipalities that received an audit, our dataset allows us to analyze a larger and potentially more representative sample of cities. Therefore, our exercise is also providing a test for the external validity of their results.[14]

Table 3 reports the results for the regression analysis. For each regression specification, we estimate the results in three samples: cities that have received an audit (column 1) similarly to Brollo et al. (2013), all cities (column 2), and cities that have

[14] For the sake of brevity we only replicate the analysis on the overall effect, omitting the threshold-specific analysis. Appendix Table D8 shows the descriptive statistics by population bracket. Brazilian municipalities in our sample receive, on average, \$3.3M BRL (about \$610K USD), while prescribed transfers are somewhat higher at \$3.7M BRL (about \$680K USD). The average level of (predicted) corruption is around 0.5 and its level does not change significantly as we move to larger cities.

never been audited (column 3). In Panel A, we find a strong first-stage effect (Equation (1)), showing that prescribed transfers positively affect actual transfers for all samples considered. In Panels B and C, we find positive and significant coefficients when estimating the reduced-form (Equation (2)) as well as two-stage-least-squares (Equation (3)). Varying the sample of interest does not significantly alter the size of the coefficient and the level of precision is stable. Notably, the magnitude of the standardized reduced-form coefficient is about four-fifths the size of that estimated by Brollo et al. (2013), and our 95% confidence interval contains the original coefficient. Thus even the magnitudes of empirical estimates using machine-learning-measured corruption are comparable to those using auditor-measured corruption.

we conducted a series of checks to probe the robustness of these results. First, to show that our results are not driven just by a larger sample, we replicate the main analysis on four random samples of 1,115 municipalities, the sample size of the original analysis by Brollo et al. (2013) (Appendix Table D9). The coefficients show some variation, but they are always positive and statistically significant. Second, we show that the instrument is not correlated with the error of the prediction model, defined as the difference between the true corruption level and the predicted one (p-value=0.212). This null is helpful because it suggests that the machine learning model's errors are not responding to the instrument – that is, the correlated factors besides corruption contributing to our prediction are not affected directly by revenue transfer shocks. Thus, using our model predictions as the outcome will still satisfy the exclusion restriction. Third, we show in Appendix Table D10 Column 1 that there is no revenue-shock effect on a corruption prediction formed with a model trained on time-invariant municipal demographic characteristics (similar to Colonnelli et al.'s (2019)). This placebo test is reassuring because the model trained on demographics does not contain budget information, so revenue shocks should not have an effect. Since our budget-trained model does have an effect, the placebo test provides additional confidence that our main model is not forming corruption predictions based on spurious correlations with demographics. Fourth, we formed predictions from our baseline model while permuting randomly the FPM transfer variable, which could be mechanically shifted by the revenue shocks instrument. The effect of revenue shocks is the same (Appendix Table D10 Column 2).

4.2. Effect of Audits on Corruption

The next empirical application uses our predicted measure of corruption to analyze the effect of auditing on subsequent corruption in an event study framework. This analysis complements Avis et al. (2018), who explore the same research question using the set of Brazilian municipalities that were (by random draw) audited twice in a cross-sectional setup. With our new measure of predicted corruption, we can overcome the data limitations of Avis et al. and extend their results. First, because of the longitudinal nature of our dataset, we can capture dynamic effects. Second, we can condition our estimates on pre-audit levels of corruption. Third, our effects are identified by a relatively larger sample of municipalities that got audited only once (rather than twice).[15]

Using the annual corruption prediction y_{ist} in municipality i of state s at year t, we take a standard event study approach and estimate the within-municipality effects of a (randomly assigned) audit. Let D_{ist}^k be a dummy variable for k years before and after an audit. We estimate

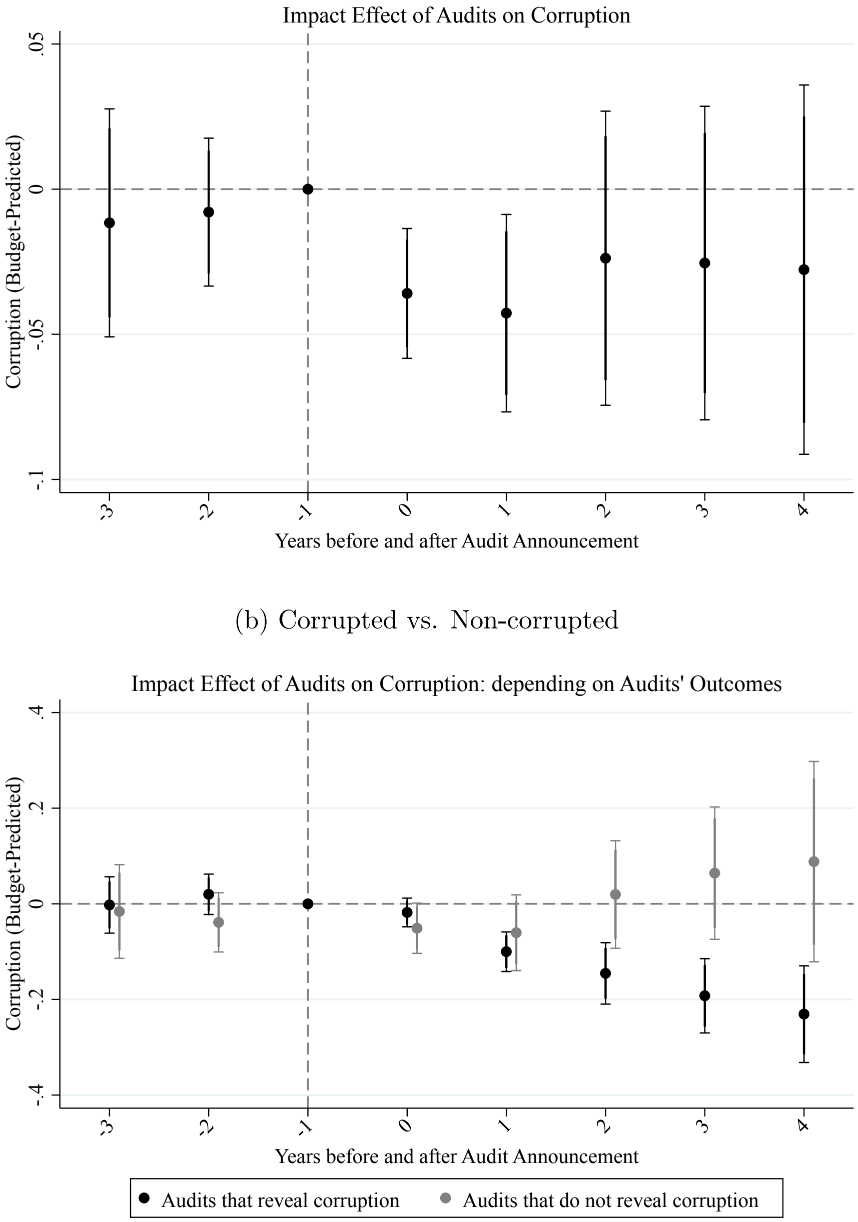
$$y_{ist} = \sum_{k=-3, k \neq -1}^{5} \beta_k D_{ist}^k + \delta_i + \lambda_t + W_{ist}'\phi + \epsilon_{ist} \tag{4}$$

where we have municipality fixed effects δ_i, year fixed effects λ_t and other controls W_{ist}, which in particular includes dummy variables indicating periods distant from when the audit took place. Because $k \neq -1$ (the year before the audit), the β_k's estimate the dynamic effects relative to the year before the audit. The identifying assumption hinges on randomness in the timing of selection into the audit program. We cluster standard errors by state, although results are the same with clustering by municipality. The sample includes 1,479 municipalities that have received an audit in the time period under analysis.

We graphically report estimates for Equation (4) in Figure 4 Panel (a), with the numerical estimates reported in Appendix Table E11. We can see that already in the year of the audit ($k = 0$), there is a sharp and statistically significant drop in predicted corruption. This effect persists for another year before becoming weaker and not statistically significant. Meanwhile, as expected given the random assignment due to the

[15] Bobonis et al. (2016) study a similar research question in Puerto Rican municipalities. The authors focus on (non random) audits of municipal accounts, finding that audits do not persistently reduce corruption in that case.

Figure 4: Dynamic Effect of Audits on Fiscal Corruption



(a) All municipalities

(b) Corrupted vs. Non-corrupted

Notes: Event study estimates for dynamic effect of audits on budget-predicted corruption. Error spikes give 95% confidence intervals, with standard error clustered by state. Top panel: all audits; bottom panel: audits that found corruption (in black); audits that did not find corruption (in grey). For the analysis on all audits leads are jointly insignificant (p-value=0.908) and lags are jointly significant (p-value=0.003). For the analysis on audits that found corruption leads are jointly insignificant (p-value=0.151) and lags are jointly significant (p-value=0.0003).

lottery, there is no statistically significant effect in the pre-announcement years.

Panel (b) reports event-study effects for the subsets of audits that find clear corruption (black points) and those that do not find corruption at all (grey points).[16] These trends look quite different. When corruption is discovered (black points), there is a much larger negative effect ranging between -1.7% and -25.8%, which is sizeable if compared with the magnitude of the treatment mean of 55.8%. The effect is persistent across subsequent years. In contrast, when the audit does not find any corruption or irregularities (grey points), there is no effect.

We report supporting analyses in Appendix E. First, we check whether the main results may be explained by post-audit budget adjustments that might mechanically take place when a municipality is found to be corrupted. In particular, we might be picking up mechanical changes due to financial penalties imposed on these municipalities. We show that this is not the case: The main results do not change if we control for total spending (per-capita) in the main regression specification, and we show that the occurrence of an audit does not affect future levels of municipal expenditures (per-capita). Second, we test the political accountability channel by checking whether the effect of the audit is stronger when local political competition is high. We find that the answer is yes: In cities where the mayor has been elected with a small margin of victory, the impact of the audit is stronger.

4.3. Discussion

These empirical exercises accomplish three goals. First, we provide evidence on the external validity of the previous work by Brollo et al. (2013) and Avis et al. (2018). Those previous results were obtained with relatively small sample sizes and for a minority of municipalities. Both sets of findings generalize to the full set of municipalities.

Second, replicating previous results provides additional validation for the use of our machine-prediction-based measure of corruption in empirical work. As highlighted especially in the Brollo et al. (2013) analysis, the empirical estimates using our machine-predicted measurements are comparable qualitatively and quantitatively to those produced using the actual audit outcomes. Together with the validations from above (in terms of capturing within-municipality changes in corruption, and with the important

[16] The former group includes those cities in which the audits discovered a positive amount of corruption (measured with the variable narrow corruption), while the latter group includes those municipalities in which the audit did not find any type of corruption.

features being mentioned in the audit reports), we gain confidence about the usefulness of our measure for empirical tasks.

To further refine this point, the empirical replication exercise provides insight into what our machine learning model is capturing, where one could consider two basic interpretations. First, it could be that corrupt municipal governments make budgeting decisions that enable corruption, such as putting slack in the right places, and our model detects that. A second explanation is that there are exogenous conditions that make it easier to engage in corruption – such as the presence of many infrastructure projects – and our model picks up budgetary responses to those conditions. The replication analysis, showing that predicted corruption responds the same way to treatments (revenue shocks and audits) as true corruption, is consistent with the first explanation. If the model were picking up (mostly) non-budget factors, then we would see no response to the causal treatments.

Third, to further highlight the empirical usefulness of the machine predictions, we have now shown that they can be used for further analysis and to provide new and relevant findings. In the event-study analysis, we go beyond Avis et al. (2018) and show more granular effects of audits on corruption, by looking over time and by the outcome of the audit. These additional insights were not available with the existing data based only on audits, given the small sample size.

5. Using Machine Learning to Guide Audit Policy

Besides extending datasets for empirical analysis, our machine predictions for corruption risk can also be used to guide policymakers. This section outlines a policy simulation for how corruption policy could be supported. We start with a baseline targeting policy based on predicted corruption risk, showing that targeted audits can detect more corruption than random audits. Second, we consider the issue of political bias in the risk scoring algorithm towards different mayor party affiliations, and analyze the performance of a politically neutral targeting policy. Third, we discuss additional caveats and complications with implementing a targeted audit system.

5.1. Targeted Audit Policy

To set the stage for targeted audits, let's first consider the performance of the status quo random audit system. Recall that there are 5563 municipalities in the dataset. In our data, the agency audited about 203 municipalities per year, with 95 corrupt municipalities detected for the average year of audits. The resulting audit probability (and therefore detection rate, regardless of whether a municipality is corrupt) is roughly 0.036.

To improve these numbers, let's consider a policy designed to maximize the number of detected corrupt municipalities in a single round of audits. Note that taking a dynamic perspective and trying to detect the most corruption over multiple rounds of audits would involve a different policy. Similarly, our baseline approach would not necessarily optimize other objectives, such as maximizing the deterrent effect of the policy. We return to some of these issues below.

We start by ranking the municipalities by corruption risk. That is, we apply the baseline gradient boosting model to the budget data for each municipality i from year t to produce $\hat{y}_{it}$. Then for each year t, we have an ordinal ranking of the municipalities (1 through 5563) by predicted probability of corruption. The proposed policy is to replace random audits with audits targeted by predicted corruption risk. Rather than sampling 203 municipalities uniformly from the distribution, the agency could audit the top 203 with the highest $\hat{y}_{it}$. These are municipalities that have a level of corruption probability higher than 0.868 in the average year.

This policy is illustrated in Figure 5. The diagonal line is at 45 degrees and indicates the predicted corruption rate at any spot in the risk distribution. The horizontal blue dashed line at 0.036 gives the audit probability under random audits, while the vertical green dashed line indicates the average threshold corruption risk (0.868, s.e. = .004) above which municipalities are targeted for audit.[17] The underlying histogram indicates the distribution of the corruption risk predictions, with the top two bins containing the approximately 203 municipalities to be targeted.

We assess the targeting policy by simulating it on our dataset and measuring its effectiveness in terms of detecting and deterring corruption. First, we assess detection by the corruption rate conditional on audit. Second, we assess deterrence by the audit rate conditional on being corrupt. Effectiveness is measured by comparing to the status quo lottery. For completeness, we simulate policies using the whole population of municipalities (treating our corruption predictions as true) as well as limiting to the municipalities in our audited sample (using the results of the true audit for evaluation). We obtain qualitatively identical results with both simulation samples.

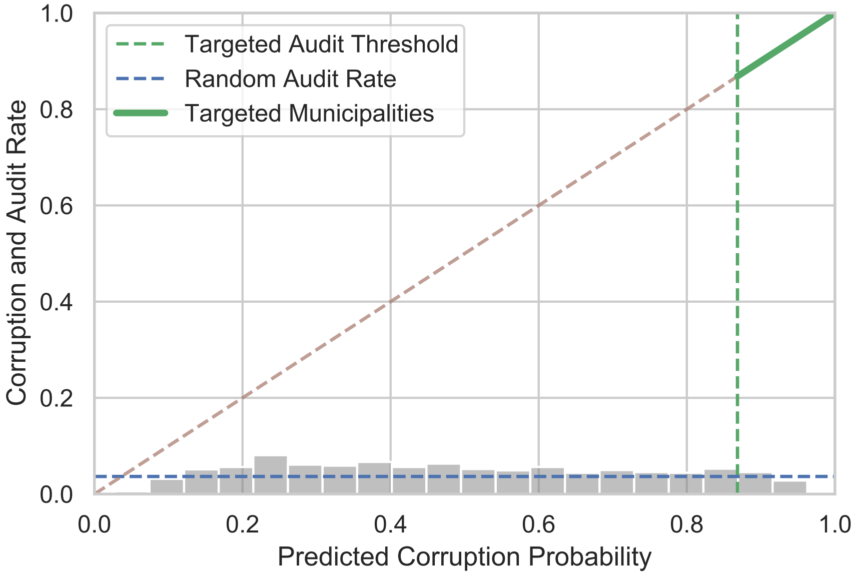
[17] The threshold is an average across years and across the five model folds.

Table 4: Performance Metrics for Targeted Auditing Policies

	Status Quo (Lottery)		Targeted Audits		Fair Targeting
	(1a)	(1b)	(2a)	(2b)	(3)
Evaluation Sample	All (Sim)	Audited	All (Sim)	Audited	Audited
Corruption Rate, if Audited	0.5133	0.4664	0.9186 (0.0033)	0.8563 (0.0163)	0.836 (0.0173)
↪ Ratio over Random Audits			1.7869 (0.0157)	1.836 (0.035)	1.793 (0.037)
Audit Rate, if Corrupt	0.036	0.036	0.0653 (0.0006)	0.0671 (0.0013)	0.0655 (0.001)

Notes: Metrics for comparing the effectiveness of audit policies: random audits (columns 1a and 1b), targeting audits to the municipalities with the highest corruption risk (columns 2a and 2b), or targeting audits with highest corruption with the constraint that all political parties are audited at the same rate (column 3). "Corruption Rate, if Audited" is the share of audited municipalities where narrow corruption is detected, for the respective policy. "Ratio over Random Audits" is the "Corruption Rate, if Audited" value for the indicated policy, divided by that value under random audits. "Audit Rate, if Corrupt" is the expected probability of being audited, if corrupt, under the various policies. In Columns 1a and 2a, "All (Sim)" indicates that statistics are produced using the full sample of municipality-year observations, treating the machine-predicted corruption probabilities as reflecting true realized corruption rates. In columns 1b, 2b, and 3, "Audited" reflects that statistics are produced in the sample of audited municipality-terms, using the true audit results. Thus, Column 1b reports the observed rates in the data. In the other columns, statistics give the mean and standard error (in parentheses) across five values for the predicted corruption risk, produced using different training-set folds.

Figure 5: Targeted Auditing Based on Corruption Risk



Notes: Illustration of targeted auditing policy. Gray histogram shows the density of the predicted corruption probability. Diagonal 45-degree line indicates predicted corruption rate. Blue horizontal dashed line shows the audit probability under random audits (=0.037). Vertical dashed green line shows the average threshold (=.868) above which a municipality is audited based on the targeting rule.

Statistics for evaluating the targeted audit policy are reported in Table 4. First, Columns 1a and 1b describe the lottery baseline. The predicted corruption rate in the whole sample (Column 1a, with "All (Sim)" for "simulated") is 0.5133. This predicted rate is quite close to the observed corruption rate of 0.4664 from the audited sample (Column 1b: "Audited").[18]

Second, Columns 2a and 2b report statistics on the expected outcomes of the targeting policy, where Column 2a (like Column 1a) simulates outcomes using predicted corruption rates from the whole sample, while Column 2b (like Column 1b) evaluates outcomes using true corruption rates from the audited sample.[19] In both simulation samples, we see substantial policy gains. In the full sample (Column 1a), the detected corruption rate of 0.9186 is almost double (1.7869×) that of the status quo policy (0.5133).

[18] Recall that the base rate in Section 3 was .422, rather than 0.4664. The difference here is that we simplify the dataset to have a single observation per audit. In Section 3, the dataset included all fiscal years checked by the audit, which slightly changes the mean corruption rate.

[19] The statistics from Column 3 come from a politically neutral audit targeting policy. We revisit these numbers in the next subsection.

In the audited sample (Column 2b), the gain is almost identical (1.836×).[20] Out of 203 audits, the agency would detect about 168 corrupt municipalities, rather than 95. In turn, a higher corruption detection rate also means a higher audit rate conditional on being corrupt. As seen in the third row, the conditional audit rate under targeting is about 0.065 (Column 2a) or 0.067 (Column 2b), again almost 2x the status quo rate of 0.036.

Overall, targeting makes a big difference in policy effectiveness. For the same number of implemented audits (and presumably the same allocation of government resources), the targeted approach detects about 80% more corrupt municipalities. Because successful audits reduce corruption (see Section 4.2 above), the targeted policy would also reduce the frequency of corrupt activities in Brazil. To achieve the same number of corruption detections as the status quo policy (95 municipalities), only 111 targeted audits are needed, down from 203 random audits. This decrease of 45%, or 91 audits per year, could imply a significant reduction in administrative expenditures.

To check robustness of these results, Appendix Table F12 reports analogous statistics to Table 4 for alternative specifications. First, we got statistically identical policy improvements using the model with alternative train/test splitting based on municipality rather than municipality-year. Second, analyzing a policy based on the alternative measure of corruption from Avis et al. (2018) obtained proportionally larger improvements on the status quo in terms of detecting corruption. Overall, the machine learning approach to support anti-corruption policy is robust to such implementation choices.

5.2. Adjusting for Political Bias in Targeted Corruption Audits

A practical strength of randomized audits is their intuitive fairness, in the sense that all municipalities are targeted equally in expectation. Given the political sensitivity of corruption, it could be that any perceptions of bias in the allocation of audits would pose an institutional barrier to implementation of anti-corruption policies. In particular, if ML-targeted auditing changed the allocation of audits across political parties, the resulting perceptions of bias could generate conflict and prevent implementation of the policy.

In this section, we show how to address these concerns. First, we show that, indeed, targeting does change the allocation of audits across parties in terms of mayoral control of

[20] Note that a counterfactual policy with the opposite goal (minimizing corruption detection) could target the lowest-risk municipalities and realize a detection rate of just 0.03.

municipalities. Second, we show how the algorithm's audit decisions can be adjusted to produce politically neutral targeting that maintains equal audit rates by party. Third, we show that this "fair targeting" policy achieves similar gains in terms of detecting corruption.

We start by exploring how the corruption prediction model treats the different political parties in our dataset. We focus on the five largest parties, as indicated by the average share of municipalities they control in our period. These parties, ranked roughly from most left-wing to most right-wing (Power and Rodrigues-Silveira, 2019), are PT, PMDB, PSDB, PTB, and DEM (formerly PFL). The distribution of municipality-terms by party is shown in Appendix Figure F8.

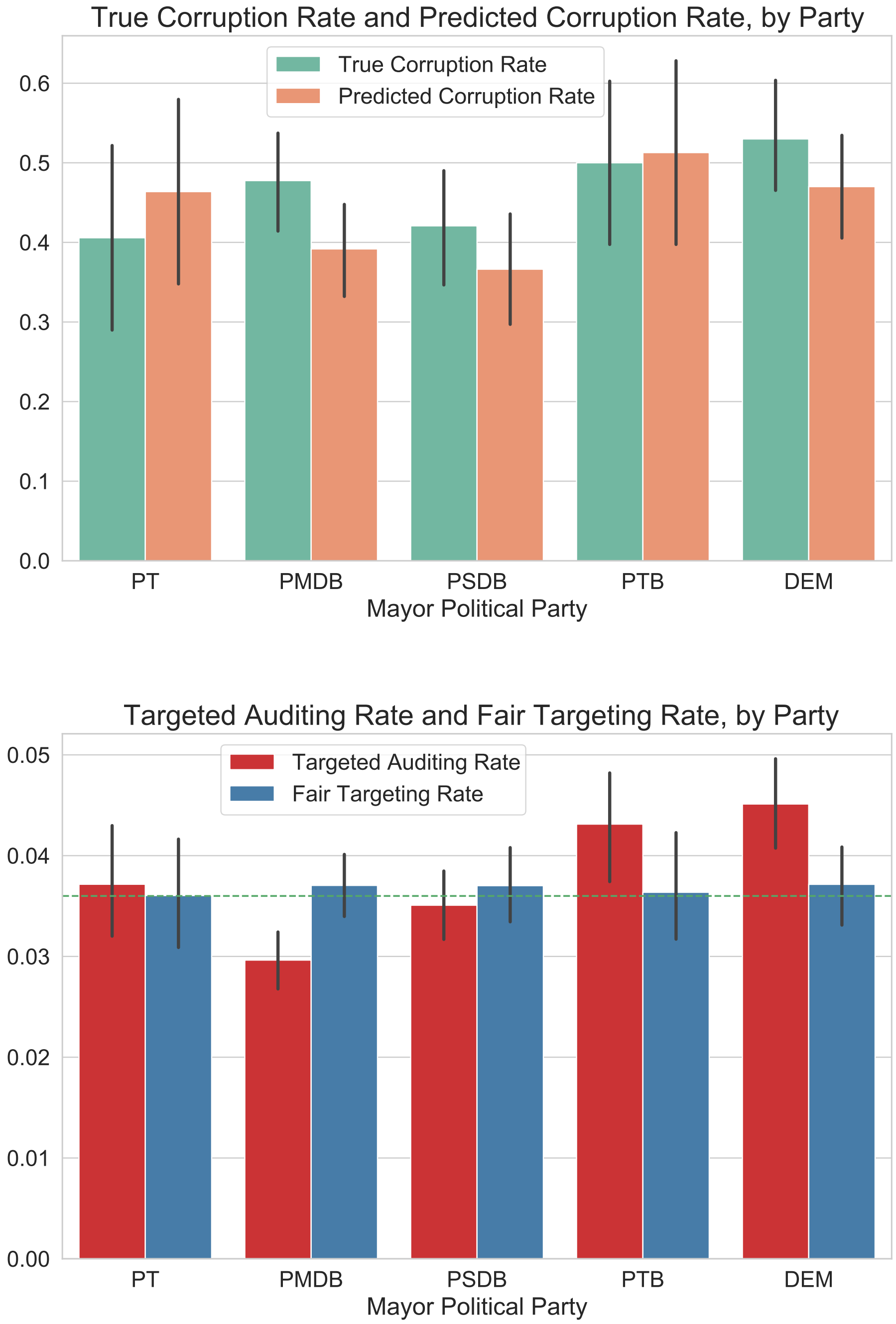
For each party, we compute the true corruption rate (from the random audits) and the predicted corruption rate (from the algorithm). These statistics are visualized in Figure 6, Top Panel. We can see that observed corruption (green bars) varies somewhat across parties. For example, PT has a relatively low corruption rate, while DEM has a relatively high corruption rate. The differences across parties are mostly reproduced in the model's predictions (orange bars), reflecting that the model is designed to treat parties the same. A notable exception is PMDB, for whom the algorithm somewhat understates the risk of corruption relative to the true rate.

These differences in predicted corruption risk would be reflected in different audit rates under a targeting system, perhaps leading to perceptions of bias and accompanying skepticism about such a program. To address such skepticism is the domain of algorithmic fairness (Barocas et al., 2019), a burgeoning literature that provides statistical definitions of fairness for evaluating automated decision processes. The classic case study is criminal risk scoring and how the scores vary across racial or ethnic groups (Chouldechova, 2017; Berk et al., 2018; Kasy and Abebe, 2020). Statistical fairness metrics assess divergences across groups in a model's predictions (e.g., error rates) or its decisions (e.g., jail/release rates).

While the literature has proposed a number of statistical fairness metrics, a standard criterion is *statistical parity.* Statistical parity requires that the probability of the negative outcome – in our case, audit rates – is equalized across groups. Given the political economy issues underlying corruption audits, statistical parity is a sensible criterion because it approximates the status quo in terms of the incidence of audits across political parties.

The algorithmic fairness literature has developed a family of approaches for adjusting

Figure 6: Corruption Risk and Targeted Auditing, by Party



True Corruption Rate and Predicted Corruption Rate, by Party

Targeted Auditing Rate and Fair Targeting Rate, by Party

Notes: Top Panel reports the true corruption rate in the audit data (in green bars) next to the predicted corruption rate from our XGBoost classifier (in orange bars), separately by the five political parties (meaning control of the mayor's office). Bottom Panel compares the auditing rates by party, under unconstrained targeting (red bars) and constrained targeting that equalizes audit rates across parties (blue bars). Horizontal dashed line gives the average audit rate in the sample. In both plots, 95% confidence interval spikes constructed by bootstrapping.

algorithmic decision procedures to achieve statistical fairness. An intuitive approach, which we follow here, is to separate the problem into a prediction step and a decision step. Rambachan et al. (2020) show that equity concerns can be addressed solely at the decision step, with the prediction step being untouched. This *post-processing* approach is distinct from the more technically complex *pre-processing* or *constrained optimization* approaches that are explored in the computer science literature (see Barocas et al., 2019, ch. 3). The advantage of the latter methods is that the model does not need access to the sensitive covariate – normally, race/ethnicity – in order to produce a fair decision. In our setting, the sensitive covariate (city mayor party affiliation) is not that sensitive after all, and it will always be available in practice. Thus we take the post-processing approach.

We propose the following politically neutral targeting policy. As noted, the prediction algorithm is not changed at all. We start with $\hat{y}_{it}$ for each municipality-year and the resulting corruption-risk ranking for all municipalities in a given year. Instead of taking the highest-ranked municipalities from the whole set, however, we produce separate rankings for each party. Within each party, we audit the same share of municipalities. Then by construction, the incidence of audits is equal across parties.

Figure 6 Bottom Panel shows the impact of fair targeting (blue bars) relative to unconstrained targeting (red bars). As intended, the fair audits have identical frequencies for each party (up to a rounding error). Comparing to the unconstrained rates, however, this fairness adjustment has significant redistributive consequences. On the one hand, PTB and DEM benefit from the introduction of fair targeting and are audited less often compared to the standard policy. On the other hand, fair targeting increases the audit risk for PMDB-controlled municipalities.

The next question is how fair targeting changes the overall effectiveness of audits, relative to unconstrained targeting. Revisiting Table 4, we see in Column 3 that the discovered corruption rate for audited municipalities is 0.836, still far higher than the random baseline (0.4664). Discovered corruption is just slightly less in magnitude than the main targeting policy, and the difference is not statistically significant. In terms of deterrence – the audit rate, conditional on corruption – the nonpartisan policy still maintains significant policy effectiveness gains: 0.065, which is still 1.793× higher than the audit rate of 0.0365 under random assignment. This is quite close to, and not statistically different from, the unconstrained targeting policy. Overall, adjusting targeted anti-corruption policies to equalize audit rates across political parties does not significantly undermine the effectiveness of those policies.

Still, policymakers should take care before implementing a politically neutral audit policy. Biased treatment of municipalities could take many forms besides political party affiliation, and adjusting the predictions for political affiliation might increase biases along the other margins. In turn, they could reduce deterrence for some municipalities.

5.3. Discussion

Beyond political neutrality, there are a number of issues worth additional discussion. First, to help contextualize the effectiveness of the proposed policies, we should consider the relevant baseline. So far, we have compared our algorithm to a baseline of random audits. Arguably, this is a weak baseline, and the model should be compared to a non-random human-targeted baseline. Bureaucrats likely have some good intuitions and experience about what budget irregularities reflect corruption, and they could guess better than randomly. It is difficult to judge how much better a human-targeted policy would work, however, as we do not have comparable data for it in the Brazilian setting.

While human targeting would likely be more accurate than the lottery, it would not come close to perfect prediction. Human targeting would be politically sensitive because it entails picking winners and losers. Thus it would be subject to political pressures, leading to biases and/or secrecy. The advantage of the algorithm is that its workings can be open-sourced for public review without subjecting it to such pressures. So a black-box algorithm might be able to produce effective targeting while getting around political economy barriers to human targeting.

Second, the policy simulation considered so far has a single round of targeted audits. At least in the short run, multiple targeted audit rounds would be possible and effective if they used the public finances data from before the first audit. But a more reasonable measure of success of targeted audits is not to identify a high number of corrupt municipalities but rather to deter corruption in the future. Subsequent to the first round of audits, the broader effect of the policy depends on how agents will learn and adapt to it. The existing model, when applied to post-targeting accounting data, may produce errors that would favor the more savvy mayors. Still, it could reduce the net marginal benefit of corrupt activities by increasing the expected cost of corrupt fiscal actions that are not easily substitutable.[21] On the other hand, the machine targeting might actually

[21] Our setting is not amenable to the "manipulation-proof machine learning" method from Björkegren et al. (2020), which requires information on the cost function over corruption activities.

spur additional corruption for municipalities who realize they have low predicted corruption. Therefore, the simulated gains from a static environment likely represent an upper bound on the longer-term deterrent effect, absent additional enhancements to the system.

In light of the behavioral responses, a question arises about how much information to publicize about audit targeting. One option would be to give full information about the policy and the associated model weights. This option would increase deterrence against corruption actions that are not easily substitutable. But it would reduce deterrence against substitutable actions, which could be easily gamed. Note that human targeting based on budget accounts would entail significant information provision, because prospectively corrupt officials could see what auditors consider in the published audit reports.

Another option would be to perform targeted audits without giving any information about how targeting is done. This system would give prospectively corrupt officials much less information than they would have under human targeting based on budgets, where they can examine the published audit reports. Presumably, under a secret system, corrupt officials could still learn how municipalities are targeted, but it is unclear whether this could be done quickly enough to allow manipulation of accounts to avoid audits. With human-targeted audits, meanwhile, there could be a problem for federal audit officials to keep up with changes in the technology of corruption. Updating bureaucratic infrastructure and rules can take years, and local officials could update their methods for corruption to stay ahead. With machine-targeted audits, the algorithm could presumably be updated every year to automatically stay on top of which budget items are newly associated with corruption.

Understanding the relevance of these factors would require additional empirical evidence, preferably through randomized interventions. The specific numbers from our simulation should be taken with some skepticism, given the previous work showing that the introduction of algorithms into decision-making can have smaller-than-expected effects (e.g., Stevenson and Doleac, 2019). There could be many reasons that an audit-targeting policy would not be as effective as outlined here.

In any case, a longer-term system of targeted audits could be strengthened by maintaining some random audits. In such a mixed system, targeted audits would be used to detect and deter corruption for the highest-risk municipalities. Random audits would be maintained for two reasons. First, even apparently low-risk municipalities (including those who are good at fooling the algorithm) would have some chance of being audited and therefore face some deterrence incentive. Second, the results of the random audits would be used to update the algorithm parameters for guiding the next round of targeted audits. Determining the optimal mix of targeted and random audits would require more information and more assumptions on the deterrence effect of both types of audits.

Beyond mixing random and targeting audits, a broader discussion could consider more refined implementation choices using predicted corruption risk. As one example, policymakers could adapt the scale of the audit to corruption risk. That is, low-risk municipalities could get a smaller team of auditors sent, while the high-risk municipalities could get a larger team. As another example, policymakers could exploit spillovers in audit effects on corruption (Avis et al., 2018) by targeting geographical clusters of high-risk municipalities.

6. Conclusion

This paper has shown that corruption in local governments can be reliably detected, predicted, and measured using public budget accounts data. We have shown that the resulting synthetic measurements can then be used in downstream empirical analysis, as we can produce the same empirical results using corruption predictions in municipalities that were never audited. In the future, we hope that the expanded datasets built with machine prediction could be broadly useful for social scientists interested in corruption, as well as other variables that the method could produce.

Beyond expanding on empirical work, the corruption predictions can be used to guide policy responses to corruption. Our counterfactual policy estimates indicate substantial gains from such a policy, even when constraining the algorithm to treat each political party equitably. We hope that this proof of concept leads to further exploration and experimentation by researchers, development organizations, and government agencies.

This research adds to the emerging literature using machine learning and other tools from data science to explore new datasets and questions (Kleinberg et al., 2015; Athey, 2018). Our method of detecting corruption has the potential to substantially expand the stock of datasets available for economists studying development, political economy, and public finance. Within Brazil, researchers will no longer be constrained to the relatively small set of municipalities that were audited. Outside of Brazil, the method could in principle be applied in any context with ground-truth labels for corruption. Something that can and should be explored is whether the corruption predictions produced in Brazil could be valid for other countries and settings.

References

Andini, M., E. Ciani, G. de Blasio, A. D'Ignazio, and V. Salvestrini (2018). Targeting with machine learning: An application to a tax rebate program in italy. *Journal of Economic Behavior & Organization 156*, 86–102.

Ash, E., M. Morelli, and R. Van Weelden (2017). Elections and divisiveness: Theory and evidence. *The Journal of Politics 79*(4), 1268–1285.

Assunção, J., R. McMillan, J. Murphy, and E. Souza-Rodrigues (2019). Optimal environmental targeting in the amazon rainforest. Technical report, National Bureau of Economic Research.

Athey, S. (2018). The impact of machine learning on economics. In *The economics of artificial intelligence: An agenda*. University of Chicago Press.

Athey, S. and S. Wager (2021). Policy learning with observational data. *Econometrica 89*(1), 133–161.

Avis, E., C. Ferraz, and F. Finan (2018). Do government audits reduce corruption? estimating the impacts of exposing corrupt politicians. *Journal of Political Economy 126*(5), 1912–1964.

Bandiera, O., A. Prat, S. Hansen, and R. Sadun (2020). Ceo behavior and firm performance. *Journal of Political Economy 0*(0), 000–000.

Bansak, K., J. Ferwerda, J. Hainmueller, A. Dillon, D. Hangartner, D. Lawrence, and J. Weinstein (2018). Improving refugee integration through data-driven algorithmic assignment. *Science 359*(6373), 325–329.

Barocas, S., M. Hardt, and A. Narayanan (2019). *Fairness and machine learning: Limitations and Opportunities.*

Battiston, P., S. Gamba, and A. Santoro (2020). Optimizing tax administration policies with machine learning. *University of Milan Bicocca Department of Economics, Management and Statistics Working Paper* (436).

Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy 76*(2), 169–217.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives 28*(2), 29–50.

Berk, R., H. Heidari, S. Jabbari, M. Kearns, and A. Roth (2018). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 0049124118782533.

Björkegren, D., J. E. Blumenstock, and S. Knight (2020). Manipulation-proof machine learning. *arXiv preprint arXiv:2004.03865*.

Bobonis, G. J., L. R. Cámara Fuertes, and R. Schwabe (2016). Monitoring corruptible politicians. *American Economic Review 106*(8), 2371–2405.

Brollo, F., T. Nannicini, R. Perotti, and G. Tabellini (2013). The political resource curse. *American Economic Review 103*(5), 1759–96.

Cavalcanti, F., G. Daniele, and S. Galletta (2018). Popularity shocks and political selection. *Journal of Public Economics 165*, 201–216.

Chen, T. and C. Guestrin (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.

Cheol, L. and J. Mikesell (2018). The impact of public officials' corruption on the size and allocation of u.s. state spending. *Public Administration Review*, 346–359.

Chong, A., A. L. De La O, D. Karlan, and L. Wantchekon (2015). Does corruption information inspire the fight or quash the hope? a field experiment in mexico on voter turnout, choice, and party identification. *The Journal of Politics 77*(1), 55–71.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data 5*(2), 153–163.

Colonnelli, E., J. A. Gallego, and M. Prem (2019). What predicts corruption? *Available at SSRN 3330651*.

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

Coviello, D. and S. Gagliarducci (2017). Tenure in office and public procurement. *American Economic Journal: Economic Policy 9*(3), 59–105.

Daniele, G. and T. Giommoni (2020). Corruption under austerity. *BAFFI CAREFIN Centre Research Paper No. 2020-131*.

De Angelis, I., G. de Blasio, and L. Rizzica (2020). Lost in corruption. evidence from eu funding to southern italy. *Italian Economic Journal*, 1–23.

Decarolis, F. and C. Giorgiantonio (2020). Corruption red flags in public procurement: new evidence from italian calls for tenders. *Questioni di Economia e Finanza, Occasional Papers* (544).

Djankov, S., R. La Porta, F. Lopez-de Silanes, and A. Shleifer (2003). Courts. *The Quarterly Journal of Economics 118*(2), 453–517.

Ferraz, C. and F. Finan (2008). Exposing corrupt politicians: The effects of Brazil's publicly released audits on electoral outcomes. *The Quarterly Journal of Economics 123*(2), 703–745.

Feurer, M., K. Eggensperger, S. Falkner, M. Lindauer, and F. Hutter (2018). Practical automated machine learning for the automl challenge 2018. In *International Workshop on Automatic Machine Learning at ICML*, pp. 1189–1232.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Gallego, J., G. Rivero, J. D. Martínez, et al. (2018). Preventing rather than punishing: An early warning model of malfeasance in public procurement. Technical report.

Gentzkow, M. and J. M. Shapiro (2010). What drives media slant? evidence from us daily newspapers. *Econometrica 78*(1), 35–71.

Gentzkow, M., J. M. Shapiro, and M. Taddy (2019). Measuring group differences in high-dimensional choices: Method and application to congressional speech. *Econometrica 87*(4), 1307–1340.

Glaeser, E. L., A. Hillis, S. D. Kominers, and M. Luca (2016). Crowdsourcing city government: Using tournaments to improve inspection accuracy. *American Economic Review 106*(5), 114–18.

Hansen, S., M. McMahon, and A. Prat (2018). Transparency and deliberation within the fomc: a computational linguistics approach. *The Quarterly Journal of Economics 133*(2), 801–870.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media.

Hessami, Z. (2014). Political corruption, public procurement, and budget composition: Theory and evidence from oecd countries. *European Journal of Political Economy 34*(C), 372–389.

Hitsch, G. J. and S. Misra (2018). Heterogeneous treatment effects and optimal targeting policy evaluation. *Available at SSRN 3111957*.

Kang, J. S., P. Kuznetsova, M. Luca, and Y. Choi (2013). Where not to eat? improving public policy by predicting hygiene inspections using online reviews. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1443–1448.

Kasy, M. and R. Abebe (2020). Fairness, equality, and power in algorithmic decision making. In *ICML Workshop on Participatory Approaches to Machine Learning*.

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan (2018). Human decisions and machine predictions. *The quarterly journal of economics 133*(1), 237–293.

Kleinberg, J., J. Ludwig, S. Mullainathan, and Z. Obermeyer (2015). Prediction policy problems. *American Economic Review 105*(5), 491–95.

Knaus, M. C., M. Lechner, and A. Strittmatter (2018). Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence. *The Econometrics Journal*.

Knittel, C. R. and S. Stolper (2019). Using machine learning to target treatment: The case of household energy use. Technical report, National Bureau of Economic Research.

Kyriacou, A. P., L. Muinelo-Gallo, and O. Roca-Sagalés (2015). Construction corrupts: Empirical evidence from a panel of 42 countries. *Public Choice 165*(1), 123–145.

Lagaras, S., J. Ponticelli, and M. Tsoutsoura (2017). Caught with the hand in the cookie jar: Firm growth and labor reallocation after exposure of corrupt practices.

Liu, C., T. T. Moldogaziev, and J. L. Mikesell (2017). Corruption and state and local government debt expansion. *Public Administration Review 77*(5), 681–690.

López-Iturriaga, F. J. and I. P. Sanz (2018). Predicting public corruption with neural networks: An analysis of spanish provinces. *Social Indicators Research 140*(3), 975–998.

Machoski, E. and J. M. de Araujo (2020). Corruption in public health and its effects on the economic growth of brazilian municipalities. *The European Journal of Health Economics*, 1–19.

Mauro, P. (1998). Corruption and the composition of government expenditure. *Journal of Public economics 69*(2), 263–279.

Mohler, G. O., M. B. Short, S. Malinowski, M. Johnson, G. E. Tita, A. L. Bertozzi, and P. J. Brantingham (2015). Randomized controlled field trials of predictive policing. *Journal of the American statistical association 110*(512), 1399–1411.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Morris, S. D. and J. L. Klesner (2010). Corruption and trust: Theoretical considerations and evidence from mexico. *Comparative Political Studies 43*(10), 1258–1285.

Mullainathan, S. and Z. Obermeyer (2019). A machine learning approach to low-value health care: wasted tests, missed heart attacks and mis-predictions. Technical report, National Bureau of Economic Research.

Mullainathan, S. and J. Spiess (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives 31*(2), 87–106.

Olken, B. A. (2007). Monitoring corruption: evidence from a field experiment in indonesia. *Journal of political Economy 115*(2), 200–249.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research 12*, 2825–2830.

Poblete-Cazenave, R. (2021). Reputation shocks and strategic responses in electoral campaigns.

Power, T. J. and R. Rodrigues-Silveira (2019). Mapping ideological preferences in brazilian elections, 1994-2018: a municipal-level study. *Brazilian Political Science Review 13*(1).

Rambachan, A., J. Kleinberg, S. Mullainathan, and J. Ludwig (2020). An economic approach to regulating algorithms. Technical report, National Bureau of Economic Research.

Rockoff, J. E., B. A. Jacob, T. J. Kane, and D. O. Staiger (2011). Can you recognize an effective teacher when you recruit one? *Education finance and Policy 6*(1), 43–74.

Stevenson, M. T. and J. L. Doleac (2019). Algorithmic risk assessment in the hands of humans. *Available at SSRN 3489440*.

Vannutelli, S. (2021). From lapdogs to watchdogs: Random auditor assignment and municipal fiscal performance in italy. *Job Market Paper: Yale University*.

Widmer, P., S. Galletta, and E. Ash (2020). Media slant is contagious. *Center for Law & Economics Working Paper Series 14*.

Winters, M. S. and R. Weitz-Shapiro (2013). Lacking information or condoning corruption: When do voters support corrupt politicians? *Comparative Politics 45*(4), 418–436.

Zamboni, Y. and S. Litschig (2018). Audit risk and rent extraction: Evidence from a randomized evaluation in brazil. *Journal of Development Economics 134*, 133 – 149.

A Machine Learning Approach to Analyze and Support Anti-Corruption Policy

APPENDIX

A. Additional Material on Model Training and Evaluation

Table A1: Balance sheets components

Year	N. of Categories								
	Assets		Liabilities		Expenditures		Revenues		
	All	Selected	All	Selected	All	Selected	All	Selected	N. of Audits
2001	56	43	46	37	43	43	52	51	0
2002	56	43	46	37	101	78	90	76	0
2003	57	44	48	39	100	77	90	76	276
2004	59	46	49	38	295	155	146	106	340
2005	63	43	52	38	298	158	151	105	300
2006	63	43	52	38	301	161	155	106	180
2007	64	43	52	38	309	161	170	108	180
2008	64	43	52	38	310	162	170	108	120
2009	80	41	57	37	331	161	198	108	180
2010	88	41	69	37	334	161	219	109	180
2011	89	41	69	37	335	161	219	109	120
2012	89	41	69	37	334	161	219	109	120

Notes: This table report the summary tabulations by year and by macro category on the total number of components of the municipal budget, the the number of components selected by XGBoost to form predictions, and the number of audits by year.

Table A2: Selected Hyperparameters and Learned Model Size

Fold	L1 Penalty	L2 Penalty	Max Tree Depth	Learning Rate	Min. Child Weight	Tree Count	Node Count
1	1	0.1	10	0.1	5	72	9461
2	1	0.1	10	0.1	3	71	11874
3	0.5	0.5	10	0.1	1	46	13135
4	2	2	10	0.1	5	97	13604
5	1	0.5	10	0.1	3	70	12467
mean	1.1	.64	10	0.1	3.4	71.4	12108.2

Notes: This table reports the hyperparameters selected for each of the 5 folds model training. Rows give the folds. L1 and L2 Penalty are regularization terms on the splitting decision that encourage smaller trees. Max Tree Depth is the max number of splits before a terminal node. Learning rate is how quickly parameters are updated during training. Minimum Child Weight is another regularization term, corresponding to the minimum number of observations required at each node. Is the number of trees grown in the resulting forest. The last column is the total number of variable splitting nodes in the forest.

Table A3: Confusion Matrices

Panel A. XGBoost

		Prediction	
		Not Corrupt	Corrupt
Truth	Not Corrupt	2573	485
	Corrupt	980	1261

Panel B. OLS

		Prediction	
		Not Corrupt	Corrupt
Truth	Not Corrupt	1243	1815
	Corrupt	961	1280

Panel C. LASSO

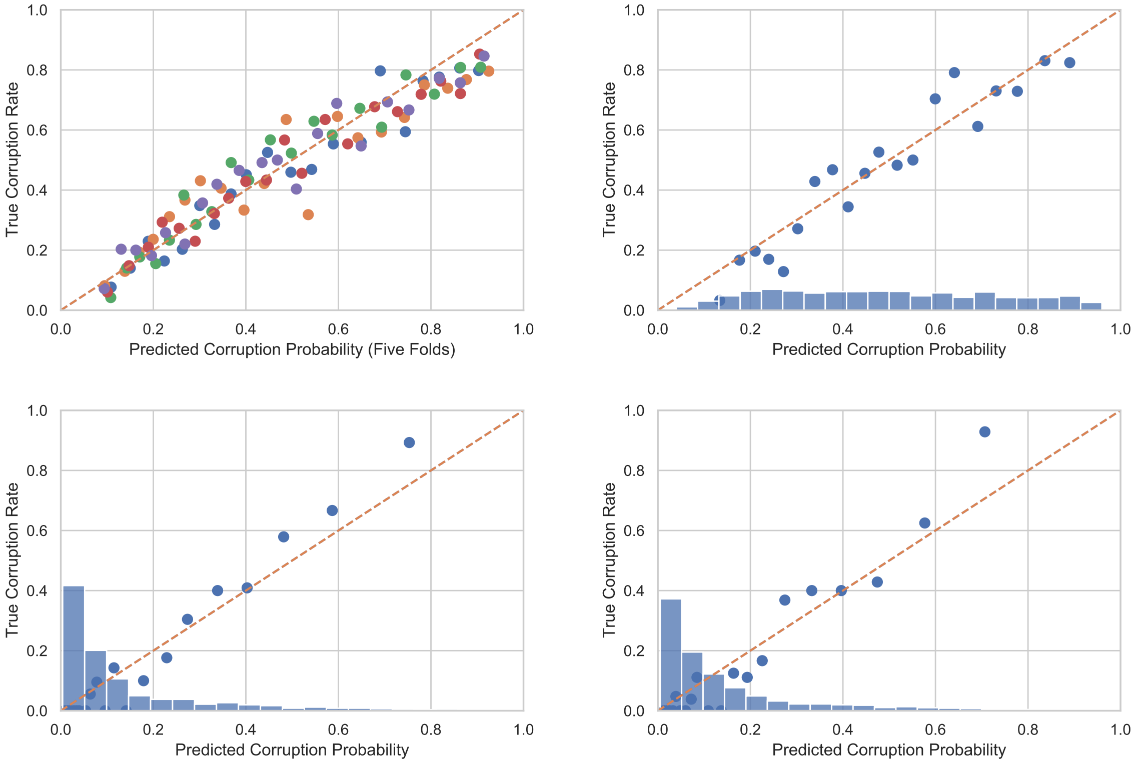
		Prediction	
		Not Corrupt	Corrupt
Truth	Not Corrupt	894	2164
	Corrupt	619	1622

Panel D. Logistic regression

		Prediction	
		Not Corrupt	Corrupt
Truth	Not Corrupt	1568	1490
	Corrupt	840	1401

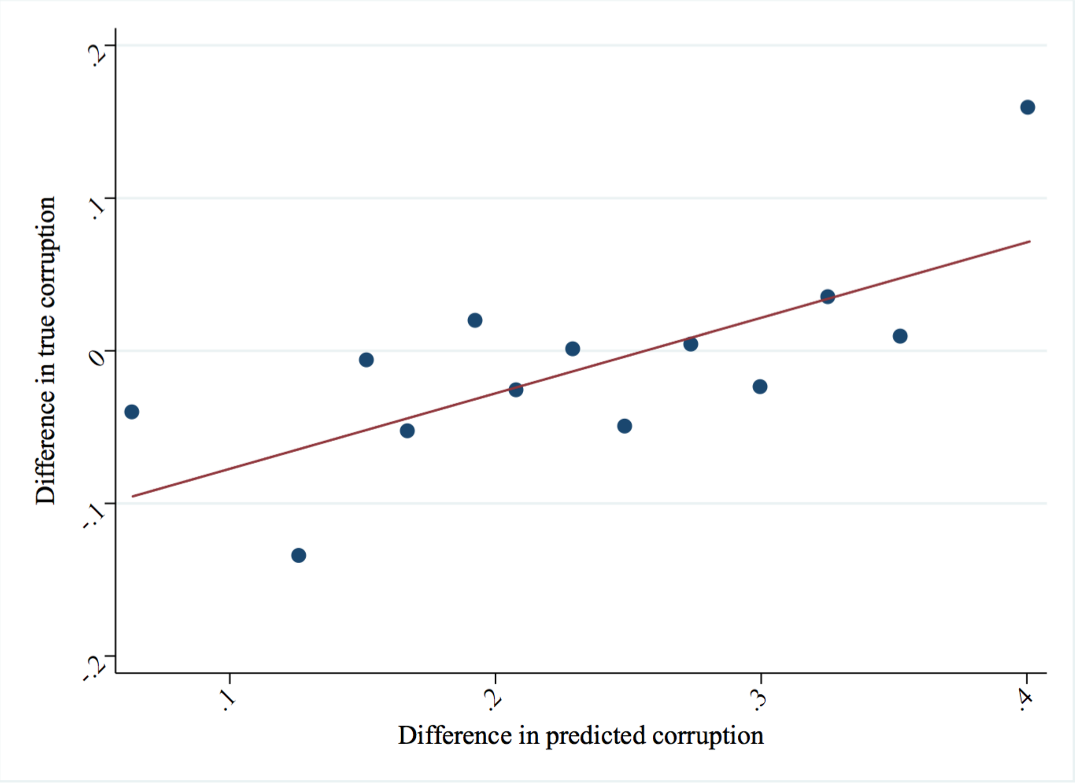
Notes: The table reports confusion matrices from the model predictions XGBoost (recall=0.562 and precision=0.722), OLS (recall=0.571 and precision=0.413), LASSO (recall=0.723 and precision=0.428) and Logistic regression (recall=0.625 and precision=0.484).

Figure A1: Additional Calibration Plots: True Corruption Rate vs. Predicted Corruption Risk



Notes: Calibration plots showing true corruption rates (marks on the vertical axis), binned by predicted corruption probability (horizontal axis). Dashed 45-degree line (in orange) demarcates perfect calibration. In top left panel, we use the baseline model, but show the calibration plot for each of the five models trained separately on different training folds. In the other panels, the blue histogram shows the density of the predicted corruption probability. Top right uses the Brollo et al corruption measure with group sampling by municipality. Bottom left and bottom right use the Avis et al corruption measure, with the bottom left using random sampling and the bottom right using municipality group sampling.

Figure A2: Difference in true and predicted corruption for cities audited twice



Notes: The figure focuses on cities that have been audited twice and it shows a binscatter between the difference over time in the true levels of corruption using the data from Brollo et al. (2013) and the predicted levels of corruption. The analysis includes the following list of fixed effects and controls: first audit year and second audit year fixed effects, mean income, share of population employed, sector of occupation (agriculture, industry, commerce, transportation, services and public administration), share with college education, poverty rate, and Gini Coefficient of income. The coefficient of the corresponding regression is 0.495 (p-value 0.095).

B. Alternative Specifications for Corruption Prediction

This appendix reports the performance metrics from some alternative corruption prediction specifications. First, to compare XGBoost model performance using not only budget factors but also fixed demographic factors, we apply random splits between training and test set by municipality, instead of by municipality-year. Appendix Table B4 shows the relative performance when we use budget data (column 1), when we add demographic characteristics (column 2), or when we use only demographic characteristics (column 3). The more conservative sampling specification in Column 1 reduces accuracy compared to the main-text specification, but it is still capturing significant predictive signal (in Column 3, AUC-ROC = 0.636 with budget and demographics). Comparing Column 1 to Column 2, we see that budget information is more predictive of corruption than demographic information.

Second, we replicate our predictive results by using the corruption measure from Avis et al. (2018). It is important to stress that there are structural differences between these two original measures of corruption. A first important difference is that the alternative measure is continuous, rather than binary. We have for each audited municipality the share of inspection orders that presented irregularities. The second difference is that we are missing the first audits, as we have information only from July 2006 through March 2013 (lotteries 22–38). Third, differently from the main main measure, with the alternative measure we do not know the exact year (or term) in which the irregularity took place. To overcome this limitation, we treat as audited the three years before the actual audit took place. Finally, to create a binary label from the continuous variable we identified as corrupted those municipalities with a share of irregularities in the top quartile of the distribution.

Despite these differences, Figure B3 shows that the predictions using the alternative corruption label are similar to those from the main analysis. They show similar rankings on average. The performance metrics are reported columns (4-7) of Appendix Table B4. Again, we find that XGBoost outperforms all the other methods. Indeed, we find accuracy metrics that are higher than those from the main analysis.

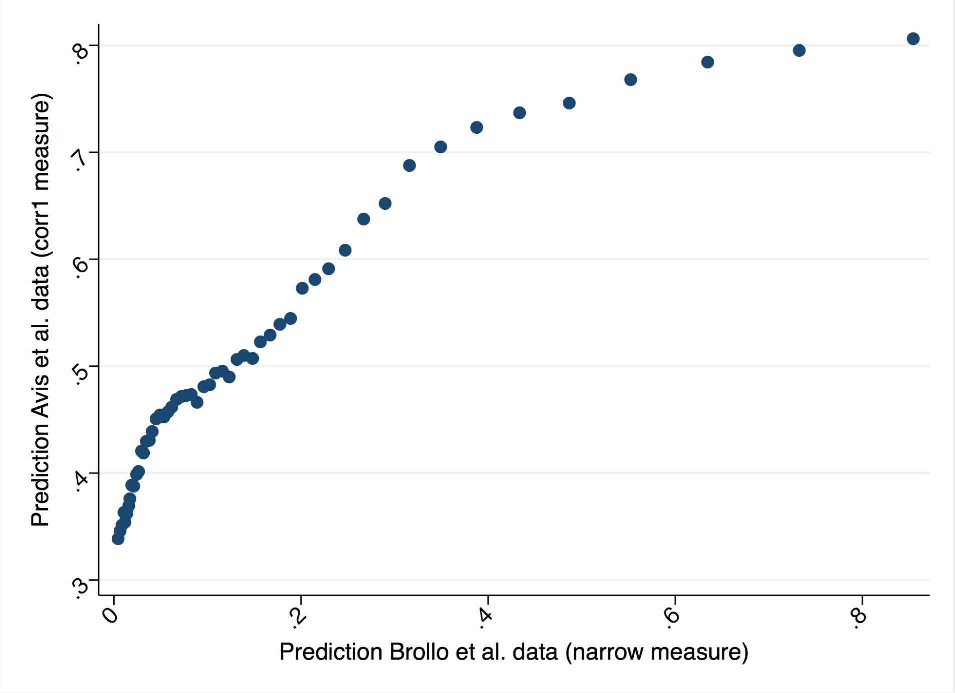
Finally, we find that most of our empirical results still hold when using the predictions from this alternative measure of corruption.

Table B4: Additional models performance

	XGBoost (municipal sampling)			Avis et al. (2018) data			
	Budget (1)	Budget + Demo (2)	Demo (3)	XGBoost (4)	OLS (5)	LASSO (6)	Logistic (7)
Accuracy	0.613	0.613	0.576	0.851	0.431	0.419	0.688
	(0.012)	(0.004)	(0.009)	(0.005)	(0.048)	(0.062)	(0.036)
AUC-ROC	0.618	0.636	0.589	0.903	0.443	0.519	0.657
	(0.015)	(0.006)	(0.012)	(0.009)	(0.065)	(0.033)	(0.010)
F1	0.486	0.498	0.476	0.635	0.311	0.375	0.485
	(0.018)	(0.007)	(0.009)	(0.017)	(0.050)	(0.028)	(0.021)

Notes: The table provides the mean and standard error (in parentheses) across five values for the prediction performance, produced using different training-set folds. In columns (1-3) we use XGBoost models with municipal sampling, and different sets of predictors: only budget components in column (1), budget components and demographic characteristics in column (2) and only demographic characteristics in column (3). In columns (4-8) we report the predictions performance as in Table 2, but using the corruption data from Avis et al. (2018).

Figure B3: Predictions from Avis et al. (2018) vs. Predictions from Brollo et al. (2013)



Notes: The figure shows a binscatter between the predictions formed using the data from Avis et al. (2018) and the ones formed using the data from Brollo et al. (2013) for all municipality-year. The correlation between the two variables is 0.531.

Table C5: Most important budget features for Corruption Prediction

				Perturbation Response		
N.	Category	Category	Weight	Mean	Min	Max
1	Tax on agricultural territorial property (ITR) (compartecipation)	Revenue	103	0.019	-0.21	0.41
2	Spending in agriculture	Expenditure	103	0.007	-0.22	0.55
3	Spending in transportation	Expenditure	95	0.007	-0.22	0.44
4	Tax on export of industrialized products (IPI) (compartecipation)	Revenue	92	0.018	-0.34	0.56
5	Budget Surplus/Deficit		84	0.029	-0.26	0.53
6	Cash	Assets	84	0.012	-0.25	0.46
7	Tax on real estate transactions (ITB)	Revenue	80	0.024	-0.21	0.37
8	Taxes	Revenue	79	0.013	-0.48	0.57
9	Deposits	Assets	76	-0.007	-0.42	0.21
10	Motor vehicle property tax (IPVA) (compartecipation)	Revenue	74	0.002	-0.33	0.45
11	Income Tax (IRRF)	Revenue	73	0.0004	-0.18	0.30
12	Transfers for the health system	Revenue	71	-0.015	-0.33	0.19
13	Civil servant per diems	Expenditure	70	-0.012	-0.42	0.42
14	Spending for legislative procedure	Expenditure	68	-0.001	-0.54	0.26
15	Revenue from assets	Revenue	67	0.011	-0.18	0.38
16	Transfers from tax on circ. of goods/services (Law 87-96)	Revenue	66.6	0.010	-0.37	0.29
17	Tax on real estate (IPTU)	Revenue	65.4	0.019	-0.22	0.36
18	Tax to fund police authority	Revenue	64.4	-0.007	-0.30	0.27
19	Direct spending (previous years)	Expenditure	63	-0.042	-0.55	0.23
20	Transfers from tax on circulation of goods/services (compartecipation)	Revenue	61.4	0.003	-0.38	0.26
21	Capital expenditure	Expenditure	61	-0.003	-0.35	0.16
22	Outstanding debt	Liabilities	60.6	-0.002	-0.26	0.28
23	Financial and non-financial liabilities	Liabilities	60.4	-0.0001	-0.24	0.29
24	Supplies (current year)	Passive	59	0.006	-0.17	0.40
25	Banks	Assets	58.2	0.004	-0.18	0.29
26	Liquid assets	Assets	57	0.002	-0.26	0.28
27	Revenue from services provided	Revenue	54	-0.007	-0.40	0.17
28	Financial liabilities	Liabilities	50.8	-0.004	-0.21	0.26
29	Non-financial liabilitie	Liabilities	50.2	-0.001	-0.22	0.24
30	Financial assets	Assets	49.4	0.003	-0.13	0.22
31	Direct spending for consulting	Expenditure	48.4	-0.014	-0.37	0.24
32	Other revenues	Revenue	48	-0.015	-0.46	0.14
33	Spending for sports and leisure	Expenditure	46.6	-0.003	-0.24	0.18
34	Supplies (previous years)	Passive	46.2	-0.0004	-0.25	0.15
35	Other deposits	Assets	46	-0.003	-0.18	0.17

Notes: List of the most important features. Weight ranks the features (budget components) by how often they are included in a decision tree contained in the ensemble classifier, averaged across the five training folds. The last three columns show the mean, minimum, and maximum values computed from the perturbation analysis described in Appendix C.1, evaluating how each individual feature affects the predicted probability of being corrupted.

C. Additional material for model interpretation

This appendix contains additional material for interpreting the predictions of the tree ensemble. Table C5 shows the list of variables (Column 2), ranked by their feature importance weight (Column 4). This weight is the average across five models (using different training-set folds) of the number of times that a constituent decision tree splits on a variable. For example, the model "uses" spending on agriculture (second row) about 103 times.

C.1. Perturbation-Based Partial Dependence Analysis

Here we illustrate the important non-linearities and interactions encoded by the tree ensemble. We take a perturbation-based partial dependence approach (see, e.g., Friedman, 2001), which works as follows. Iterating over each observation i in the dataset, we form the predicted change in $\hat{Y}_i$ from perturbing a budget feature j by one standard deviation, either up or down. This perturbation is done for each model m (that is, models trained on a different training-set fold), so we obtain a dataset of values $\Delta\hat{Y}_{ijm}$ (with values inverted for negative perturbations).

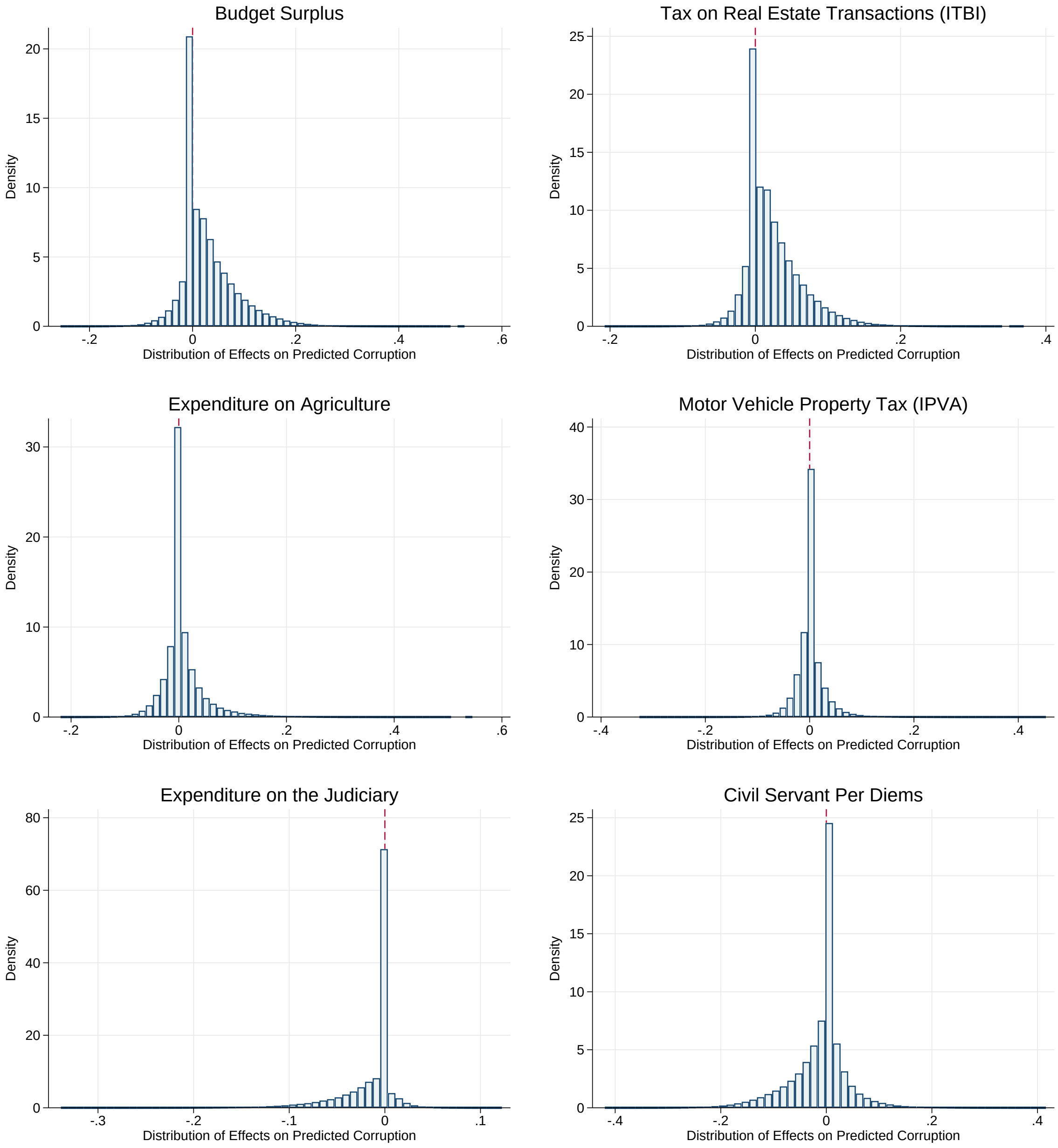
With five folds and a positive/negative perturbation per fold, we observe ten deltas for each observation in the dataset. With 5563 municipalities and 12 years of data per municipality, we produce about 650,000 deltas in total for each of 797 feature variables. What can these distributions of predicted changes in corruption rates tell us? If corruption were linearly related to features, we would expect the distribution of $\Delta\hat{Y}_{ijm}$ to have the same sign for a given feature j. If the algorithm captures important non-linearities, non-monotonicities, and/or interactions, then the responses would have both positive and negative density.

Overall, the perturbation results are consistent with a highly non-linear and contingent predictive relationship. Of the 446 predictive variables, only 19 variables have an always-weakly-positive relation, and only 20 variables have an always-weakly-negative relation. Columns 6 and 7 of Table C5 show the minimum and maximum values for the perturbation responses $\Delta\hat{Y}_{ijm}$ by variable, for the most important features. We can see that for all of these variables, the minimum is negative and the maximum is positive. So for the most pivotal variables, shifts could have either a positive or negative relation to predicted corruption depending on the status quo values.

To help illustrate this contingent partial dependence, Figure C4 shows distributions of the perturbation response $\Delta\hat{Y}_{ijm}$ for a selection of variables j. Each graph contains a relatively wide distribution of possible responses, again indicating a non-linear, interaction-heavy relationship. For example, it is intuitive that judiciary spending (bottom left panel) is mostly negatively related to corruption risk. However, the effect is non-monotonic and there are some positive values in the response distribution.

In Table C5 Column 5, we report the mean value of the feature perturbation response. Bearing in mind all of the caveats mentioned so far, this column shows the average direction and magnitude of the model's perceived association between the indicated variable and corruption risk. Positive values indicate that higher values of this variable

Figure C4: Distributions of Predicted Responses to Perturbing Model Features



Notes: The figure shows histograms of $\Delta\hat{Y}_{ijm}$, produced using the perturbation approach described in the text, for a selection of variables j.

tend to reflect greater corruption risk, while negative values indicate that higher values of this variable tend to reflect lower corruption risk.

As mentioned in the main text, it is consistent with some previous literature that (3) expenditures on transportation (Hessami, 2014) and (7) real estate and construction (Kyriacou et al., 2015) are positively correlated with corruption. In addition, it is intuitive that higher (18) spending on policing is negatively related to corruption. But other positive variables are inconsistent with the previous literature. For example, our model suggests that having a budget surplus rather than deficit (5) is positively associated with corruption, which goes against the findings in Liu et al. (2017). Overall, these additional results do not modify the central point that the model's functional form is complex and one cannot identify one-to-one relationships between a budget factor and predicted corruption.

C.2. Counting Budget Feature Mentions in Audit Reports

The municipal audit reports are published on the web site of the CGU, `auditoria.cgu.gov.br`, in a search engine interface. We programmatically downloaded the full library of audit reports for our time period as PDF files. The corpus contains 2,062 reports. The PDFs were in machine-readable Portuguese and therefore straightforward to extract as plain text using the python package pdfminer.

We performed some mild pre-processing on the report texts. Punctuation and capitalization were removed. The resulting pre-processed corpus consists of 2,062 documents, each containing on average 26,743 words and 173,648 characters. In total, the corpus contains over 55 million words.

The next step is to identify mentions of relevant budget factors. Our dataset of budget features has a codebook with a variable label and short description for each budget item. For example, the budget item "Outras TrConvMun" is described as "Outras Transferências de Convênios dos Municípios" ("Other Transfers from Municipalities"). Both the label and the description are included in our pattern matching lexicon, after being pre-processed in the same way as the corpus (punctuation and capitalization removed). The lexicon contains 1,141 items as sometimes the label and description are the same. On average, the pre-processed items contain 28 characters and are 4 words long.

Table C6: Budget features most often mentioned in the audit reports

N.	Category	Category	Mention
1	Health expenditure	Expenditure	190,610
2	Assets	Assets	69,328
3	Spending in labour	Expenditure	59,835
4	Spending in education	Expenditure	56,563
5	Spending in adminstration	Expenditure	49,858
6	Cash	Assets	35,553
7	Spending in transportation	Expenditure	32,176
8	Spending in social services	Expenditure	29,633
9	Spending in basic health	Expenditure	28,499
10	National fund for education development	Revenue	19,148
11	Spending in culture	Expenditure	15,776
12	Spending in primary education	Expenditure	15,049
13	Supply spending	Expenditure	10,427
14	Spending in agriculture	Expenditure	9,615
15	Permanent assets	Assets	8,132
16	Spending in communication	Expenditure	8,062
17	Spending in social security	Expenditure	8,015
18	Spending in sanitation	Expenditure	6,739
19	Spending in the employment fund	Expenditure	5,501
20	Current spending in other contributions	Expenditure	4,510
21	Spending in transfers	Expenditure	4,388
22	Spending in telecommunication	Expenditure	4,273
23	Spending in energy	Expenditure	3,961
24	Spending in kindergarten	Expenditure	3,799
25	Stocks	Assets	3,758
26	Spending in tourism	Expenditure	3,527
27	Spending in high school	Expenditure	2,791
28	Transportation services	Revenue	2,100
29	Spending in health surveillance	Expenditure	2,096
30	Spending in adult education	Expenditure	1,944
31	Taxes	Revenues	1,942
32	Spending in electric energy	Expenditure	1,881
33	Spending in industry	Expenditure	1,848
34	Spending in leisure	Expenditure	1,819
35	Spending in urban infrastructures	Expenditure	1,796

Notes: List of the features most often mentioned in the audit reports, as described in Appendix C.2.

Finally, we counted the total mentions of each budget feature in the corpus of reports, limiting to exact matches of the pre-processed strings. Summary statistics on these matches are reported in Table C6. For example, health expenditures are mention almost 200,000 times. 28% of the budget variables are mentioned in the reports. Conditional on being mentioned at all, a budget factor is mentioned 4,465 times on average, or about twice per report.

D. Additional Material: Effect of Revenue Shocks on Corruption

Table D7: Population thresholds for Inter-Government Transfers

Population interval	FPM coefficient
Below 10,189	0.6
10,189−13,584	0.8
13,585−16,980	1
16,981−23,772	1.2
23,773−30,564	1.4
30,565−37,356	1.6
37,357−44,148	1.8
44,149−50,940	2
Above 50,940	from 2.2 to 4

Notes: These coefficients have been introduced by *Decreto-lei* n. 1,881, 27 august 1981.

Table D8: Descriptive statistics for the Revenue Shocks Analysis

	FPM transfers			
Population	Actual transfers	Theoretical transfers	Predicted Corruption	N
(1)	(2)	(3)	(4)	(5)
6,793 − 10,188	19.655	21.200	.442	1,429
10,189 − 13,584	25.642	28.771	.500	1,076
13,585 − 16,980	31.888	36.316	.527	805
16,981 − 23,772	38.445	44.019	.543	1,083
23,773 − 30,564	44.223	51.082	.529	629
30,565 − 37,356	50.869	58.113	.521	380
37,357 − 44,148	57.376	66.468	.510	253
44,149 − 50,940	62.389	72.368	.498	154
Total	33.440	37.930	.502	5,809

Notes: The sample includes all Brazilian municipalities with population in the interval 6,793-50,940. Population is the number of inhabitants. Actual and theoretical FPM transfers expressed in R$100,000 at 2000 prices.

Table D9: Replication Brollo et al. (2013) with random samples

Random sample:	First (1)	Second (2)	Third (3)	Fourth (4)
Panel A. First Stage				
Theoretical transfers	0.7649*** (0.0215)	0.7100*** (0.0247)	0.6810*** (0.0485)	0.7344*** (0.0177)
Panel B. Reduced Form				
Theoretical transfers	0.0048*** (0.0007)	0.0042*** (0.0007)	0.0049*** (0.0007)	0.0038*** (0.0008)
Panel C. 2SLS				
Actual transfers	0.0063*** (0.0010)	0.0059*** (0.0010)	0.0072*** (0.0011)	0.0052*** (0.0010)
N. Observations	1115	1115	1115	1115

Notes: Effects of FPM transfers on (predicted) corruption measures. The four columns display the analysis focusing on four different random samples with 1,115 observations. Panel A reports the estimates of the first-stage analysis, the dependent variable is *actual transfers*. Panel B reports the estimates of reduced form analysis, the dependent variable is *predicted corruption*. Panel C reports the estimates of the 2sls estimates, the dependent variable is *predicted corruption* and *actual transfers* is instrumented with *theoretical transfers*. Column headings indicate the sample of municipalities included. All regressions controls for a third-order polynomial in normalized population size, term dummies, and macro-region dummies. Robust standard errors clustered at the municipal level are in parentheses: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table D10: Effect of Revenue Shocks on Corruption - Alternative Predictions

Dep. var.: Predicted corruption	All cities	
	Prediction demographics (1)	Prediction budget (without FPM) (2)
Panel A. Reduced Form		
Theoretical transfers	-0.0002	0.0045***
	(0.0008)	(0.0003)
Panel B. 2SLS		
Actual transfers	-0.0003	0.0065***
	(0.0011)	(0.0005)
N. Observations	5808	5808

Notes: Effects of FPM transfers on (predicted) corruption measures: column (1) contains the analysis with the predictions built using as predictors a set of municipal demographic characteristics, and column (2) contains the analysis with the predictions built with budget predictors where FPM transfers are permuted randomly. Panel A reports the estimates of reduced form analysis, the dependent variable is *predicted corruption.* Panel B reports the estimates of the 2sls estimates, the dependent variable is *predicted corruption* and *actual transfers* is instrumented with *theoretical transfers.* The sample includes all Brazilian municipalities. All regressions controls for a third-order polynomial in normalized population size, term dummies, and macro-region dummies. Robust standard errors clustered at the municipal level are in parentheses: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

E. Additional Material: Effect of Audits on Corruption

In this Appendix we discuss a series of additional results for the event study analysis. First, we check whether post-audit budget adjustments may explain the decline in predicted corruption levels after the audit. We provide two tests. First, we estimate the main model controlling for total expenditure, expressed in per-capita terms. This test is reported in Figure E5 and the results are similar to the ones of the main model, reported in Figure 4. Secondly, we estimate the main model using as dependent variable the amount of total expenditure (per-capita): Figure E6 shows this test and it suggests that the audit does not have any significant effect on future levels of total expenditure. This result holds for the full sample and for the sample of corrupted and non-corrupted cities.

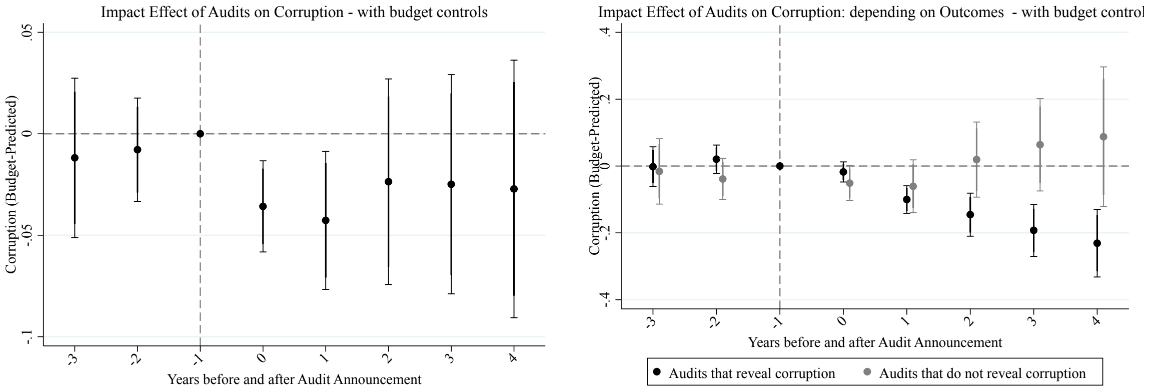
Second, we test the channel of political accountability. In particular, we aim to study whether the effect of the audit on future corruption is stronger where local political accountability is high and we focus on the variable margin of victory. This test is shown in Figure E7, which reports the analyses conducted with the full sample. The figures show that the effect is stronger in cities where the mayor won with a small margin of victory – below the median level – compared to cities where she won with a high margin – above the median level. This result suggests that the audit has a larger impact where the electoral competition is more pronounced. Overall, these results provide some evidence that political accountability affects the impact of an audit on future corruption.

Table E11: Coefficient Estimates for Event Study Analysis

	All cities	Cities with corruption	Cities without corruption
	(1)	(2)	(3)
Year pre4 and behind	-0.0171	-0.0287	-0.0052
	(0.0245)	(0.0427)	(0.0748)
Year pre3	-0.0118	-0.0024	-0.0164
	(0.0190)	(0.0287)	(0.0476)
Year pre2	-0.0078	0.0203	-0.0390
	(0.0124)	(0.0205)	(0.0302)
Audit year	-0.0358***	-0.0177	-0.0506*
	(0.0109)	(0.0145)	(0.0254)
Year post1	-0.0429**	-0.1002***	-0.0597
	(0.0166)	(0.0200)	(0.0387)
Year post2	-0.0238	-0.1456***	0.0205
	(0.0246)	(0.0311)	(0.0545)
Year post3	-0.0253	-0.1924***	0.0659
	(0.0262)	(0.0376)	(0.0672)
Year post4	-0.0276	-0.2307***	0.0903
	(0.0308)	(0.0490)	(0.1018)
Year post5	-0.0156	-0.2585***	0.1581
	(0.0418)	(0.0620)	(0.1185)
Years post6 and more	-0.0364	-0.3260***	0.1756
	(0.0478)	(0.0711)	(0.1294)
N. Observations	17252	8895	3086
Adjusted R^2	0.535	0.510	0.538

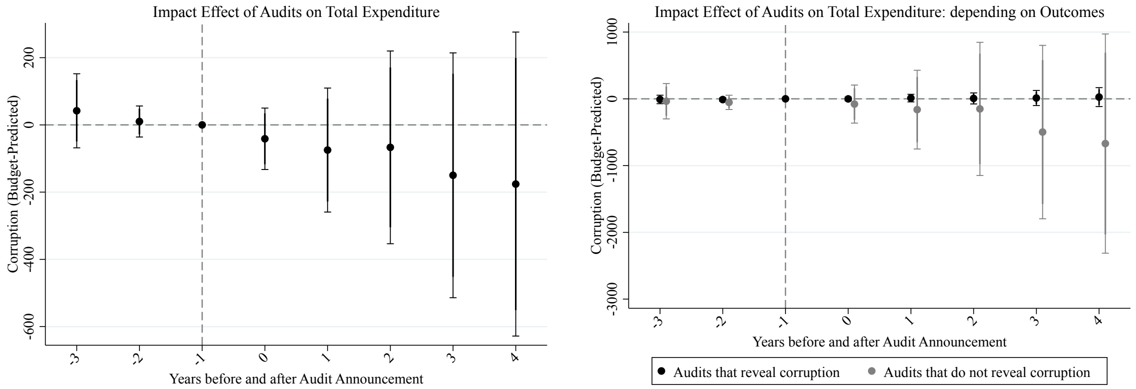
Notes: The dependent variable is (predicted) corruption measure - binary. The sample includes all the cities that receive an audit for the period 2001-2012. Column (1) includes the complete sample, Column (2) includes the sample of cities in which the audit discovered corruption (according to the definition of narrow corruption) and Column (3) includes the sample of cities in which the audit did not discover any type of corruption. The specification includes city and year fixed effects. Robust standard errors clustered at the state level are in parentheses: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure E5: Dynamic effect of the audits - Controlling for total expenditure



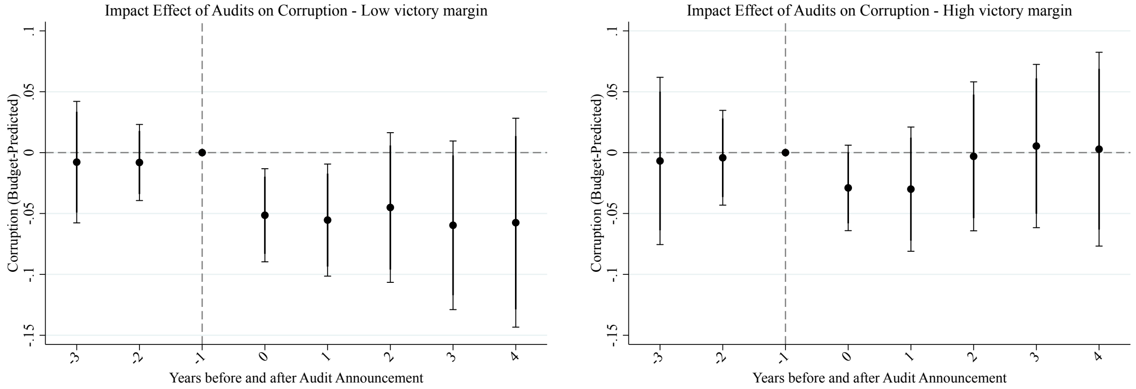
Notes: Event study estimates for dynamic effect of audits on budget-predicted corruption. Error spikes give 95% confidence intervals, with standard error clustered by state. Left panel: all audits; right panel: audits that found corruption (in black); audits that did not find corruption (in grey). This regressions include as additional control municipal total expenditure.

Figure E6: Dynamic effect of the audits on total expenditure



Notes: Event study estimates for dynamic effect of audits on municipal total expenditure. Error spikes give 95% confidence intervals, with standard error clustered by state. Left panel: all audits; right panel: audits that found corruption (in black); audits that did not find corruption (in grey).

Figure E7: Dynamic effect of the audits - Margin of victory



Notes: Event study estimates for dynamic effect of audits on budget-predicted corruption. Error spikes give 95% confidence intervals, with standard error clustered by state. In the left panel are considered only municipalities where the mayor won with a low margin of victory (below the median); In the right panel are considered only municipalities where the mayor won with a high margin of victory (above the median)

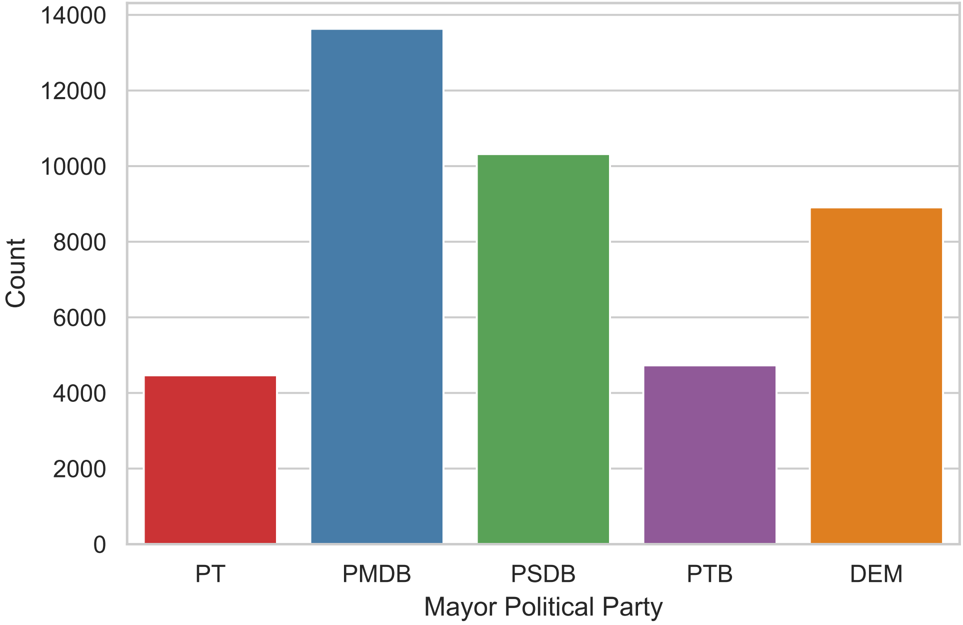
F. Additional Material on Audit Policy Support

Table F12: Performance Metrics for Targeted Auditing Policies, Additional Specifications

	Brollo et al. (2013) Corruption Labels			*Avis et al. (2018) Corruption Labels*		
	Random Audits	Targeted Audits		Random Audits	Targeted Audits	
	(1)	(2)	(3)	(4)	(5)	(6)
Train/Test Split		Muni-Year	Muni		Muni-Year	Muni
Corruption Rate, if Audited	0.4664	0.8563	0.8231	0.19	0.6743	0.6725
		(0.0163)	(0.0217)		(0.0186)	(0.0257)
# Corrupt Munis Detected	94.8	174.0861	167.3261	40.4	137.0670	136.7004
		(3.3225)	(4.45)		(3.7793)	(5.2294)
Audit Rate, if Corrupt	0.036	0.0671	0.0645	0.036	0.1241	0.1237
		(0.0013)	(0.0017)		(0.0034)	(0.0047)
↪ Ratio to Lottery		1.836	1.7648		3.3954	3.3863
		(0.035)	(0.0465)		(0.0936)	(0.1295)
Min Audit # Equivalent		110.9143	115.5913		60.1115	60.4993
		(2.0543)	(3.0415)		(1.7429)	(2.4499)

Notes: Metrics for comparing the effectiveness of audit policies. Columns 1 through 3 use the main label of corruption from Brollo et al. (2013). Columns 4 through 6 use the alternative label of corruption from Avis et al. (2018). Columns 1 and 4 report the true rates under random audits. Columns 2, 3, 5, and 6 report the results from targeting audits, with Columns 2 and 5 using the main train/test sampling by munipality-year, and Columns 3 and 6 using the alternative grouped splitting by municipality. The rows report the different politcy outcomes. "Corruption Rate, if Audited" is the share of audited municipalities where narrow corruption is detected. "# Corrupt Munis Detected" is the number of corrupt municipalities detected, out of the 203 audits implemented. "Audit Rate, if Corrupt" is the expected probability of being audited if corrupt. "Ratio to lottery" is the "Audit Rate, if Corrupt" value for the indicated policy, divided by that value under random audits. "Min Audit # Equivalent" is the number of audits needed under targeting to detect the same number of corrupt municipalities detected under the lottery system. For the audit-targeting statistics, we report the mean and standard error (in parentheses) across five values for the predicted corruption risk, produced using different training-set folds.

Figure F8: Distribution of Party Control of Municipalities



Notes: Number of municipality-year observations for each party, in terms of the affiliation of the mayor in that municipality.