Fundação Getulio Vargas
Escola de Matemática Aplicada

Laura Sant'Anna Gualda Pereira

Learning about Corruption:
A Statistical Framework for working with Audit Reports

Rio de Janeiro
2018

Laura Sant'Anna Gualda Pereira

Learning about Corruption:

A Statistical Framework for working with Audit Reports

Dissertação submetida à Escola de Matemática Aplicada como requisito parcial para a obtenção do grau de Mestre em Modelagem Matemática

Área de Concentração: Modelagem e Análise da Informação

Orientador: Eduardo Fonseca Mendes

Rio de Janeiro
2018

Acknowledgment

Firstly, I would like to thank my family and friends for the emotional support throughout the development of this work. In special, I thank Tom for his comprehension and for helping me believe in my abilities during my Masters.

I am truly grateful to my thesis advisor, Professor Eduardo Fonseca Mendes, for his patient, rigorous and encouraging guidance over the past two years. I would like to extend my thanks to all the professors of the School of Applied Mathematics at FGV. Each of you played an important role in my development as a researcher.

I would also like to thank all the staff of FGV for their kindness on my toughest moments.

I wish to thank CGU and Rafael Velasco, researcher at FGV/CTS, for the data.

Finally, I am also grateful to CAPES for the financial support.

Abstract

Quantitative studies aiming to disentangle public corruption effects often emphasize the lack of objective information in this research area. The CGU Random Audits Anti-Corruption Program, based on extensive and unadvertised audits of transfers from the federal government to municipalities, emerged as a potential source to try to fill this gap. Reports generated by these audits describe corrupt and mismanagement practices in detail, but reading and coding them manually is laborious and requires specialized people to do it. We propose a statistical framework to guide the use of text data to construct objective indicators of corruption and use it in inferential models. It consists of two main steps. In the first one, we use machine learning methods for text classification to create an indicator of corruption based on irregularities from audit reports. In the second step, we use this indicator in a regression model, accounting for the measurement error carried from the first step. To validate this framework, we replicate an empirical strategy presented by Ferraz et al. (2012) to estimate effects of corruption in educational funds on primary school students' outcomes, between 2006 and 2015. We achieved an expected accuracy of 92% on the binary classification of irregularities, and our results endorse Ferraz et al.. findings: students in municipal schools perform significantly worse on standardized tests in municipalities where was found corruption in education.

Resumo

Estudos quantitativos em corrupção política enfatizam a falta de informações objetivas nessa área de pesquisa. O Programa de Fiscalização por Sorteios Públicos da CGU se baseia em auditorias não anunciadas das transferências do Governo Federal para municípios, e aparece como uma potencial solução para essa lacuna. Relatórios gerados durante essas auditorias descrevem com detalhe práticas de corrupção e de má gestão pública. No entanto, a análise manual desses relatórios é penosa e requer o conhecimento de especialistas. Nós propomos um *framework* estatístico para guiar o uso desses dados textuais na construção de indicadores objetivos de corrupção e em modelos de inferência. O *framework* consiste em duas etapas gerais. Na primeira, usamos métodos de aprendizagem de máquinas para classificação das irregularidades constatadas durante as auditorias. Na segunda etapa, construímos um indicador de corrupção baseado na classificação e o utilizamos em um modelo de regressão, ajustando pelo erro de medida derivado da primeira etapa. Para validar essa metodologia, nós replicamos a estratégia empírica apresentada por Ferraz et al. (2012) para estimar o efeito da corrupção em fundos educacionais nos resultados escolares de alunos do Ensino Fundamental, entre os anos de 2006-2015. Nós obtemos uma acurácia média de 92% na classificação binária de irregularidades, e nossos resultados corroboram com os encontrados em Ferraz et al.: estudantes de escolas municipais apresentam resultados significativamente piores em testes padronizados se estudam municípios com indícios de corrupção na área de educação.

Contents

1. Introduction

Political corruption has been a major issue in Brazil over the last few years. On the news, almost everyday a Business Executive or a Congressman is indicted in Brasilia, Rio de Janeiro or São Paulo. Most are part of scandals involving bribes with so many zeros that goes beyond our understanding - and it might be just the tip of the iceberg. On the bright side, Brazilians were taken by an apparent wave of willingness to fight against corruption. As researchers in Brazil, we know how we can start our contribution: assisting on the disclosure and efficient use of information about corruption. For us it is clear that, reliable data and robust modelling approaches are essential for public awareness and for guiding effective public policies to help extinguish corruption.

A considerable challenge in quantitative studies in corruption relies on how to measure it, since corrupt acts are, by definition, not easy to be traced and accounted for. The use of perception indicators is widespread among researchers, but their papers often include a note discussing its imprecision and emphasizing the need of more reliable and objective measurements of corruption. Current literature includes political, economical and social sciences approaches doing cross-national comparisons based on indices such as the Transparency International's Corruption Perception Index (CPI)[1] and the World Bank's CPIA Transparency, Accountability and Corruption in the Public Sector rating [2].

The Random Audits Anti-Corruption Program (*Programa de Fiscalização por Sorteios Públicos*) introduced by the Comptroller General of Brazil (CGU, henceforth) in 2003 appeared as a singular source of information to take quantitative analysis in corruption to an upper level. Reports generated by each audit describe irregular practices related to corruption and mismanagement in the use of Federal transfers in detail. In this study we propose a framework for working with text data in inferential modelling and we use information disclosed by this program as a direct application.

Our framework consists in two steps. First, we propose a machine learning approach to predict classes for a dataset of irregularities detected through CGU Anti-Corruption Program from 2006 to 2015. Next, we describe how one can use the resulting classified data to construct an indicator variable of corruption for audited municipalities and sequentially use it for inferential modelling. Since this variable carries an error from the text classification task, we use a Modified Least

[1] International Transparency (2017)

[2] World Bank (2017)

Squares (MLS) approach to adjust the coefficient estimate associated to the variable of corruption for measurement error. On the top of proposing a robust methodology, we have two additional goals: to release a dataset of categorized irregularities that can be used as objective measurement of corruption and to motivate any researcher interested in using text data in inferential models.

To exemplify the use of our framework, we choose a case study that uses manually coded CGU program's audit reports from 2003 to 2005 to extend. We take the empirical strategy proposed in~Ferraz et al. (2012) to estimate the effects of public corruption in students outcomes. We construct a binary variable of corruption in education based on the dataset of classified irregularities and use municipal and schooling data related to three consecutive mayoral terms, from 2005 to 2016.

This study is organized as follows. In the next section we describe the CGU Random Audits Anti-Corruption Program, its impact on research and the dataset of irregularities in which this study is based. Section 3 provides a brief background on the methods adopted in our framework. In section 4, we propose the framework and describe the case study used to exemplify it. Section 5 contains results for the classification task and the inferential model. In section 6 we conclude and propose next steps.

2. CGU Random Audits Anti-Corruption Program

In 2001, the Brazilian Federal Government instituted a national anti-corruption agency (*Corregedoria Geral da União*) to assist in activities of internal control, including public auditing and engagement against corruption. In 2003, its activities were embodied by the Comptroller General of Brazil (*Controladoria Geral da União*, CGU henceforth), created according to the Brazilian Law n. 10,683[3]. Since then, CGU is part of a federal government initiative aiming to evaluate the use of federal funds transferred to local governments to execute public programs: the Random Audits Anti-Corruption Program (*Programa de Fiscalização por Sorteios Públicos*).

This program had forty editions until 2015[4], having investigated 1,915 Brazilian municipalities[5]. The first edition started with a random sample of five municipalities and this number increased until sixty municipalities per round from the tenth edition onward. The program is designed to audit municipalities with small population size: depending on the edition, only municipalities with less than 100,000 or 500,000 inhabitants could be drawn. Municipalities within the threshold of 500,000 inhabitants represent about 70% of Brazilian population and more than 99% of the municipalities[6]. Each edition's sample is geographically representative, and municipalities are selected for auditing through the Brazilian lottery system - which is announced only a few days prior to the audit takes place.

For each municipality drawn, Service Orders are expedited to guide inspections in federal transfers associated to Ministries of Education, National Integration, Health, Cities, Social Development, Agricultural Development, Sport and others, according to the edition. The audit process includes physical and document analysis, photographic records and questionnaires/interviews with inhabitants. At the end of each audit, an extensive report - ranging from 30 to 200 pages long - is consolidated with a narrow description of anomalies detected and strict guidelines on how to deal with them. Reports are sent to respective ministries and published on CGU webpage[7]. An example of Service Order can be found on Appendix A.

These reports are so rich in information on the misuse of public funds, that relevant analyses on side effects of corruption were possible through their use as data

[3] For institutional details about CGU, see: http://www.cgu.gov.br/sobre/institucional/historico

[4] Since 2015, it was transformed into a broader program that selects municipal and state governments for inspection, referred in portuguese as *Programa de Fiscalização em Entes Federativos*. We choose to limit our analysis to the original program, for consistency.

[5] 2,241 audits were already conducted within this program, but some municipalities were selected in more than one edition.

[6] Instituto Brasileiro de Geografia e Estatística (2017)

[7] Look for "Relatório de Fiscalização Sorteio de Municípios" on the field "Título do Relatório" on: https://auditoria.cgu.gov.br/

source. Ferraz and Finan (2008) were pioneers and discussed its reliability towards capturing corrupt activities. In addition to the unadvertised aspect[8] of the audits, they also found evidence that municipalities audited before or after local elections were not subject to different auditing processes and showed that "(...)mayors with more political power, those affiliated with higher levels of government, and those who obtained larger campaign contributions did not receive preferential audits." (Ferraz and Finan, 2008)[9]. In their study, they take advantage of the randomness of audits and timing of its reports disclosure to investigate whether electoral outcomes of candidates for reelection are affected when voters know about candidates' corrupt practices during their first mandate. In particular, they built an indicator based on the *number of violations associated with corruption* (Ferraz and Finan, 2008, p. 710) and compare municipalities audited before and after 2004 municipal elections, controlling for local media presence. They define corruption as "(...) any irregularity associated with fraud in procurement, diversion of public funds, or over-invoicing" (Ferraz and Finan, 2008, p. 710) and this definition is adopted in the other studies mentioned in this section and in our framework. Their findings suggest that public release of audit reports before local elections contributed to reduce the expected probability of reelection in 17% in municipalities where two or more anomalies associated to corruption were found. Since their study was restricted to mayors competing on 2004 elections, their analysis is restricted to 373 reports from municipalities that fit into this condition and were audited until July, 2005.

Ferraz and Finan (2011) investigate if electoral accountability - through the possibility of reelection - is associated to less corrupt practices among politicians. They coded audit reports for 496 municipalities drawn for audit until the beginning of 2004 to create three measures of corruption: *total amount of resources related to corrupt activities*, *number of irregularities related to corruption* and *share of service items associated with corruption*(Ferraz and Finan, 2011, p. 1283). Their study found evidence that mayors with reelection incentives are associated to a smaller amount of resources misused due to corruption than mayors with no incentives - on average 27%.

Ferraz et al. (2012) build three objective measures of corruption in educational resources transferred to local governments based on the audit reports: *proportion of items with corruption in education*, *share of audited resources with corruption in education* and *indicator of corruption in Education* (Ferraz et al., 2012, p.721).

[8] In fact, it is likely that the unadvertised aspect of the audits lost part of its strength after few years of the program's existence, as discussed by Lichand et al. (2016).

[9] For a detailed description of the CGU Anti-Corruption Program and its possible limitations, see Ferraz and Finan (2008) and Ferraz and Finan (2011)

This study groups indicators of corruption for 366 municipalities randomly selected from 10 lotteries. Their main empirical findings are that "(...) student test scores on a national standardized exam and pass rates are significantly lower, and dropout rates are significantly higher in municipalities where corruption is prevalent" (Ferraz et al., 2012). We give a detailed description of their empirical strategy in the Methodology section, since we propose to extend it using our corruption indicator.

Mondo (2016) creates a panel data set containing a corruption indicator based on the *count of all individual instances of corruption identified in the audit reports*(Mondo, 2016, p. 482) and relevant characteristics for 140 municipalities audited at least two times between 2003 and 2013[10]. Besides providing a unique dataset, she uses it to evaluate electoral accountability, analyzing if municipalities had a lower level of corruption when being audited for the second time, but found no statistical evidence.

Similarly, other analyses were accomplished through indicators based on manual coding of audit reports. Bologna and Ross (2015) used corruption indicators constructed by Ferraz and Finan (2011) and found evidence that corruption is related to a lower level of entrepreneurial business activities in Brazilian municipalities. Brollo and Troiano (2016) use audit reports to analyze men and women characteristics as policymakers and find that, on averagem, women are less likely to be involved in corruption activities, hire less temporary public employees in electoral years and have a lower probability of reelection, when compared to male mayors. Lichand et al. (2016) constructed indicators of corruption in health-related transfers to show that, although the program had effect on reducing irregularities associated to corruption, it was not accompanied by an improvement in municipal health basic services. In fact, they found that "public spending fell by so much that corruption per dollar spent actually increased" (Lichand et al., 2016, p. 1), supporting the "greasing-the-wheels" hypothesis about corruption(Méon and Sekkat, 2005).

Since 2016, CGU began to release a *on-demand* dataset containing irregularities found from the 20th edition of the Anti-Corruption Program, normally used for internal purposed. Two important studies relied on it to build corruption indicators. Caldas et al. (2016) used basic text mining techniques[11] to reclassify *severe* and *moderate*-level[12] irregularities found in audits conducted between 2006 and 2010

[10] Provided that at least one election for mayor was held and that the mayor in power during the first audit had a successor running for election (or was running for reelection) (Mondo, 2016, p. 480)

[11] Broadly speaking, they gathered a list of terms that signalize corruption and classified irregularities containing these terms as corruption (Caldas et al., 2016, p. 249)

[12] According to CGU criteria, which is not explicitly released.

(20th to 33rd editions) according to corruption. Their indicator consists in the *share of irregularities classified as corruption*, and they found that, in Brazilian municipalities, higher levels of corruption are usually associated to a higher amount of public expenditure towards Health and Education. These findings might be related to Brazilian discretion laws on public spending. Avis et al. (2016) take irregularities classified as *severe* and *moderate* by CGU and classify as evidence of corruption. They calculate the *number of corrupt acts per Service Order*(Avis et al., 2016) for municipalities audited between 2006 and 2013 (22nd to 38th editions) and found evidence that the program has contributed to reduce corruption in municipalities previously audited in past editions of the program.

2.1. Dataset of Irregularities from Audit Reports

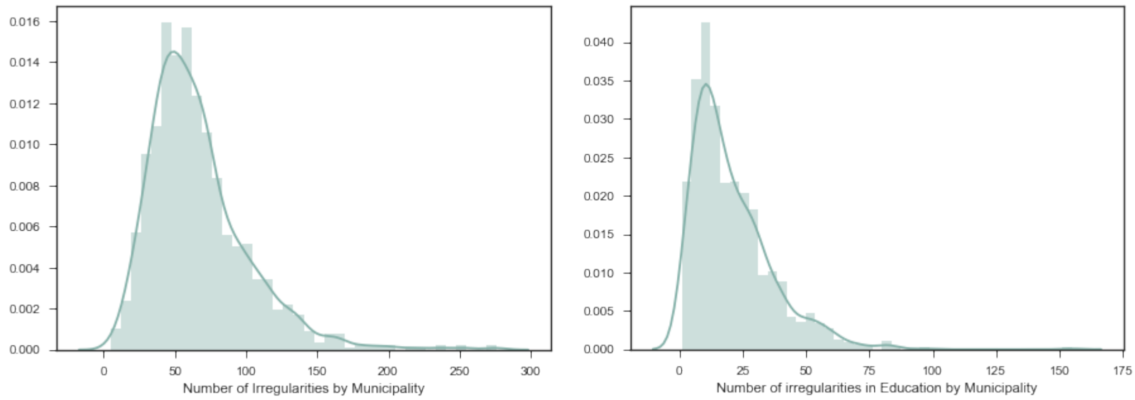
The first version of this project was based on the idea of extracting irregularities from each audit report published by CGU. We automatically[13] downloaded 2,204 reports from the CGU website and began to study its vocabulary and document structure to extract information. Conveniently, after this process we got access to a dataset maintained by CGU for internal use, released *on-demand* through the Access to Information Law (LAI). It contains information on irregularities found in each Service Order (OS) expedited for audits carried from the 20th to the 40th edition of the Anti-Corruption Program. It has basic details of each OS, such as the amount of resources investigated, the related ministry and program, a short description of each irregularity and its severity level (*severe*, *moderate* or *administrative*). In total, we have 81,715 irregularities found through execution of 29,821 OS in 1,223 municipalities randomly selected for audit between 2006 and 2015.

On average, 24 OS are dispatched for each municipality and an average of 3,7 irregularities by OS are reported. 85% of the irregularities found are associated to OS expedited for transfers from the ministries of Health, Education and Social Development. Considering any year, an average of R$13 million in federal transfers are audited in each municipality. On average, 66.8 irregularities are found per municipality and Education accounts for almost a third of it, with an average of 20.7. Figure 1 shows the distribution of the number of total irregularities and specifically in Education.

Above all, we are interested in two variables from this dataset: the short description and the severity level. Avis et al. (2016); Caldas et al. (2016) use the same dataset, but for a smaller period, to construct indicators of corruption based on *moderate* and *severe* irregularities. Based on the analysis of a sample from this

[13] Through a web crawler, available at https://github.com/lauragualda/CorruptionProject

Figure 1. Number of irregularities by municipality, total and in Education



data, we believe that this classification might not be the best separator of irregularities associated to corruption. For instance, among the 81, 715, only 7% are classified as *administrative*, the majority (79%) is *moderate* and 14% are *severe*. Thus, most of the irregularities could be misclassified as corruption. We found some irregularities where the the short description has equal meaning (and often the exact same text), but different severity level. For example, an irregularity associated to the Brazilian government social welfare program *Bolsa Família* "Famílias beneficiárias não localizadas"[14] was classified as *moderate* when found in Ibirapuã (BA) and as *severe* in Presidente Tancredo Neves(BA), both in the 28th edition of the CGU program. Similarly, the irregularity "Famílias beneficiárias do Programa Bolsa Família não localizadas"[15] was found in Goianésia(PA) on the 40th edition, and classified as *administrative*.

Therefore, we chose to take a different approach. Our goal is to categorize irregularities according to their short description into well delimited classes, following a structure similar to Lichand et al. (2016). We firstly defined twelve characteristic classes, but for the supervised learning task they were grouped in 10 broader classes and later into two: evidence of corruption or evidence of mismanagement. On Appendix A we provide a small description and one example of irregularity assigned to each of the classes. For instance, we follow Ferraz and Finan (2008) and consider classes "Fraud in procurement" (FL), "Diversion of public funds" (DR) and "Over-invoicing" (SF) as related to corruption, On Appendix A, we highlight examples of similar irregularities assigned to one of these three categories but classified into different severity levels by CGU.

[14] "Families benefiting not located"

[15] "Families benefiting from Bolsa Família program not located"

3. Background

In this section we introduce the tasks used in our framework, namely: supervised learning for classification of text data and approaches for correcting measurement error in binary explanatory variables in regression models.

3.1. Supervised Learning for Text Classification

Relevant studies in corruption were already developed using the reports from CGU audits, but using machine learning techniques a lot can be done to improve the use of text data in this research area. The task of reading and manually coding reports requires special training and an impractical amount of time and attention. Therefore most researchers resorted to taking a random subsample from the total reports. It is an option, but not the best: one can be missing useful information from the rest of the sample. With the growing use of machine learning to automate repetitive decision-making tasks, the dataset of irregularities from the CGU Anti-Corruption program is a good use case for trying an automated classification process. In this section, we briefly introduce a machine learning supervised approach for text classification.

We typically resort to machine learning algorithms when we want to predict an outcome measurement - either quantitative or categorical - using a set of available features. In a supervised learning problem, we have a training set, where both outcome and features are observable for all observations and we use it to build a prediction model (or learner), to try to predict the outcome for unseen objects. We want to train a model as accurate as possible when making predictions for unseen observations. (Hastie et al., 2001, p. 1-2)

Broadly speaking, text classification is the task of classifying documents $d \in \mathbb{X}$ into elements of a set of classes $\mathbb{C}$. When we have a set of documents, $\mathbb{D} \subset \mathbb{X}$, for which true outcomes are known, we can think of this task as a supervised learning problem. Using a machine learning algorithm, we use training subset to learn a classification function, h, that maps documents to classes using a set of features. We will take as example one irregularity detected in Curaçá (BA) in 2007, where the true class is "Procurement Issue" to illustrate this process:

(d, c) = (Frustração do caráter competitivo de licitação por cobrança abusiva para retirada do Edital e publicidade insuficiente.[16], *Procurement Issue*).

The first difference between a text and a numerical classification problem relies on the definition of features. A simple approach to extracting features from

[16] Self-translated as: "Restriction on the competitive nature of bidding due to improper charging for taking the Notice of Bidding and insufficient publicity."

a document is through a *bag of words* model, where each document is represented as a vector X of binary variables. In this case each element x_i represents a word from the documents' *corpus*, and $x_i = 1$ if the word exists in the document d and 0, if not. Natural extensions are N-gram models, where each element is a sequence of n ordered words, and a binary variable indicates its presence or absence in the document. It is often useful when the appearance order of words matter (eg. when a negative precedes a word). When non-text features are also available, when can also combine them to test their contribution on improving the task accuracy.

In all of these cases, pre-processing methods are necessary to transform raw structured text into useful information (Vijayarani et al., 2015). Common techniques that were performed in our framework, are:

- Tokenization: Split text into a set of single words (tokens or unigrams),
- Punctuation removal: Unless it characterizes specific sentiment (eg.: !!!), punctuation is usually coincidental across documents,
- Stopwords removal: Removal of most common words, articles, prepositions from the corpus (eg. in Portuguese: a, o, para). Besides regular ones, it is also important to note specific stopwords for each particular corpus,
- Stemming: Conversion of words into a root form (eg. in Portuguese: *contrato* and *contratual* should be consider the same token, *contrat*).

Take our example given above. After pre-processing (following Portuguese patterns), we get the following features according to a *bag of words* and a bigram model:

[*'frustração', 'caráter', 'competitivo', 'licitação', 'cobrança', 'abusiva', 'retirada', 'edit', 'publicidad', 'insuficient', ('frustração', 'caráter'), ('caráter', 'competitivo'), ('competitivo', 'licitação'), ('licitação', 'cobrança'), ('cobrança', 'abusiva'), ('abusiva', 'retirada'), ('retirada', 'edit'), ('edit', 'publicidad'), ('publicidad', 'insuficient')*]

The final representation of the features space is a matrix, where each row is a document and each of the columns is a feature (in our example, a word or a bigram). Depending on the richness and variety of the vocabulary and length of the documents, it is efficient to store it as a sparse matrix, since most of the elements will be zero. When a *corpus* has too many words, it is convenient to adopt a strategies for feature selection, varying according to the machine learning algorithm.

In machine learning tasks it is convenient to consider and compare several learning algorithms, since each performs differently depending on the the nature of data. This comparison is usually based on evaluation metrics obtained through k-fold cross-validation on training set. The k-fold cross-validation process consists in splitting the training set in k subsets and for each of the k, we use $k-1$ folds to train a model and the remaining as test.

Performance metrics used in our framework are conveniently illustrated through a confusion matrix. Consider the following one representing a Yes-No classification problem:

		Predicted Label	
		Yes	No
True label	Yes	True Positive	False Negative
	No	False Positive	True Negative

We are interested in three performance metrics: accuracy, precision and recall. Accuracy is simply the proportion of observations for which classes were correctly assigned (True Positive + True Negative/Total). Precision is the proportion of observations assigned to a class that truly belong to it (True positive/True Positive + False Positive) and recall is the proportion of observations correctly assigned to a class given the total of observations that belong to it (True Positive/True Positive + False Negative) Manning et al. (2008). Once we evaluate the performance of each classifier on the training set, we choose one to be applied on the test set and obtain an estimation of its out-of-sample error.

3.1.1. Machine Learning classifiers

In this section, we briefly describe the machine learning algorithms used in our framework. Since we want to motivate its use among researchers in any area, we emphasize two simple and intuitive probabilistic models used for classification tasks: Naive Bayes and Multinomial Logistic Regression. Besides those, we summarize the idea behind the other three classifiers used for benchmark, namely Support Vector Machines for classification and two ensemble methods: Random Forests and Voting.[17]

Naive Bayes

Consider a training set of N pairs (d, c) , where d represents a document and c a class. A Naive Bayes learned model attributes to each d on the test set a class

[17] It is not in the scope of this work to precisely demonstrate each of these algorithms. For more details see Abu-Mostafa et al. (2012).

c^* with the highest posterior probability $\mathbb{P}(c^*|d)$, ie:

$$c^* = \underset{c_i \in \{c_1,\ldots,c_K\}}{\operatorname{argmax}} \mathbb{P}(c_i|d). \tag{3.1}$$

Define V as the *corpus* and X as the features vector. A given $x_i \in X$ takes 1 if the corresponding feature w_i occurs in the document and 0 otherwise. Note that by Bayes theorem $\mathbb{P}(c|d)$ can be written as:

$$\mathbb{P}(c|d) = \frac{\mathbb{P}(d|c)\mathbb{P}(c)}{\mathbb{P}(d)}, \tag{3.2}$$

which implies that:

$$\begin{aligned} c^* &= \underset{c_i \in \{c_1,\ldots,c_K\}}{\operatorname{argmax}} \mathbb{P}(c_i|d) \\ &= \underset{c_i \in \{c_1,\ldots,c_K\}}{\operatorname{argmax}} \frac{\mathbb{P}(d|c_i)\mathbb{P}(c_i)}{\mathbb{P}(d)} \\ &= \underset{c_i \in \{c_1,\ldots,c_K\}}{\operatorname{argmax}} \mathbb{P}(d|c_i)\mathbb{P}(c_i). \end{aligned} \tag{3.3}$$

It means that, we only need to estimate $\mathbb{P}(d|c_k)$ and $P(c_k)$ for $k = 1, ..., K$ in order to classify each document. The main assumption of the Naive Bayes model is that features are independent in each given class (hence the term "Naive"), which simplifies our estimation problem. Thus, using Bayes Theorem we can write $\mathbb{P}(d|c_k)$ as:

$$\begin{aligned} \mathbb{P}(d|c_k) &= \mathbb{P}(X|c_k) \\ &= \mathbb{P}(x_1, ..., x_{|V|}|c_k) \\ &= \prod_{i=1}^{|V|} x_i\mathbb{P}(w_i|c_k) + (1 - x_i)(1 - \mathbb{P}(w_i|c_k)), \end{aligned} \tag{3.4}$$

where $P(w_i|c_k)$ is the probability that the word w_i exists in the document given its class is c_k. It is estimated by computing the ratio between the number of times w_i appeared in the documents of class c_k and the number of such documents in the training set. $P(c_k)$ is the prior probability of the class k, and it can be estimated by the fraction of documents of class k and the total of documents in the training set. After estimating $P(c_k)$ and $P(w_i|c_k)$ for $k = 1, \cdots, K$, the algorithm assigns to each document d a class c^* solving equations (4) and (3).

Even though the hypothesis of conditional independence of features behind the Naive Bayes classifier might seem too strong in our context, we chose it due to its simplicity and relatively fair performance, when comparing to sophisticated classifiers (Hastie et al., 2001, p. 211).[18]

Multinomial Logistic Regression

Instead of using Bayes Theorem to calculate $\mathbb{P}(c|d)$, the Logistic Regression estimate the conditional probability directly, modelling the posterior probabilities as linear functions in the feature vector X:

$$\log \frac{\mathbb{P}(c_i|X = x)}{\mathbb{P}(c_K|X = x)} = \beta_{i0} + \beta_i^T x, \tag{3.5}$$

where $i = 1, 2, ..., K - 1$.
From (5), we have :

$$\mathbb{P}(c_K|X = x) = \frac{1}{1 + \sum_{\ell=1}^{K-1} \exp(\beta_{\ell 0} + \beta_\ell^T x)}, \tag{3.6}$$

for C_K and for $i = 1, 2, ..., K - 1$:

$$\mathbb{P}(c_i|X = x) = \frac{\exp(\beta_{i0} + \beta_i^T x)}{1 + \sum_{\ell=1}^{K-1} \exp(\beta_{\ell 0} + \beta_\ell^T x)}. \tag{3.7}$$

Then, a Multinomial Logistic Regression is fit maximizing the likelihood function:

$$\ell(\beta) = \sum_{i=1}^{K} \log \mathbb{P}_{c_i}(x_i, \beta), \tag{3.8}$$

where $\mathbb{P}_k(x_i, \beta) = \mathbb{P}(c_k|X = x, \beta)$[19].

The Multinomial Logistic Regression is a another simple probabilistic model and it usually outperforms the Naive Bayes classifier, since it does not rely on the hypothesis that features are independent inside each class. Particularly in a text classification problem, we expect that features inside each class are not independent, since they represent the occurrence of words (or n-grams).

Support Vector Machines

Support Vector Machines (SVM) learning methods were introduced by Vapnik (1995) and are based on the general task of finding a hyperplane that best separates observations based on a set of features. They are designed to handle

[18] For a detailed description and examples on the Naive Bayes classifier, see (Hastie et al., 2001, p. 210-211).

[19] Note that there is no closed solution for the maximum likelihood estimation.

high-dimensional data, therefore, consistently achieve good accuracy on text classification tasks at a lower computational cost and easier tuning than other established classifiers (Joachims, 1998).

Random Forest Ensemble Method

Random Forests are a class of ensemble methods proposed by Breiman (2001) that uses bootstrap sampling to build different Decision Trees and combine their predictions to construct a stronger classifier. It takes a considerable time to train this model and its predictions are usually not interpretable, but it has the advantage of maintaining low bias of a single Tree while reducing variance, thus reducing chances of overfitting.

Voting Ensemble Method

The process of classification performed by a Voting ensemble method is pretty intuitive: it classifies each observation based on the majority voting of a set of different machine learning classifiers. Particularly in this work, we performed a *soft voting* considering the four classifiers mentioned above. In that case, this model assigns to observation y the predicted class c_i that maximizes the combination of probabilities calculated by each of the four classifiers, given a set of weights:

$$\hat{y} = \underset{i}{\operatorname{argmax}} \sum_{m=1}^{4} \omega_m p_{im}.$$

3.2. Measurement Errors in Binary Explanatory Variables

The second part of our framework consists in using a dataset of categorized irregularities to construct an indicator of corruption in Brazilian municipalities. This indicator is based on the binary classification of each irregularity as associated to *corruption* or not (*mismanagement*, following the literature). The problem is, we cannot simply assume that an indicator constructed from this data is accurately measured. It is based on irregularities classified according to a machine learning process and we know that predictions are not 100% accurate. Therefore, to use this indicator in any inferential model, we must resort to econometric methods for correcting the measurement error bias.

In this work, we focus on the construction of a binary indicator of corruption. Thus, in this section we derive the bias in the slope estimator $\hat{\beta}_{OLS}$ in a linear regression when a binary explanatory variable is subject to measurement error and present established approaches to deal with it.

Following Aigner (1973); Bollinger (1996); Johnston (1997); Savoca (2000), we begin by examining the case of a simple linear regression:

$$y = \beta x + \epsilon, \tag{3.9}$$

where y is a continuous variable and x is measured with an error u such that $z = x + u$. In our context, we observe z: an indicator variable of irregularities associated to corruption, classified according to a supervised learning task. There are four possible scenarios for x, z, u:

Table 1. Measurement error possibilities

z	**x**	**u**
1	1	0
1	0	1
0	1	-1
0	0	0

Let r_1 be the probability of observing $z = 1$ when $x = 0$ (false positive) and r_0 the probability of $z = 0$ when $x = 1$ (false negative). The conditional probabilities for u are:

$$\begin{aligned}
\mathbb{P}[u = 1|x = 0] &= \mathbb{P}[z = 1|x = 0] = r_{10}, \\
\mathbb{P}[u = 0|x = 0] &= \mathbb{P}[z = 0|x = 0] = 1 - r_{10}, \\
\mathbb{P}[u = -1|x = 1] &= \mathbb{P}[z = 0|x = 1] = r_{01}, \\
\mathbb{P}[u = 0|x = 1] &= \mathbb{P}[z = 1|x = 1] = 1 - r_{01}.
\end{aligned}$$

Now consider P_x the proportion of $x = 1$ in the population and P_z the proportion on the sample subject to mismeasurement. Then the error expectation is given by:

$$\mathbb{E}[u] = r_{01}(1 - P_x) - r_{10}P_x = r_{10} - (r_{01} + r_{10})P_x, \tag{3.10}$$

which implies that $\mathbb{E}[u] = 0 \quad \Leftrightarrow \quad P_x = \frac{r_{10}}{r_{01}+r_{10}}$, classical errors-in-variables models (where $\mathbb{E}[u] = 0$). When this is not the case, the sample proportion is a biased estimate for the population proportion. Indeed, we can calculate an expression for the covariance of x and u, and observe that they are negatively correlated:

$$Cov(x, u) = -r_{10}P_x - P_x[r_{01} - (r_{01} + r_{10})P_x] = -(r_{01} + r_{10})P_x(1 - P_x). \tag{3.11}$$

Similarly, we can write an expression for the variance of u:

$$Var(u) = r_{10} + (r_{01} - r_{10})P_x - [r_{10} - (r_{01} + r_{10})P_x]^2. \tag{3.12}$$

Now, we can rewrite the regression (3.9) as:

$$y = \beta z + (\epsilon - \beta u), \tag{3.13}$$

and the estimator $\hat{\beta}_{OLS}$ has probability limit given by:

$$\begin{aligned} plim\hat{\beta}_{OLS} &= \beta + \frac{Cov(z, \epsilon - \beta u)}{Var(z)} \\ &= \beta - \beta\frac{Cov(z, u)}{Var(z)} \\ &= \beta\Big[\frac{Var(x+u) - Cov(x+u, u)}{Var(z)}\Big] \\ &= \beta\Big[\frac{Var(x) + Var(u) + 2Cov(x, u) - Cov(x, u) - Var(u)}{Var(z)}\Big] \\ &= \beta\Big[\frac{Var(x) - Cov(x, u)}{Var(z)}\Big] \\ &= \beta\Big[\frac{P_x(1-P_x)(1-r_{01}-r_{10})}{P_z(1-P_z)}\Big]. \end{aligned} \tag{3.14}$$

As proven by Aigner (1973), Bollinger (1996) and Savoca (2000), the slope estimator asymptotically tends to zero as the sum of classification errors get closer to one. Aigner (1973) proved that the absolute value of the bias term is always less than one and when more than half of the population were misclassified in the observable variable, the OLS coefficient would have opposite sign.

Savoca (2000) presents an extensive review of methods for correcting measurement error bias in binary variables. The author comments that, the obvious solution of instrumental variables to treat the correlation between the explanatory variable and the error term would only lead to a consistent estimator for β if information on misclassification rates were available. But still, finding a instrument for the classification error in binary variables can be rather difficult, if possible. For cases when information on mismeasurement rates are unavailable, Klepper (1988) has derived consistent bounds for the true slope based on the strong assumption that $r_{01} = r_{10} < 0.5$. Bollinger (1996) proposed tighter bounds under the assumption that $r_{01} + r_{10} < 1$. Another alternative is to incorporate the misclassification rates into the likelihood function to obtain asymptotically efficient estimators.

The last and preferred alternative outlined by Savoca (2000) is the modified least squares estimator (MLS), first proposed by Johnston (1997) in his book for econometric applications. It requires extra-sample information on misclassification rates to estimate the covariance matrix between the observable variable z and the error, u. We adopt this approach in our framework, so we present derivations proved by Johnston (1997) in the format of Savoca (2000).

Consider the original regression model, the mismeasurement model and its corresponding regression model in matrix form, where Y and E are $N \times 1$, Z and U are $N \times K$ and β is $K \times 1$:

$$\begin{aligned} Y &= X\beta + E \\ Z &= X + U \\ Y &= Z\beta + E - U\beta. \end{aligned} \tag{3.15}$$

The OLS estimator for β is given by:

$$\hat{\beta}_{OLS} = (I - (Z'Z)^{-1}Z'U)\beta + (Z'Z)^{-1}Z'E. \tag{3.16}$$

Given E independent of X and U and $\Omega = \Sigma_{ZZ}^{-1}\Sigma_{ZU}$ we have:

$$plim(\hat{\beta}_{OLS}) = (I - \Omega)\beta. \tag{3.17}$$

We can use sample moments to estimate Σ_{ZZ}^{-1} and extra-sample data to estimate Σ_{ZU}, to get the consistent estimator for β is a linear regression subject to mismeasurement in a binary independent variable:

$$\hat{\beta}_{MLS} = (I - \hat{\Omega})^{-1}\hat{\beta}_{OLS} \tag{3.18}$$

Where $\hat{\beta}_{MLS}$ is the vector with slope estimators corrected for mismeasurement error in binary variables. To compute an estimate for the standard error of $\hat{\beta}_{MM}$ and evaluate the coefficient statistical significance, we can use the delta method.

4. Methodology

We propose a two-step framework for working with text data in econometric models. The first step comprises Automate Text Classification tasks, such as manual classification, data mining and supervised learning to classify a set of irregularities from CGU Anti-Corruption Program. The second one consists in constructing a binary variable of corruption for municipalities based on the classified irregularities and using it in inferential models, correcting for the measurement error carried from the first step. To illustrate and evaluate the framework we replicate an empirical strategy proposed by Ferraz et al. (2012). The authors use Audit reports from CGU Anti-Corruption program to estimate the effect of "corrupt use of education public funds" (Ferraz et al., 2012) on students outcomes in municipal schools. In this section we describe the framework in detail and present the model specification and data used on the case study.

4.1. The Framework

Step 1: Automate Text Classification

The input for this step is the dataset of 81,715 irregularities and the list of classes described in Section 2. We wish to learn a function that maps each irregularity to one class with a reasonable accuracy. As output, we expect to have each of the documents - in this work, irregularities found in the audited municipalities - assigned to a class and to calculate an estimate of the classification error. This stage can be split in two parts:

(i) Manual Classification of Irregularities

Our goal is to perform supervised learning for classification. The first step consists in building a representative sample of documents classified according to Appendix A and Ferraz and Finan (2008); Lichand et al. (2016). For this study, we randomly selected 200 irregularities from each edition of the program, adding up to 4,200, or approximately 5% of the entire dataset.[20]

(ii) Supervised Learning for Text Classification

Once we have a sample of classified irregularities, we begin the task of supervised learning. First, it is important to separate the unclassified part of our data - for which we do not know the classes and will only apply the learned model by the end of this stage - and the classified sample, that will be used for training, testing

[20] The process of manual classification was taken entirely by the author to reduce variance, but indeed might be subject to a personal bias. For future studies, a review in this step is recommended.

and evaluating the machine learning models. The classified sample must be split into train and test (in our case, 70%/30%), taking in account classes weights.

We consider four models as described in Section 3: Naive Bayes, Multinomial Logistic Regression, SVM and Random Forest. In addition to that, we also used a Voting ensemble method, combining the average predicted probabilities calculated by the four original classifiers for prediction. We perform the following steps:

1. Text mining: text pre-processing, exploratory analysis, feature generation
2. Perform 10-fold cross-validation for choosing each algorithm hyper-parameters (eg. number of text features, regularization term, n-gram range). Evaluate model performance when non-textual features are also included.
3. Train each classifier using the hyper-parameters that minimize the cross-validation error.
4. Compute the accuracy of each classifier using 10-fold cross-validation.

After calculating the performance of each model on the training set, we select one of them to predict classes for the entire dataset. We evaluate its performance on the test set and compute the confusion matrix to calculate accuracy, precision and recall to be used on the next step of the framework.

The joint probability distribution obtained from the confusion matrix is particularly important for the model estimation step. Let $\widehat{D_k}$ $(1 \leq k \leq 81,715)$ be the predicted class of irregularity k and D_k the true class. In a binary classification problem, it will be convenient to write the joint probabilities in the normalized confusion matrix, as in:

		Predicted label	
		0	1
True label	0	π_{00}	π_{01}
	1	π_{10}	π_{11}

Step 2: Model Estimation with Measurement Error

From the previous step, we have a set of irregularities, their classes predicted by a machine learning classifier and an estimate of the classification error. We want to use the predicted irregularities' classes to construct an indicator variable for corruption, C_i, that is 1 if at least one irregularity associated to corrupt practices was found in municipality i, and 0 otherwise. Once we have this variable, aggregated at the municipal level, we would like to use it as explanatory variable in a linear regression model. Since it is subject to measurement error from the classification step, its associated coefficient must be corrected. We divide this part of the framework in three steps:

(i) Constructing an indicator variable of Corruption

Our first goal is to build a indicator variable of corruption for each municipality i based on the irregularities found during their audit. Note that, for each municipality i we have N_i irregularities D_j $(1 \leq j \leq N_i); \sum N_i = 81,715)$ classified as 1 (associated to corrupt practices) or 0. An indicator variable of corruption for municipality i is:

$$C_i = \begin{cases} 1, & \text{if } \exists \ \ D_{i,j} = 1 \\ 0, & \text{otherwise.} \end{cases} \tag{4.1}$$

Next, we need an estimator for the measurement error carried from the automate classification step. Let $\widehat{C_i}$ be the predicted indicator of corruption and C_i the true label for municipality i. We can calculate the false positive rate $r_{10,i}$ for each municipality i as a function of the confusion matrix obtained in the previous step:

$$\begin{aligned} r_{10,i} = \mathbb{P}(\widehat{C_i} = 1 | C_i = 0, N_i) &= \mathbb{P}(\exists \{\widehat{D_{ij}} = 1\} | \nexists \{D_{ij} = 1\}) \\ &= 1 - \mathbb{P}(\nexists \widehat{D_{ij}} = 1 | \nexists D_{ij} = 1) \\ &= 1 - \mathbb{P}\Big(\bigcap_{j=1}^{N_i} \{\widehat{D_{ij}} = D_{ij}\} | D_{ij} = 0 \ \ \forall D_{ij} \Big) \\ &= 1 - \prod_{j=1}^{N_i} \mathbb{P}(\widehat{D_{ij}} = D_{ij} | D_{i1} = 0, D_{i2} = 0, \cdots, D_{iN_i} = 0) \\ &= 1 - \prod_{j=1}^{N_i} \frac{\mathbb{P}(\widehat{D_{ij}} = 0 \cap D_{ij} = 0)}{\mathbb{P}(D_{ij} = 0)} \\ &= 1 - \Big(\frac{\pi_{00}}{\pi_{00} + \pi_{10}} \Big)^{N_i}. \end{aligned} \tag{4.2}$$

Taking the expected value on N, we obtain the expected false positive rate r_{10} for all municipalities:

$$\begin{aligned} r_{10} &= \mathbb{E}[\mathbb{P}(\widehat{C}_i = 1 | C_i = 0, N)] \\ &= \mathbb{E}\Big[1 - \Big(\frac{\pi_{00}}{\pi_{00} + \pi_{10}}\Big)^N\Big] \\ &= 1 - \mathbb{E}\Big[\Big(\frac{\pi_{00}}{\pi_{00} + \pi_{10}}\Big)^N\Big]. \end{aligned} \tag{4.3}$$

Analogously, the expected false negative rate (r_{01}) is:

$$r_{01} = 1 - \mathbb{E}\Big[\Big(\frac{\pi_{11}}{\pi_{11} + \pi_{01}}\Big)^N\Big]. \tag{4.4}$$

A natural question is how to estimate these expected misclassification rates. The first and most natural way to do it is to estimate these quantities directly on the text classification task. We calculate $r_{01,i}$ and $r_{10,i}$ from the confusion table constructed using the predicted $\widehat{C}_i$ and the true value C_i for each municipality i. Then, we estimate r_{10} by $\widehat{r}_{10} = \frac{1}{K}\sum_{i=1}^{n} r_{10,i}$ and $\widehat{r_{01}}$ analogously. We suggest the following cross-validation scheme:

1. Group irregularities by municipality;
2. Divide the municipalities, randomly, in K groups $\mathcal{G}_k$, $k = 1, ..., K$;
3. Perform K-fold cross-validation selecting $K - 1$ groups of municipalities each round;
4. At each round k, predict $\widehat{C}_i$ for $i \in \mathcal{G}_k$ using the model trained on $\cup_{j \neq k}^{K} \mathcal{G}_j$;
5. Calculate the confusion matrix for predicting C_i and obtain $r_{10,k}$ and $r_{01,k}$;
6. Calculate $\widehat{r}_{10} = \frac{1}{K}\sum_{k=1}^{K} r_{10,i}$ and $\widehat{r}_{01}$ analogously.

This approach is equivalent to performing cross-validation in a post-stratified sample by municipality. We are still estimating a model for the irregularities, the only change is how the cross-validation is performed. Using this scheme, we predict C_i for each municipality "out-of-sample", thus avoiding optimism bias in the construction of the confusion table. Note that π.s are not needed since we estimate r_{10} directly.

Another way to estimate (4.3) and (4.4) is to assume a distribution for N. The advantage of this approach is that we can have smaller variance and a larger sample of municipalities to use. Here, we make the following assumption: *the number of irregularities is independent and identically distributed across municipalities.* [21] We assume the number of irregularities follow a Negative Binomial distribution with parameters α and p, i.e., $N \sim NB(\alpha, p)$, where $\alpha > 0$ and $0 < p < 1$.

[21] This condition can be replaced by "*the number of irregularities is independent and identically distributed within each ministry across municipalities.*"

The Negative Binomial distribution can be written as a Poisson-Gamma mixture, i.e., $N|\lambda \sim \mathsf{Poisson}(\lambda)$ and $\lambda \sim \mathsf{Gamma}(\alpha, \frac{p}{1-p})$. To be more precise in our calculations, we took only municipalities that have at least one irregularity in Education, then we should drop $N = 0$. We define $\tilde{N} = N - 1$ and suppose $\tilde{N} \sim \mathsf{NegativeBinomial}(\alpha, p)$. Using linearity and the law of iterated expectation, $\mathbb{E}[q^{\tilde{N}+1}] = q\,\mathbb{E}[\mathbb{E}[q^{\tilde{N}}|\lambda]]$, where the expectation inside is taken with respect to $\tilde{N}|\lambda \sim \mathsf{Poisson}(\lambda)$

$$\mathbb{E}[q^{\tilde{N}}|\lambda] = \sum_{n=0}^{\infty} q^n \frac{e^{-\lambda}\lambda^n}{n!} = \frac{e^{-\lambda}}{e^{-q\lambda}} \sum_{n=0}^{\infty} \frac{e^{-q\lambda}(q\lambda)^n}{n!} = e^{-\lambda(1-q)}.$$

The random variable $\lambda \sim \mathsf{Gamma}(\alpha, \frac{p}{1-p})$, then,

$$\begin{aligned}
\mathbb{E}[e^{-\lambda(1-q)}] &= \int_0^{\infty} e^{-\lambda(1-q)} \frac{\lambda^{\alpha-1} e^{-\lambda \frac{1-p}{p}}}{\Gamma(\alpha)\left(\frac{p}{1-p}\right)^{\alpha}} \mathrm{d}\lambda \\
&= \left(\frac{1-p}{p}\right)^{\alpha} \left(\frac{p}{p(1-q)+(1-p)}\right)^{\alpha} \int_0^{\infty} \frac{\lambda^{\alpha-1} e^{-\lambda \frac{p(1-q)+(1-p)}{p}}}{\Gamma(\alpha)\left(\frac{p}{p(1-q)+(1-p)}\right)^{\alpha}} \mathrm{d}\lambda \\
&= \left(\frac{1-p}{1-pq}\right)^{\alpha}. \qquad (4.5)
\end{aligned}$$

Hence, it follows that

$$\mathbb{E}\left[q^{\tilde{N}+1}\right] = q\,\mathbb{E}\left[\mathbb{E}\left[q^{\tilde{N}}|\lambda\right]\right] = q\left(\frac{1-p}{1-pq}\right)^{\alpha},$$

for any $0 < q < 1$.

The problem lies in choosing α and p. These values can be directly estimated using maximum likelihood from the whole set of irregularities grouped by municipality.[22] Suppose one estimates $\hat{\alpha}$ and $\hat{p}$ and can use the plug-in estimators:

$$\hat{r}_{01} = 1 - \pi_0 \left(\frac{1-\hat{p}}{1-\hat{p}\pi_0}\right)^{\hat{\alpha}} \quad \text{and} \quad \hat{r}_{10} = 1 - \pi_1 \left(\frac{1-\hat{p}}{1-\hat{p}\pi_1}\right)^{\hat{\alpha}}, \qquad (4.6)$$

where $\pi_0 = \frac{\pi_{00}}{\pi_{00}+\pi_{01}}$ and $\pi_1 = \frac{\pi_{11}}{\pi_{10}+\pi_{11}}$ are taken from the confusion matrix for the classification of irregularities.

[22] If using the condition that number of irregularities are independent and identically distributed across municipalities within each ministry, one has to count the number of irregularities within each ministry in each municipality.

(ii) Model Estimation adjusting for Measurement Error

Consider now that we have the following population regression model:

$$\begin{aligned} Y &= X\beta + E \\ &= [Z \quad : \quad W]\beta + E \\ &= Z\alpha + W\theta + E, \end{aligned} \tag{4.7}$$

where Y is a $N \times 1$ vector of observations on the dependent variable, X is a $N \times K$ matrix of explanatory variables and E is a $N \times 1$ vector of stochastic errors. We split the vector X in two parts for convenience: Z is a $N \times 1$ vector of observations from a binary variable indicating corruption and W is a $N \times (K-1)$ matrix with observations for the other $K-1$ explanatory variables. β is the vector of K coefficients, where α is the coefficient associated to variable Z and θ is a vector with coefficients for variables in W.

Our main interest is to have an estimate for α: the average effect of corruption on the variable Y. We can estimate this regression based on observable variables W and Z. But our observation of Z is the indicator of corruption C built from the automate classification step, therefore it is measured with an error u, according to:

$$C = Z + U_C \tag{4.8}$$

For controlling purposes we assume that observations for the $K-1$ variables in W have no measurement error. Thus we can write the matrix of mismeasured variables X_M as:

$$\begin{aligned} X_M &= X + U \\ &= [Z \quad : \quad W] + [U_C \quad : \quad 0] \\ &= [Z + U_C \quad : \quad W] \\ &= [C \quad : \quad W]. \end{aligned} \tag{4.9}$$

And the regression model based on observable variables can be written as:

$$\begin{aligned} Y &= (X_M - U)\beta + E \\ &= X_M\beta + E - U\beta. \end{aligned} \tag{4.10}$$

Therefore, we have measurement error in a binary variable in a multivariate framework. Since we have information on its misclassification rates, we can compute the modified least squares estimator (MLS) proposed by Johnston (1997), correcting

OLS estimate for β and obtaining a consistent estimator(Johnston, 1997; Savoca, 2000) in the presence of measurement error:

$$\hat{\beta}_{MLS} = (I - \hat{\Omega})^{-1}\hat{\beta}_{OLS}, \tag{4.11}$$

where $\hat{\Omega}$ is an estimator for $\Omega = \Sigma_{X_M}^{-1}\Sigma_{X_M,U}$.

Our purpose is to be able to make inference on the coefficient associated to the variable of corruption. Let $\hat{\alpha}_{MLS}$ be the MLS estimator for the coefficient associated to the mismeasured variable of corruption and θ_{MLS} a vector of coefficients for the other $K-1$ explanatory variables in W. We can rewrite our equation of adjustment in a partitioned matrix form for convenience:

$$\hat{\beta}_{MLS} = \left[\begin{array}{c} \hat{\alpha}_{MLS} \\ \hdashline \hat{\theta}_{MLS} \end{array}\right], \quad \hat{\beta}_{OLS} = \left[\begin{array}{c} \hat{\alpha}_{OLS} \\ \hdashline \hat{\theta}_{OLS} \end{array}\right],$$

$$\Sigma_{X_M}^{-1} = \left[\begin{array}{c:c} \sigma_C^2 & \sigma_{C,W} \\ \hdashline \sigma_{W,C} & \Sigma_W \end{array}\right]^{-1}, \quad \Sigma_{X_M U} = \left[\begin{array}{c:c} \sigma_{C,U_C} & \sigma_{C,0} \\ \hdashline \sigma_{W,U_C} & \Sigma_{W,0} \end{array}\right],$$

where we assumed that variables in W have no measurement error. Thus, $\sigma_{C,0} = 0$ and $\sigma_{W,0} = 0$. Besides, considering that the error U_C is a consequence of the automated classification task[23], it is not correlated with the idiosyncratic variables in W, ie. $\sigma_{W,U_C} = 0$.

Recall that we can calculate the inverse of $\Sigma_{X_M}^{-1}$ blockwise[24]. Thus term associated to the first block of $\Sigma_{X_M}^{-1}$ can be written as:

$$(\sigma_{C,W}\Sigma_W^{-1}\sigma_{W,C})^{-1}. \tag{4.12}$$

[23] Since public auditors are highly qualified professionals selected through a rigorous national public tendering, we expect that the vocabulary used in the reports does not vary across audits in the same edition.

[24] See Bierens HJ (2014):

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{bmatrix}$$

Therefore, $\hat{\alpha}_{MLS}$ can be calculated as:

$$\begin{aligned}
\hat{\alpha}_{MLS} &= (1-\hat{\Omega}_{11})^{-1}\cdot\hat{\alpha}_{OLS} \\
&= \left(1-\frac{\sigma_{C,U_c}}{\sigma_C^2-\sigma_{C,W}\Sigma_W^{-1}\sigma_{W,C}}\right)^{-1}\cdot\hat{\alpha}_{OLS} \\
&= \left(1-\frac{\sigma_{Z+U_c,U_c}}{\sigma_C^2-\sigma_{C,W}\Sigma_W^{-1}\sigma_{W,C}}\right)^{-1}\cdot\hat{\alpha}_{OLS} \\
&= \left(1-\frac{\sigma_{Z,U_c}\sigma_{U_c}^2}{\sigma_C^2-\sigma_{C,W}\Sigma_W^{-1}\sigma_{W,C}}\right)^{-1}\cdot\hat{\alpha}_{OLS},
\end{aligned} \tag{4.13}$$

where σ_{Z,U_c} and $\sigma_{U_c}^2$ can be estimated using out-of-sample information for P, r_{01}, r_{10} as shown in (3.12) and (3.11) and $(\sigma_{C,W}\Sigma_W^{-1}\sigma_{W,C})^{-1}$ can be estimated using the inverse of the sample covariance matrix.

4.2. Case Study: Does Corruption affect Educational Outcomes?

As an application of this framework we propose replicating the empirical analysis in Ferraz et al. (2012), that estimates the relationship between corruption in educational transfers and schooling outcomes. Our measure of corruption in is the main variable used by (Ferraz et al., 2012, p .722): an binary variable indicating detection of corruption in Education. We constructed it using the automated classification of audits' irregularities and other control variables were selected as described by Ferraz et al.. Our analysis has three main differences. We do not build variables of corruption based on the share of audited resources and proportion of irregularities with corruption[25]. Our sample covers a larger number of municipalities audited in three different political terms, from 2006 to 2015. Our study has a methodological focus: we restrict to statistical significance results and do not wish to extend the substantial discussion taken by Ferraz et al. on the causes and consequences of political corruption in Education.

4.2.1. Model Identification

To estimate how the corrupt use of public education resources affects students achievements, we take the exact identification strategy proposed by Ferraz et al. (2012). In their approach, the academic achievement of a student i, attending school s in municipality m in grade g ($A_{i,s,m,g}$) can be expressed as the sum of two components: a share (δ) of the academic achievement obtained until grade *g-1* ($A_{i,s,m,g-1}$) and a share (γ_g) of the amount of resources effectively invested in their education,

[25] It would require some adjustments on the measurement error correction term, but this is indeed a natural extension

given by the difference in the amount of education funds that should be directed to their grade and school and the amount that is diverted $(Y_{m,g,s} - C_{m,g,s})^{26}$. The parameter δ represents the share of learning absorbed from the previous grade and γ_g captures the effect of educational resources on student achievements. Thus, it can be written by the following expression (Ferraz et al., 2012, p. 719):

$$A_{i,s,m,g} = \delta A_{i,s,m,g-1} + \gamma_g (Y_{m,g} - C_{m,g}). \tag{4.14}$$

Under this specification, Ferraz et al. write the expression for educational achievement accomplished by the end of fourth grade, assuming a single measure of corruption for the previous four years as:

$$A_{i,s,m,4} = \delta^4 A_{i,s,m,0} + \sum_{g=1}^{4} \delta^{4-g} \gamma_g Y_{m,g} - \sum_{g=1}^{4-g} C_m. \tag{4.15}$$

Taking the average for students within a school, Ferraz et al. arrive at the equation we wish to estimate empirically:

$$A_{s,m,4} = \alpha + \beta C_m + Z'_m \theta_1 + X'_{s,m} \theta_2 + \epsilon_{s,m}, \tag{4.16}$$

where $A_{s,m,4}$ is a vector of average student achievements for each school s, C_m if a measure of corruption in education, Z_m is a vector of municipal characteristics, including municipality expenditure in primary school, to approximate $\sum_{g=1}^{4} \delta^{4-g} \gamma_g Y_{m,g}$ and $X_{s,m}$ is a vector of students and their families characteristics to account for the initial student achievement $A_{i,s,m,0}$.

Ferraz et al. (2012) assume that $\mathbb{E}[C_m \epsilon_{s,m} | X_m Z_{s,m}] = 0$ and thus "(...) the coefficient β captures the discounted cumulative effects of corruption on student performance since the first grade." (Ferraz et al., 2012, p. 719)

4.2.2. Data

To test our framework we relied on three different public data sources, in addition to the set of irregularities from CGU Random Audits Program. For educational information we used surveys and basic education assessment data organized by the Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP[27]), the research agency linked to the Ministry of Education. For demographic and

[26] Ferraz et al. (2012) discusses that the schooling subscript s can be omitted, assuming that education resources and corruption practices affect every schools evenly.

[27] http://portal.inep.gov.br/web/guest/about-inep

institutional municipal indicators we gathered data from the Instituto Brasileiro de Geografia e Estatística (IBGE) and for data on municipal spending towards primary education, we accessed the Information System on Public Budget in Education (SIOPE), hosted by the National Fund for Education Development (FNDE). In this section we describe each data source in detail. A complete list of variables used in our models and summary statistics tables are available on Appendix B. Each dataset built for this study, as well as the scripts in Python for coding the variables, are available on a public repository[28].

Educational outcomes and school characteristics

To replicate the study proposed by Ferraz et al. (2012), we collected data from two distinct surveys and assessments of Brazilian primary education conducted by INEP: Prova Brasil/SAEB and the School Census. We gathered educational indicators for each penultimate mandate year, since we have information on irregularities found in three political terms (2005-2008, 2009-2012, 2013-2016).

The National Basic Education Assessment System (SAEB) is conducted by INEP since 1995 and is administered every two years. It has two kinds of educational assessment: ANEB, which is sample-based and Prova Brasil (since 2005), a population-based assessment of public schools. It consists of standardized tests covering Portuguese (focusing on reading questions) and Mathematics (focusing on problem solving) and questionnaires asking students, teachers and school principals about their demographic, socioeconomic and studying/working conditions. Students of every public school in the final year of each educational level (primary, secondary and high school) participate on the assessments. Following Ferraz et al. (2012), from this source, we used test scores, students socioeconomic indicators and answers from teachers and principals towards what is a major concern in their school.

We used School Census data, a survey answered by every school in Brazil through an online system. It compiles information about achievement rates, teachers and infrastructure in each school, mainly to assist as tool for guiding public policy in education. We collected information on approval and dropout rates as well as about existence of computer and science labs and sanitation.

Variables from these sources were aggregated at the school level and cover results for pupils in their 5th year of primary education in municipal public schools located at the audited municipalities.

[28] https://github.com/lauragualda/CorruptionProject

Municipal characteristics

We used three different population surveys from IBGE. We took the proportion of urban population and the Gini index for each municipality from the 2010 Census. The GDP per capita estimates for each municipality are presented by IBGE in the annual publication *Gross Domestic Product of Municipalities*. Information about each mayor and institutional characteristics (eg. existence of Education council and if parents actively participate on it) were taken from the Survey of Basic Municipal Information (Munic). Since it covers different topics each year, we used variables from different years inside each term (in the period 2005-2015). We accessed the Information System on Public Budget in Education (SIOPE) to obtain the municipal spending by pupil in primary education for the period 2008-2016. In this system, each municipal and state government is responsible for providing trustworthy information on their spending of resources in education. Since no raw data in a friendly format was available online, we developed a web crawler to extract all the information from *html* pages of each municipality. The crawler script and the consolidated dataset, including information not used in our study are also available in the public repository. Monetary variables of GDP per capita and municipal spending were deflated using the Extended National Consumer Price Index (IPCA).

5. Results

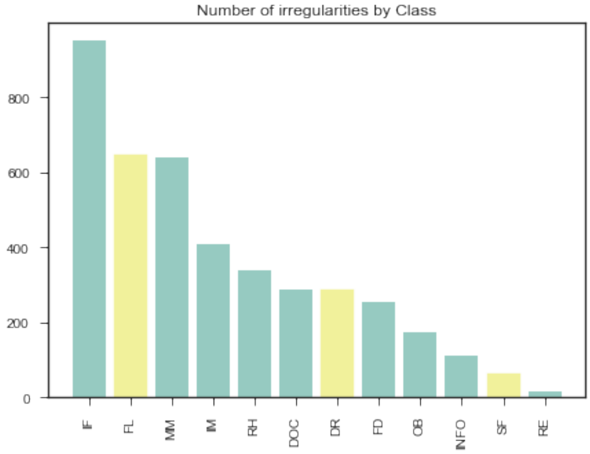
5.1. Classifying Irregularities from Audit Reports

We present results for the automate classification step proposed in our framework in two parts: manual classification and supervised learning, comparing Multiclass and Binary classification tasks.

5.1.1. Manual classification

The manual classification step consisted in reading a sample of irregularities and assigning them to one of twelve pre-defined classes, following Section 2 and Appendix A. We defined a representative sample towards edition, to control for possible variation in the text use and public administration areas audited: 200 irregularities were randomly selected from each of the 21 editions covered in the dataset. On total, we classified 4,166 irregularities [29] according to the distribution in Figure 2. Classes FL, DR and SF (in red) aggregate irregularities related to corruption, following the literature(Ferraz and Finan, 2008). Figure 2 shows the frequency of irregularities assigned to each class. We can see that class *IF* - that groups irregularities related to poorly evaluated supply services and facilities - accounts for approximately 25% of the classified sample. From the classes associated to corruption, procurement related flaws (*FL*) is the most frequent.

Figure 2. Number of irregularities manually assigned to each class



Before proceeding to the supervised learning task, interesting aspects can be highlighted connecting the text classification to other variables in the Irregularities data. About 80% of the irregularities in our sample are *Moderate*, but few classes

[29] 34 had no description, hence had to be discarded.

concentrate most of the 604 *Severe* irregularities. For instance, 47.3% of them were assigned to classes FL, DR or SF. Classes FD and OB also have high shares: 11.1% and 10.1%, respectively, and all the others account for less than 10%. Looking at each class separately, 47% of the irregularities classified as SF and 45% of those assigned to DR are *Severe*. Looking at distribution of irregularities across ministries, the only class with a unexpected behavior is FD: 95% of the classes are associated to the Ministry of Social Development and Fight against Hunger; whereas in all others ministries of Education and Health are related to more than 50% of the irregularities.

Perhaps the most relevant characteristic to analyze after the manual classification is the distribution of words in each class. It is expected that they are aligned to the classes prior delimitation. On Table 2, we have the three most frequent tokens and their relative frequency in each class. From it, we might expect that most of these tokens, when used as features on the supervised learning task, will be useful for splitting the classes. For example, the word *licitatório* rises in 232 irregularities, from which 223 are classified as FL (therefore $\mathbb{P}$(licitarório|FL)=0.96) and *superfaturamento* only occurs in SF class (16 times in the 62 irregularities).

Table 2. Three most frequent (stemmized) tokens in each class

Class	**frequent tokens** (relative frequency)		
FL	licitatório (0.35)	processo (0.3)	licitação (0.23)
SF	valor (0.35)	sobrepreço (0.32)	preço (0.31)
DR	**recurso** (0.47)	**despesa** (0.45)	pagamento (0.29)
FD	**programa** (0.62)	beneficiário (0.56)	família (0.53)
IM	conselho (0.49)	municip (0.38)	social (0.24)
MM	**recurso** (0.42)	**ausência** (0.26)	notificação (0.15)
IF	control (0.16)	escolar (0.14)	**programa** (0.12)
OB	obra (0.63)	execução (0.32)	objeto (0.16)
RH	saúd (0.32)	profissional (0.25)	equip (0.21)
DOC	**ausência** (0.28)	**despesa** (0.25)	documento (0.24)
INFO	divergência (0.33)	censo (0.27)	beneficiário (0.25)
RE	contrapartida (0.81)	estadu (0.56)	**ausência** (0.5)

On the other hand, *recurso*, *programa* and *ausência* are the three most frequent tokens among all irregularities[30], and their high presence in irregularities of

[30] See Appendix A, Words and Bigrams Frequency

two or more classes might not contribute much to the supervised learning task. For instance, the word *recurso* appears in more than a half of FD irregularities, but it only accounts for 22% of the total appearance of this word in our sample.

5.1.2. Automated classification

For the supervised learning step, we observed that classes SF and RE, although well delimited in vocabulary, were too small to be significantly representative. Therefore, we chose to merge SF into class DR and RE into class INFO, ending up with 10 classes in a Multiclass classification problem. In addition to that, since our primary goal is to differentiate irregularities associated to corruption to those best defined as mismanagement, we also performed a binary classification task, where irregularities marked as FL, DR or SF receive 1 and the others receive 0. Next, we present results for these two tasks.

Both tasks follow steps proposed in Section 4.1. Irregularities short description were transformed in features through a pre-processing step, where stop-words[31] from Portuguese vocabulary were removed, some words were reduced to their root form and features were constructed taking 1 and 2-grams. Besides text features, we also tested the inclusion of two binary features generated from information contained in the dataset: to indicate the related ministry and to indicate the severity level of each irregularity. The classified sample was randomly split into a training set with 2,916 irregularities and a test set with 1,250. We take a stratified sample regarding ministry and severity level, to ensure that we have fair representation in every fold during cross-validation.

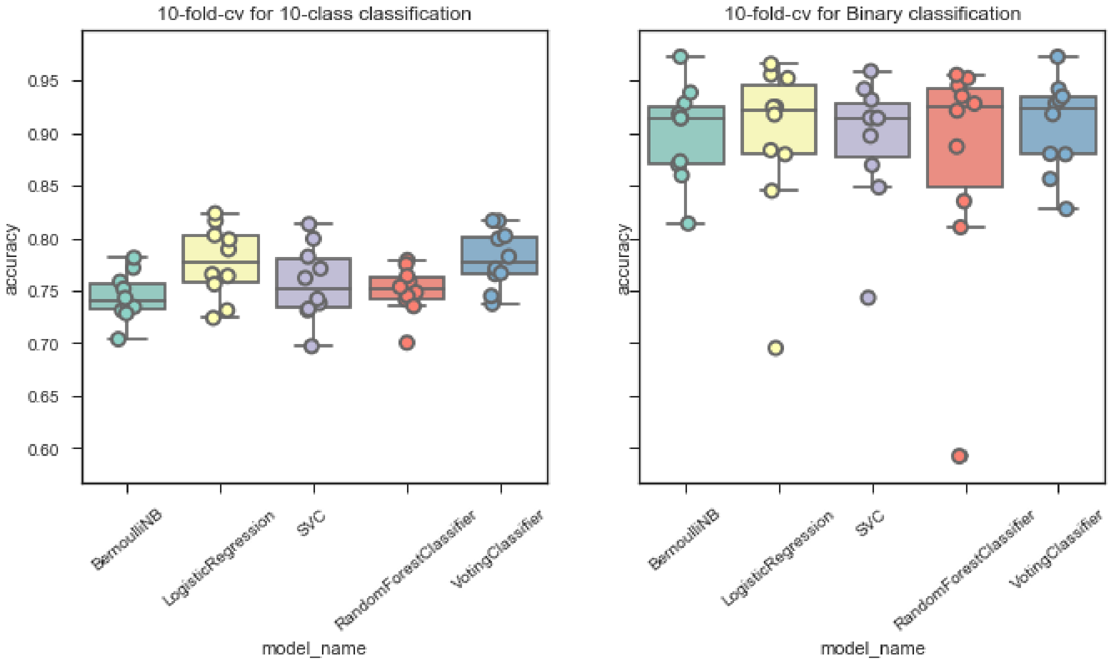
For each of the four machine learning algorithms - Naive Bayes, Multinomial Logistic Regression, SVM and Random Forest - we used 10-fold cross-validation for tuning hyper-parameters. After selecting a reasonable setting we compared their performance through 10-fold cross-validation on the training set and respective accuracy values are reported in Figure 3. For both tasks we also used a Voting ensemble method that chooses the class that maximizes the sum of predicted probabilities for each single classifier.

We can see that Multinomial Logistic Regression outperforms the other three classifiers on the Multiclass problem. On the binary classification task, all classifiers have similar performance and can correctly predict classes of about 92% of the irregularities. Tables 20 and 21 on Appendix C contain final hyper-parameters

[31] We used NLTK module for Portuguese corpus and added 'n°', 'N.°', 'r', specific to our set of documents.

setting for each classifier, and the average accuracy on the train and test set obtained using them.

Figure 3. 10-fold cv for Supervised Learning classification tasks



To predict classes for the entire set of irregularities, we chose the Voting classifier. In addition to good performance, we observed an advantage on combining the four classifiers: while the Naive Bayes tends to be bolder and incurs in more False Positive predictions, the other three classifiers tend to be more conservative and have a higher False Negative rate. Therefore, their combination resulted in a satisfactory equilibrium of precision and recall Tables 20 and 21 contain precision and recall values for the Voting model on the test set. On Appendix C we present confusion matrices, useful for the next step.

Table 3. Classification report on the Test set - 2 classes

class	**precision**	**recall**	**f1-score**	**class size**
FL+DR+SF	0.83	0.85	0.84	298
other	0.95	0.95	0.95	953
avg/total	0.92	0.92	0.92	1251

Table 4. Classification report on the Test set - 10 classes

class	**precision**	**recall**	**f1-score**	**class size**
FL	0.87	0.84	0.85	194
DR+SF	0.66	0.72	0.69	104
FD	0.89	0.88	0.89	76
IM	0.91	0.86	0.88	122
MM	0.76	0.69	0.72	192
IF	0.81	0.87	0.84	285
OB	0.67	0.83	0.74	52
RH	0.83	0.79	0.81	101
DOC	0.76	0.76	0.76	86
INFO+RE	0.82	0.69	0.75	39
avg/total	0.81	0.8	0.8	1251

5.2. Estimating effects of Corruption in Schooling outcomes

We describe results for the second step of our framework in three parts. We begin by describing how we constructed a municipal indicator variable of corruption and estimate necessary parameters for using it in inferential models. Next, we present the specification of parameters used for measurement error adjustment. Then we estimate and compare coefficients obtained using the corruption variable on the model proposed in Equation (4.16). As educational outcomes we use standardized scores on Math and Portuguese tests, Failure rates and Dropout rates, as proposed by Ferraz et al. (2012).

5.2.1. Constructing a Municipal Indicator of Corruption

The second step of our framework begins with the construction of a municipal indicator of corruption in Education, following the methodology proposed in Section 4. In our study, a municipality is classified as having corruption in Education if at least one irregularity found during the audit fits into three criteria:

1. It is associated to a OS expedited by the Ministry of Education,
2. It is predicted as FL, SF or DR on the automated classification task,
3. It is classified as *Severe* by CGU.

According to this indicator, 40.45% of the 1,223 Brazilian municipalities randomly selected for inspection had at least one anomaly suggesting public corruption in Education. (Ferraz et al., 2012, p. 717) find 35% on their sample 365 municipalities. Figure 4 shows the distribution of Math and Portuguese scores split by our indicator of corruption and year. Despite the noise introduced by the measurement error, a difference is visible: median scores in municipalities not classified as corrupt are higher than third percentile scores, in both tests and in the three years. This difference is less striking when looking at Failure and Dropout rates, as shown in Figure 5.

Figure 4. Prova Brasil scores in Mathematics and Portuguese, split by term and incidence of corruption in Education

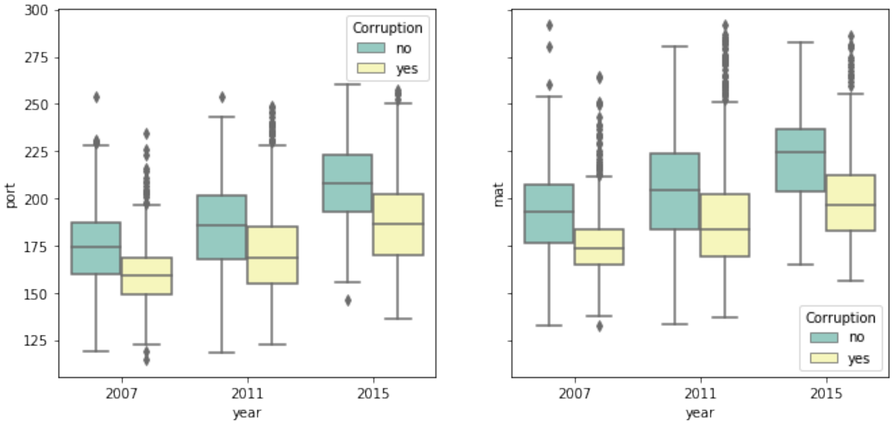
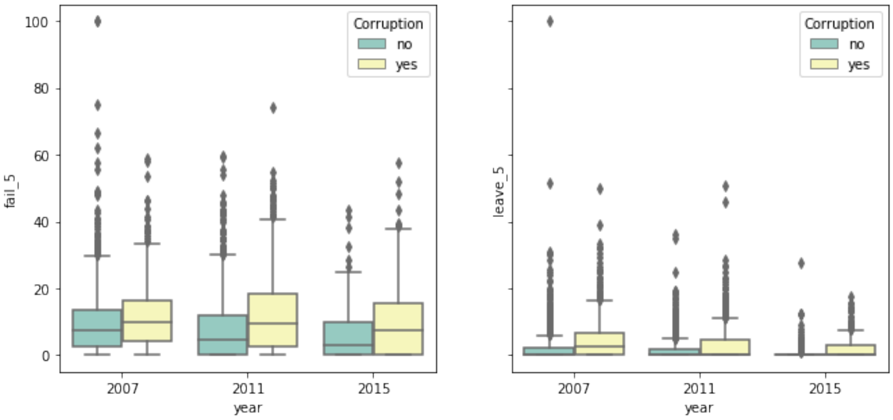


Figure 5. Failure and Dropout rates (%), split by term and incidence of corruption in Education



5.2.2. Model estimation adjusting for Measurement Error

To be able to correctly use this variable in an inferential model, we need estimates for its misclassification rate, r_{01} and r_{10} and extra-sample information on the true occurrence of corruption in Brazilian municipalities, P. We took the "Proportion of municipalities with corruption in education" value of 35% found by Ferraz et al. to account for P. To estimate the misclassification rates, we take the two possible approaches proposed in Section 4: *Negative Binomial*, (1), and *Classification task*, (2).

To use the Negative Binomial approach for estimating misclassification rates, we first need to test the adherence of the *Number of irregularities in Education*[32], estimating α and p from maximum likelihood on the data. We do not reject the hypothesis that the random variable is generated according to this distribution, therefore we can use the estimated $\hat{\alpha}$ and $\hat{p}$ to calculate the estimators for $\hat{r}_{01}$ and $\hat{r}_{10}$, following Equation (4.6).

For estimating $\hat{r}_{01}$ and $\hat{r}_{10}$ directly from the text classification task, we followed the cross-validation process described in Section 4. The confusion matrix of the test set aggregated by municipality is on Appendix D.

The term of adjustment (4.13) can be calculated using these estimates and the inverse covariance matrix term (4.12). The latter is estimated through data covariance matrix. On Table 5, we present the estimates for error rates, for the term on the inverse-covariance matrix (term inv-cov.) and the calculated term of adjustment.

Table 5. Parameters specification to adjust for Measurement error

Parameter	**Negative Binomial** (1)	**Classification task** (2)
P	0.35	0.35
r_0	0.055	0.078
r_1	0.186	0.104
term inv-cov.	5.35	5.35
term of adjustment	1.67	1.38

[32] See Adherence Test for Number of Irregularities in Education in Appendix D

5.2.3. Estimating effects of Corruption in Schooling outcomes

The last part of our framework consists in obtaining an asymptotically efficient estimator for the effect of corruption on different schooling outcomes. We estimate OLS regression models using four dependent variables: standardized Math scores, standardized Portuguese scores, Failure rates and Dropout rates. Following Ferraz et al. (2012), in addition to the mismeasured variable of corruption, we include school and municipal characteristics for control. Precisely, school characteristics include principals' survey answers, dummy variables to account for existence of computer lab, science lab and sanitation and students rates based on dummies of gender, color, age, parents' qualification, living conditions and computer at home [33]. Municipal characteristics are Gini, proportion of urban population, *log* of total population, *log* of GDP per-capita and *log* of expenditure in primary school per child at prices of 2007. Since we aggregate data for three consecutive political terms, we include dummy variables for term (2007, 2011 or 2015), for mayor mandate (first or second) and for identifying if the municipality had already been audited before (Avis et al., 2016; Ferraz and Finan, 2011).

Tables 6 and 7 report OLS and MLS estimates (considering strategies (1) and (2) for estimating misclassification rates) for the effect of corruption in municipal schools' average performance on standardized Math and Portuguese tests, taken by students in 5th grade. We find a statistically significant signal that endorses Ferraz et al. (2012) results: schools in municipalities where corruption in Education was detected have worse performance on Math and Portuguese tests, when compared to schools in municipalities where irregularities associated to corruption in Education were not found.

Table 6. Effects of Corruption on Mathematics Standardized Test Scores

	OLS Estimation	**MLS Estimation**	
		(1)	(2)
Corruption in Education	-0.1098*** (0.025)	-0.183*** (0.042)	-0.152*** (0.035)
School and Municipal characteristics	Yes	Yes	Yes

Dependent Variable: Standardized Math Test scores among 5th year students.
Robust SE. Adjusted R-squared for OLS Estimation: 0.497. $N = 4670$

[33] We chose not use teacher's survey answers because the questions of interest were not asked on the 2015 survey. Besides that, Ferraz et al. (2012) also includes rates based on dummies of electrical power and sanitation on students' houses, but this question is no longer in their survey.

Table 7. Effects of Corruption on Portuguese Standardized Test Scores

	OLS Estimation	**MLS Estimation**	
		(1)	(2)
Corruption in Education	-0.111*** (0.024)	-0.185*** (0.04)	-0.153*** (0.033)
School and Municipal characteristics	Yes	Yes	Yes

Dependent Variable: Standardized Portuguese Test scores among 5th year students.
Robust SE. Adjusted R-squared for OLS Estimation: 0.534. $N = 4670$

Ferraz et al. find that "corruption in education is associated with a significant decrease of 0.35 standard deviations in test scores"(p.722) when controlling only for school characteristics. When municipal characteristics are includes, the decrease is about 0.29 for Math scores and 0.277 for Portuguese. We found lower coefficients using both methods for estimating misclassification rates: a decrease of 0.18 standard deviations in both test scores is expected considering approach (1) and of 0.15 according to approach (2). We believe that difference in time spanned and sample size in our estimations account for part of this difference. While Ferraz et al. (2012) use data for 1,488 schools from 366 municipalities audited from 2003 to 2005, our sample has test scores for 4,670 schools from 1,001 municipalities audited during three consecutive political terms (from 2006 to 2015).

Tables 8 and 9 present estimates for the effect of corruption in educational attainment rates among students from 5th grade in municipal schools. We found that dropout rates are less than 0.5 percentage point higher in municipalities where corruption in Education was detected. Although shy, this effect is significant. Failure rates are approximately 1 percentage point higher, but this effect is less significant, also endorsing Ferraz et al. (2012) findings.

Table 8. Effects of Corruption on Failure Rates

	OLS Estimation	**MLS Estimation**	
Explanatory Variables		(1)	(2)
Corruption in Education	0.006* (0.003)	0.009** (0.005)	0.008** (0.004)
School and Municipal characteristics	Yes	Yes	Yes

Dependent Variable: Failure rates among 5th year students.
Robust SE. Adjusted R-squared for OLS Estimation: 0.175. $N = 4657$

Table 9. Effects of Corruption on Dropout Rates

	OLS Estimation	**MLS Estimation**	
		(1)	(2)
Corruption in Education	0.003** (0.001)	0.005*** (0.002)	0.004*** (0.001)
School and Municipal characteristics	Yes	Yes	Yes

Dependent Variable: Dropout rates among 5th year students.
Robust SE. Adjusted R-squared for OLS Estimation: 0.209. $N = 4657$

We obtained results consistent with Ferraz et al. (2012) when using our indicator variable of corruption and following a MLS approach to adjust for measurement error. Some robustness checks are necessary before drawing conclusions about causal effects of corruption in schooling outcomes and this is a topic well discussed by Ferraz et al.. For now, our goal was simply to demonstrate an application of our framework and provide evidence that it achieves consistent results.

6. Conclusion

Corruption is one of the most discussed issues by Brazilian population and an eminent research topic among economists, political scientists, social scientists and recently, even between mathematicians. Quantitative studies in this area are important to help disentangle the effects of public corruption on the supply and quality of basic services. Understanding these consequences is essential to guide effective public policies. A general complaint in the literature (Fisman and Golden, 2017) is the lack of objective measures of corruption to serve as input for consistent studies. The CGU Random Audits Anti-Corruption Program has contributed to this cause by providing detailed reports revealing corruption and mismanagement irregularities detected during unadvertised inspections.

In this study, we proposed a framework guide the efficient use of these reports as source of information for quantitative analyses. Our methodology assembles two distinct areas of statistics: machine learning for text classification and inferential modelling. We use a supervised learning classifier to predict classes for a set of irregularities detected during audits and sequentially use them to construct a municipal indicator of corruption. On the binary classification we obtained an expected accuracy of 92% and were able to predict most of the irregularities related to corruption. To use this indicator as explanatory variable in a regression model, we proposed an approach using Modified Least Squares to account for measurement error using misclassification rates from the first step.

We presented a first application for this framework based on the empirical approach proposed in Ferraz et al. (2012) to estimate the relationship between public corruption in educational funds and schooling outcomes. We used our indicator of corruption in education and data for in 4,670 schools across 1,001 municipalities audited from 2006 to 2015 to estimate if standardized test scores and achievement rates vary in municipalities where corruption was detected. We cover a different period and a different set of audited municipalities, but our results endorse Ferraz et al. findings: schooling outcomes are significantly worse in municipalities where corruption in education was revealed. For instance, our results suggest that students' standardized scores on Math and Portuguese national tests (Prova Brasil), were on average 0.15 standard deviations lower in corrupt municipalities.

Overall, we achieved satisfactory results using the proposed framework, both on the machine learning classification step and on the regression model estimation, and we hope that this contribution is able to motivate the use of text data in inferential models. However, we understand that it has some limitations that must be addressed in future works. The framework was initially designed to construct and deal with measurement error in a binary indicator, but a natural extension is to adapt it for other variables, such as proportion and mean. We proposed a particular use case based on audit reports, but its use can be extended to any source of text data, accounting for differences in structure and vocabulary.

References

Abu-Mostafa, Y. S., Magdon-Ismail, M., and Lin, H.-T. (2012). *Learning From Data*. AMLBook.

Aigner, D. J. (1973). Regression with a binary independent variable subject to errors of observation. *Journal of Econometrics*, 1(1):49–59.

Avis, E., Ferraz, C., and Finan, F. (2016). Do government audits reduce corruption? estimating the impacts of exposing corrupt politicians. Working Paper 22443, National Bureau of Economic Research.

Bierens HJ (2014). The inverse of a partitioned matrix.

Bollinger, C. R. (1996). Bounding mean regressions when a binary regressor is mismeasured. *Journal of Econometrics*, 73(2):387 – 399.

Bologna, J. and Ross, A. (2015). Corruption and entrepreneurship: evidence from brazilian municipalities. *Public Choice*, 165(1):59–77.

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

Brollo, F. and Troiano, U. (2016). What happens when a woman wins an election? evidence from close races in brazil. *Journal of Development Economics*, 122(C):28–45.

Caldas, O., Costa, C., and Pagliarussi, M. (2016). Corrupção e composição dos gastos governamentais: evidências a partr do programa de fiscaização por sorteios públicos da controladoria-geral da união. *Revista Administração Pública*, 50(2):237–264.

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

Ferraz, C. and Finan, F. (2011). Electoral accountability and corruption: Evidence from the audits of local governments. *American Economic Review*, 101(4):1274–1311.

Ferraz, C., Finan, F., and Moreira, D. B. (2012). Corrupting learning: evidence from missing federal education funds in brazil. *Journal of Public Economics*, 96(Issues 9–10):712–726.

Fisman, R. and Golden, M. A. (2017). *Corruption: What Everyone Needs to Know*. Number 9780190463977 in OUP Catalogue. Oxford University Press.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

Instituto Brasileiro de Geografia e Estatística (2017). Population estimates for brazilian municipalities.

International Transparency (2017). Corruption perception index.

Joachims, T. (1998). Text categorization with suport vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning*, ECML '98, pages 137–142, London, UK, UK. Springer-Verlag.

Johnston, J. (1997). *Econometric Methods.* McGraw Hill Higher Education, 4th edition.

Klepper, S. (1988). Regressor diagnostics for the classical errors-in-variables model. *Journal of Econometrics*, 37(2):225 – 250.

Lichand, G., Lopes, M., and Medeiros, M. (2016). Is corruption good for your health? (job market paper). Working paper.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval.* Cambridge University Press, online edition.

Mondo, B. V. (2016). Measuring political corruption from audit results: A new panel of brazilian municipalities. *European Journal on Criminal Policy and Research*, 22:477–498.

Méon, P.-G. and Sekkat, K. (2005). Does corruption grease or sand the wheels of growth? *Public Choice*, 122(1/2):69–97.

Savoca, E. (2000). Measurement errors in binary regressors: An application to measuring the effects of specific psychiatric diseases on earnings. *Health Services and Outcomes Research Methodology*, pages 149–164.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory.* Springer-Verlag New York, Inc., New York, NY, USA.

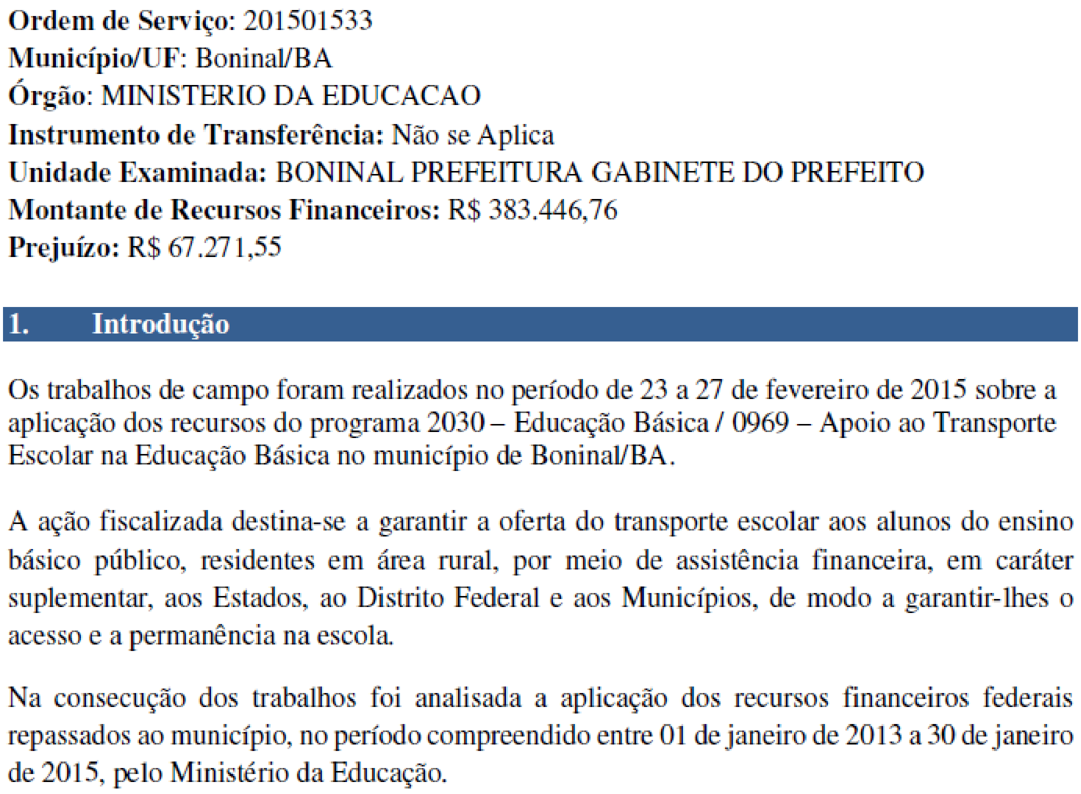
Vijayarani, S., Ilamathi, M. J., and Nithya, M. (2015). Preprocessing techniques for text mining - an overview. *International Journal of Computer Science & Communication Networks*, 5(1):7–16.

World Bank (2017). Cpia transparency, accountability and corruption in the public sector rating.

Appendices

A. Irregularities data

Figure 6. Example of Service Order in Education executed in Boninal(BA) during an audit on the 40th edition for the CGU Anti-Corruption Program



Ordem de Serviço: 201501533
Município/UF: Boninal/BA
Órgão: MINISTERIO DA EDUCACAO
Instrumento de Transferência: Não se Aplica
Unidade Examinada: BONINAL PREFEITURA GABINETE DO PREFEITO
Montante de Recursos Financeiros: R$ 383.446,76
Prejuízo: R$ 67.271,55

1. Introdução

Os trabalhos de campo foram realizados no período de 23 a 27 de fevereiro de 2015 sobre a aplicação dos recursos do programa 2030 – Educação Básica / 0969 – Apoio ao Transporte Escolar na Educação Básica no município de Boninal/BA.

A ação fiscalizada destina-se a garantir a oferta do transporte escolar aos alunos do ensino básico público, residentes em área rural, por meio de assistência financeira, em caráter suplementar, aos Estados, ao Distrito Federal e aos Municípios, de modo a garantir-lhes o acesso e a permanência na escola.

Na consecução dos trabalhos foi analisada a aplicação dos recursos financeiros federais repassados ao município, no período compreendido entre 01 de janeiro de 2013 a 30 de janeiro de 2015, pelo Ministério da Educação.

Five irregularities are associated to this Service Order (in Portuguese)[34]:

- "Falta de comprovação documental das despesas realizadas em 2013 e 2014."
- "O Conselho do Fundeb não atua no acompanhamento da execução do Programa Nacional de Transporte Escolar - Pnate."
- "Exigências indevidas com repercussão na competitividade do certame e direcionamento em licitação do transporte escolar (Pregão n.º 012/2013)"
- "Intermediação irregular na prestação do serviço de transporte escolar."
- "Pagamento por serviços não realizados, no montante de R$13.577,27, referente à despesa do transporte escolar do mês de junho de 2013."

[34] For instance, these would be classified as FL, DR, DR, IM and IF, respectively.

Classes description

Table 10. Classification of irregularities - part 1

Class	**Example** (in Portuguese)
(1/FL) Procurement related flaws eg. Favored vendor, lack of publicity, irregular receipts, evidence for ghost firms, documents set with different dates, no realization, irregular class	"Direcionamento de processos licitatórios e de dispensa, favorecendo parentes do Prefeito Municipal"
(2/SF) Over-invoicing/ Over-princing/ Off-the-record payments	"Superfaturamento de 17,62% na prestação de serviços de retífica de motores realizada em ônibus escolares com recursos do PNATE"
(2/DR) Resource Diversion eg. Unconfirmed payments, Diversion of resources for other goal, for other goals within the same Ministety, for other goals within Program, Under-application of resources	"Despesa de 5.444,50 não comprovada por nota fiscal válida"
(3/FD) Frauds in execution of public programs	"Servidor estadual beneficiário do Programa Bolsa Família com indícios de renda per capita superior à estabelecida na legislação para a permanência no Programa"

Table 11. Classification of irregularities - part 2

Category/Class	**Example** (in Portuguese)
(4/IM) Councils and public secretaries	"O Plano Municipal de Saúde não tem estrutura e conteúdo conforme legislação"
(5/MM) Mismanagement practices directly related to the municipal government eg. irregular composition, irregular operation, poor infrastructure and work conditions	"Prefeitura não notificou os partidos políticos e entidades municipais quanto ao recebimento de recursos relativos ao Contrato de Repasse n^o 180.314–72/2006"
(6/IF) Poorly evaluated supply services and facilities eg. precarious facilities, signs and logos not properly set, lack of medical supplies, lack of books and meals in schools	"Falta de livros nas escolas"
(7/OB) Unfinished projects and unaccomplished goals in public construction	"O objetivo do programa não foi atendido, uma vez que, a obra de construção da concha acústica do município de Palestina do Pará permanece paralisada"
(8/RH) Human Resources anomalies eg. Professionals that don't fulfill work time requirements, staff training, staff composition, Public servants' payments	"Equipe de Saúde da Família com composição incompleta"
(9/DOC) Incomplete documentation or inadequate account keeping	"Ausência de comprovação de contrapartida municipal, no montante de R$ 38.069,16"
(10/INFO) Lack or divergence on the publicly released and the audited information	"Divergência entre o números de alunos informados pela Prefeitura e pelo Educacenso"
(10/RE) Absence of State's government participation in public programs	"Secretaria de Estado da Saúde não repassa integralmente os valores pactuados para medicamentos básicos"

Irregularities associated to Corruption by Severity level

Severe irregularities assigned to classes related to corruption:

- Caicó_040_63: Superfaturamento, no valor de R$ 16.245,27. - **SF**
- Itambacuri_040_51: Itarantim_038_5: Irregularidades nos processos licitatórios: concorrência fictícia. - **FL**
- São João_029_3: Não disponibilização de documentações comprobatórias de débito efetuados na conta específica do Programa Vigilância Epidemiológica em Saúde. - **DR**

Moderate irregularities assigned to classes related to corruption:

- Dois Irmãos das Missões_037_24: Indicação de superfaturamento na construção de muro de contenção, objeto do primeiro aditivo contratual. Ausência de orçamento prévio e ausência de avaliação técnica do atendimento do projeto por parte da Prefeitura. Construção oferece risco a moradores. - **SF**
- Itambacuri_040_51: Restrição ao caráter competitivo na realização de licitação para construção da Escola Municipal Irmã Germana. - **FL**
- Itaparica_03_80: Falta de comprovação documental de despesas realizadas no montante de R$ 297.006,65 com recursos do FUNDEB em 2008. - **DR**

Administrative irregularities assigned to classes related to corruption:

- Nova Olímpia_039_62: Pagamento indevido decorrente de superfaturamento por serviços pagos e não executados que pode ultrapassar o valor de R$ 3.783,12 (três mil e setecentos e oitenta e três reais e doze centavos). - **SF**
- Virgem da Lapa_020_15: Ausência de processo licitatório na aquisição de gêneros alimentícios. - **FL**
- Bela Vista da Caroba_036_12: Ausência da documentação de suporte à movimentação financeira da conta do programa. - **DR**

Frequent Words and Bigrams

Figure 7. 40 most frequent words in all irregularities

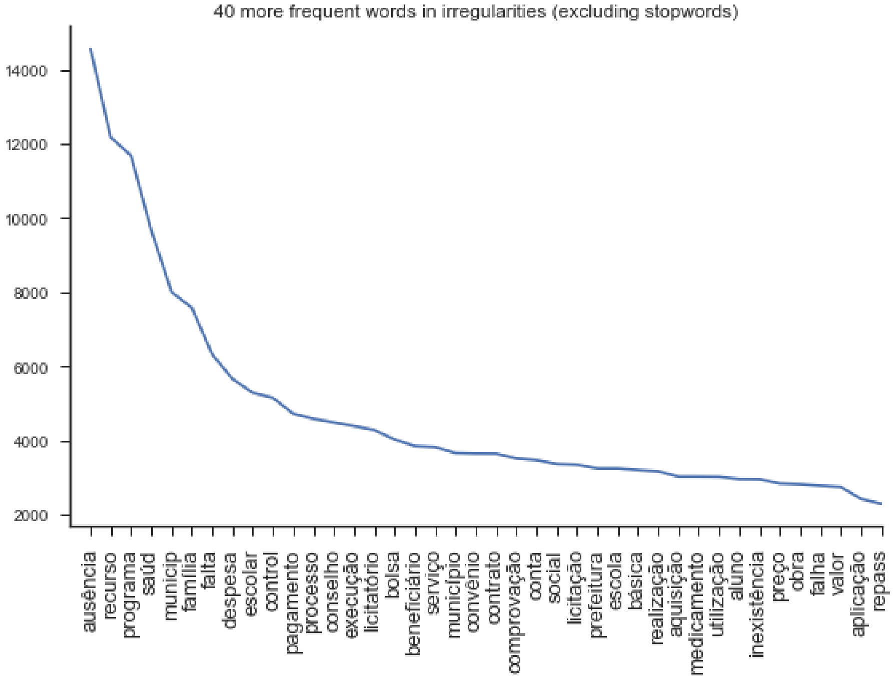
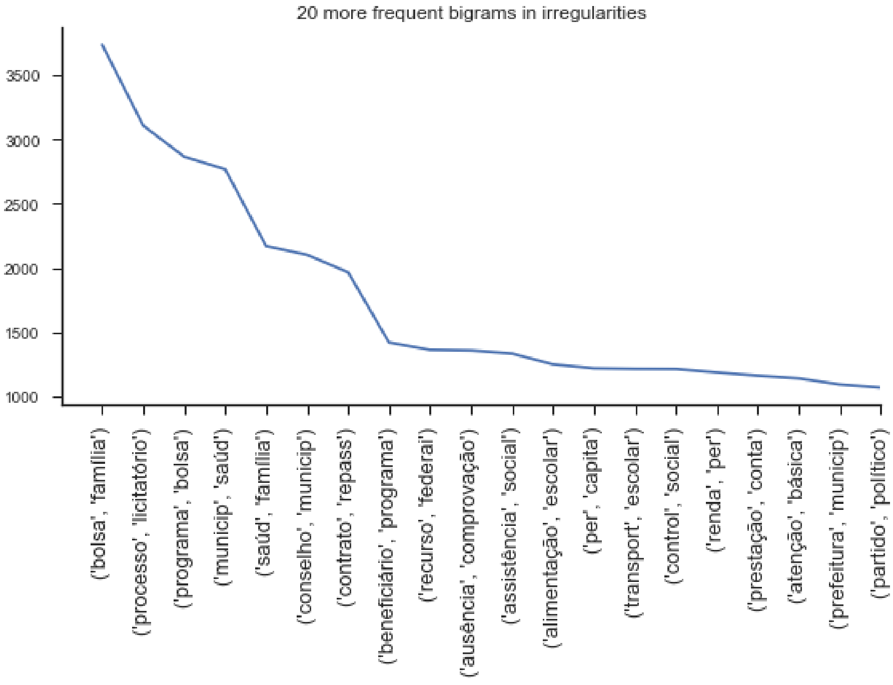


Figure 8. 20 most frequent bigrams in all irregularities



B. Case Study data

Table 12. School characteristics

Variable	Source	Period
Portuguese Test Scores - 5th year	INEP (Prova Brasil/SAEB)	2007, 2011, 2015
Mathematics Test Scores - 5th year	INEP (Prova Brasil/SAEB) (Prova Brasil/SAEB)	2007, 2011, 2015
Number of test participants	INEP (Prova Brasil/SAEB)	2007, 2011, 2015
Approval rate - 5th year (municipal schools)	INEP (School Census)	2007, 2011, 2015
Approval rate - primary education (municipal schools)	INEP (School Census)	2007, 2011, 2015
Failure rate - 5th year (municipal schools)	INEP (School Census)	2007, 2011, 2015
Failure rate - primary education (municipal schools)	INEP (School Census)	2007, 2011, 2015
Dropout rate - 5th year (municipal schools)	INEP (School Census)	2007, 2011, 2015
Dropout rate - primary education (municipal schools)	INEP (School Census)	2007, 2011, 2015
School has a computer lab	INEP (School Census)	2007, 2011, 2015
School has a science lab	INEP (School Census)	2007, 2011, 2015
School has sanitation	INEP (School Census)	2007, 2011, 2015
Number of municipal schools	INEP (School Census)	2007, 2011, 2015

Table 13. Students characteristics

Variable	**Source**	**Period**
% male students	INEP (Student Survey, Prova Brasil/SAEB)	2007, 2011, 2015
% white students	INEP (Student Survey, Prova Brasil/SAEB)	2007, 2011, 2015
% of students who are 8 years old or younger	INEP (Student Survey, Prova Brasil/SAEB)	2007, 2011, 2015
% of students who are 9 years old	INEP (Student Survey, Prova Brasil/SAEB)	2007, 2011, 2015
% of students who are 10 years old	INEP (Student Survey, Prova Brasil/SAEB)	2007, 2011, 2015
% of students who are 11 years old	INEP (Student Survey, Prova Brasil/SAEB)	2007, 2011, 2015
% of students who are 12 years old	INEP (Student Survey, Prova Brasil/SAEB)	2007, 2011, 2015
% students with a home computer	INEP (Student Survey, Prova Brasil/SAEB)	2007, 2011, 2015
% of students living with both parents	INEP (Student Survey, Prova Brasil/SAEB)	2007, 2011, 2015
% students living with 6 or more people	INEP (Student Survey, Prova Brasil/SAEB)	2007, 2011, 2015
% of mothers with a high school degree	INEP (Student Survey, Prova Brasil/SAEB)	2007, 2011, 2015
% of fathers with a high school degree	INEP (Student Survey, Prova Brasil/SAEB)	2007, 2011, 2015

Table 14. Principals and Teachers characteristics

Variable	**Source**	**Period**
% of schools where elections are held for principal	INEP (Student Survey, Prova Brasil/SAEB)	2007, 2011, 2015
% of schools with training courses provided to teachers	INEP (Student Survey, Prova Brasil/SAEB)	2007, 2011, 2015
School principal considers lack of financial resources as serious concern	INEP (Student Survey, Prova Brasil/SAEB)	2007, 2011, 2015
School principal considers lack of schooling supplies as serious concern	INEP (Student Survey, Prova Brasil/SAEB)	2007, 2011, 2015
School principal considers lack of teachers as serious concern	INEP (Student Survey, Prova Brasil/SAEB)	2007, 2011, 2015
School principal considers disciplinary problems as serious concern	INEP (Student Survey, Prova Brasil/SAEB)	2007, 2011, 2015
% of teachers with College degree	INEP (Student Survey, Prova Brasil/SAEB)	2007, 2011, 2015
% of teachers considering lack of financial resources as serious concern	INEP (Student Survey, Prova Brasil/SAEB)	2007, 2011, 2015
% of teachers considering lack of schooling supplies as serious concern	INEP (Student Survey, Prova Brasil/SAEB)	2007, 2011, 2015
% of teachers considering lack of teachers as serious concern	INEP (Student Survey, Prova Brasil/SAEB)	2007, 2011, 2015
% of teachers considering disciplinary problems as serious concern	INEP (Student Survey, Prova Brasil/SAEB)	2007, 2011, 2015

Table 15. Municipal characteristics

Variable	**Source**	**Period**
% urban population	IBGE (Brazilian Census)	2010
Gini	IBGE (Brazilian Census)	2010
GDP per capita - prices of 2007 (IPCA)	IBGE (PIB dos Municípios)	2007, 2011, 2015
Expenditure in primary school per child - prices of 2007 (IPCA)	SIOPE (Relatório de Indicadores	2008, 2011, 2015
Mayor is a male	IBGE (Munic)	2005, 2009, 2013
Mayor on second term	IBGE (Munic)	2005, 2009, 2013
Mayor has college degree	IBGE (Munic)	2005, 2009, 2013
Mayor Party	IBGE (Munic)	2005, 2009, 2013
Community helps in the maintenance of municipal schools	IBGE (Munic)	2006, 2009, 2014
Municipality had school enrollment campaign	IBGE (Munic)	2006, 2009, 2014
Municipality uses participatory budgeting	IBGE (Munic)	2006, 2009, 2014
Municipality has an education council	IBGE (Munic)	2006, 2009, 2014
Parents are active in the education council	IBGE (Munic)	2006, 2014
Municipality has a intergovernmental consortium	IBGE (Munic)	2005, 2009, 2015
Schools receive support from private sector	IBGE (Munic)	2005, 2009

Table 16. Summary Statistics: Schools and Students characteristics

Variable	**N**	**Mean**	**SD**	**Min**	**Max**
Portuguese Test Scores - 5th year	4670	176.82	23.61	114.38	260.64
Mathematics Test Scores - 5th year	4670	194.30	26.40	132.30	291.68
Standardized Portuguese Test Scores	4670	-0.22	1.04	-3.10	4.50
Standardized Mathematics Test Scores	4670	-0.19	1.06	-2.80	5.13
Number of test takers	4670	52.62	34.22	5.00	318.00
Approval rates - 5th year	4657	0.88	0.12	0.00	1.00
Failure rates - 5th year	4657	0.10	0.10	0.00	1.00
Dropout rates - 5th year	4657	0.02	0.05	0.00	1.00
% male students	4670	0.51	0.10	0.00	1.00
% white students	4670	0.28	0.16	0.00	1.00
% of students who are 8 years old or younger	4670	0.00	0.01	0.00	0.28
% of students who are 9 years old	4670	0.04	0.06	0.00	0.48
% of students who are 10 years old	4670	0.38	0.18	0.00	0.96
% of students who are 11 years old	4670	0.31	0.13	0.00	1.00
% of students who are 12 years old	4670	0.13	0.08	0.00	0.64
% students with a home computer	4670	0.30	0.22	0.00	1.00
% of students living with both parents	4670	0.63	0.13	0.00	1.00
% students living with 6 or more people	4670	0.27	0.14	0.00	1.00
% of mothers with a high school degree	4670	0.17	0.11	0.00	0.80
% of fathers with a high school degree	4670	0.14	0.10	0.00	0.67
% of schools where elections are held for principal	4670	0.21	0.41	0.00	1.00
% of schools with computer lab	4670	0.53	0.50	0.00	1.00
% of schools with science lab	4670	0.03	0.18	0.00	1.00
% of schools with sanitation	4670	0.98	0.14	0.00	1.00

Table 17. Summary Statistics: Principals and Teachers Surveys

Variable	**N**	**Mean**	**SD**	**Min**	**Max**
Principals Survey					
% training courses are provided to teachers	4670	0.50	0.50	0.00	1.00
% lack of financial resources is a serious concern	4670	0.09	0.28	0.00	1.00
% lack of schooling supplies is a serious concern	4670	0.05	0.23	0.00	1.00
% lack of teachers is a serious concern	4670	0.06	0.24	0.00	1.00
% disciplinary problems are a serious concern	4664	0.09	0.28	0.00	1.00
Teachers Survey					
% with College degree	4670	0.84	0.29	0.00	1.00
% lack of financial resources is a serious concern	4059	0.07	0.19	0.00	1.00
% lack of schooling supplies is a serious concern	4059	0.08	0.21	0.00	1.00
% lack of teachers is a serious concern	4059	0.07	0.19	0.00	1.00
% disciplinary problems are a serious concern	4059	0.10	0.23	0.00	1.00

Table 18. Summary Statistics: Municipal characteristics

Variable	**N**	**Mean**	**SD**	**Min**	**Max**
GDP per capita (R$, 2007)	1036	15,947.53	23,218.92	2,396.24	51,3134.20
Population	1036	25,338	47,325.75	805	571,149
% urban population	1036	0.65	0.21	0.09	1.00
Gini	1036	0.51	0.06	0.33	0.78
Mayor is a male	1036	0.92	0.28	0.00	1.00
Mayor is on 2nd term	1036	0.36	0.48	0.00	1.00
Mayor has College degree	1036	0.46	0.50	0.00	1.00
Community helps in maintenance of schools	1036	0.50	0.50	0.00	1.00
Municipality has School enrollment campaign	1036	0.45	0.50	0.00	1.00
Municipality uses participatory budgeting	1036	0.94	0.24	0.00	1.00
Municipality has Education council	1036	0.73	0.44	0.00	1.00
Parents and Teachers participate in Education council	542	0.64	0.48	0.00	1.00
Municipality has a intergovernmental consortium in Education	1036	0.22	0.41	0.00	1.00
Municipality has a private support in Education	883	0.15	0.36	0.00	1.00
Expenditure in primary school per child (R$,2007)	1036	5,178.70	2,371.17	1,298.59	37,045.79
Number of Municipal Schools	1036	22.58	27.78	1.00	222.00
Municipality has Private School	1036	0.47	0.50	0.00	1.00
Avg. Failure rate in Private Schools - 5th year	492	0.02	0.04	0.00	0.38
Avg. Dropout rate in Private Schools - 5th year	522	0.02	0.06	0.00	1.00

Table 19. Summary Statistics: Municipal characteristics from Irregularities data

Variable	**N**	**Mean**	**SD**	**Min**	**Max**
Number of irregularities	1073	68.41	35.24	5.00	276.00
Number of *Severe* irregularities	1073	10.00	11.95	0.00	92.00
Number of *Moderate* irregularities	1073	53.85	29.72	5.00	249.00
Number of *Administrative* irregularities	1073	4.57	5.41	0.00	42.00
Number of OS in Education	1073	6.06	2.59	1.00	34.00
Number of Irregularities in Education	1073	21.41	16.27	0.00	155.00
Number of *Severe* irregularities in Education	1073	3.27	5.65	0.00	41.00
Number of irregularities associated to corruption	1073	2.18	3.86	0.00	37.00
Municipality has irregularities associated to corruption in Education	1073	0.46	0.50	0.00	1.00
Municipality was audited more than once	1073	0.08	0.27	0.00	1.00

C. Supervised Learning Performance metrics

Table 20. ML classification of irregularities - 10 classes

Classifier	**avg 10-cv accuracy**	**test accuracy**	**hyper-parameters**
Naive Bayes	0.744	0.777	$\alpha = 0.01$

Multinomial Logistic Reg.	0.777	0.794	penalty=lasso
			C=5

SVM	0.757	0.778	kernel=linear
			C=1

Random Forest	0.75	0.781	estimators=1000
			criterion=gini

Voting	0.78	0.803	weights=[1,1,1,1]

Table 21. ML classification of irregularities - 2 classes

Classifier	**avg 10-cv accuracy**	**test accuracy**	**hyper-parameters**
Naive Bayes	0.9	0.913	$\alpha = 0.01$

Multinomial Logistic Reg.	0.894	0.934	penalty=lasso
			C=5

SVM	0.893	0.922	kernel=linear
			C=1

Random Forest	0.876	0.922	estimators=1000
			criterion=gini

Voting	0.907	0.923	weights=[3,1,1,1]

Confusion matrices

Figure 9. Multiclass confusion matrix - Voting classifier on the Test set

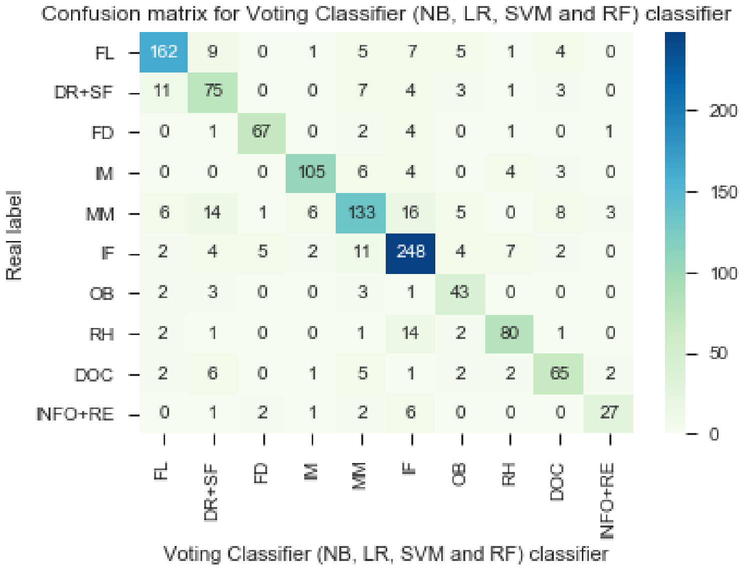
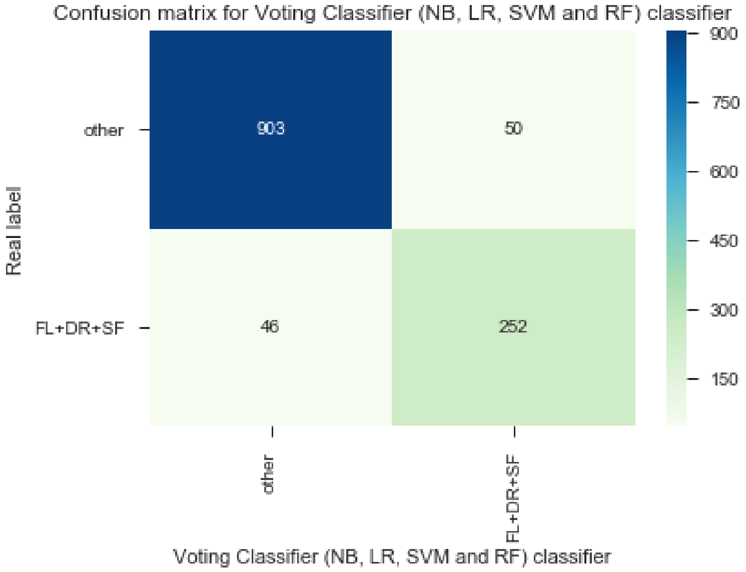


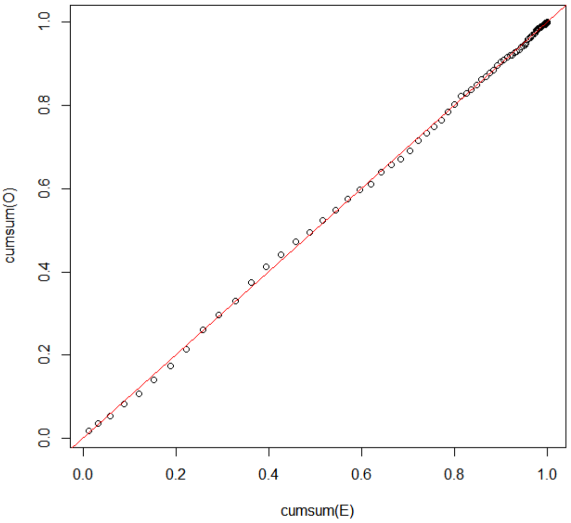
Figure 10. Binary confusion matrix - Voting classifier on the Test set



D. Measurement Error Adjustments

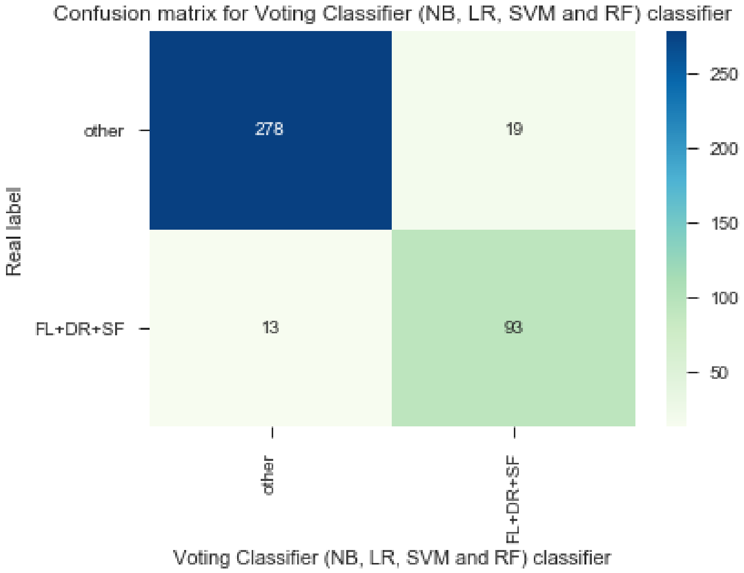
Adherence Test for Number of Irregularities in Education

Figure 11. Quantile-Quantile Plot comparing theoretical Negative Binomial distribution to Number of Irregularities in Education, α=1.73, p=0.08



Confusion Matrix for classification of Municipalities

Figure 12. Confusion matrix for classification of Municipalities based on irregularities predictions


Confusion matrix for Voting Classifier (NB, LR, SVM and RF) classifier
Real label
other
FL+DR+SF
other
FL+DR+SF
Voting Classifier (NB, LR, SVM and RF) classifier