Learning about Corruption:

A Statistical Framework for working with Audit Reports

by Laura Sant'Anna and Duda F. Mendes

Content

Motivation

Quantitative studies in corruption are essential to track externalities and to guide effective public policies to fight against it.

But corrupt acts are hard to be traced and accounted for, therefore data about it is scarce.

- Cross-country indicators of corruption are usually based on perception, such as the *CPI* from Transparency International and the *CPIA transparency, accountability, and corruption in the public sector rating* from World Bank.

Motivation

In Brazil, CGU was created in 2003 to assist in activities of internal control, such as public auditing and fighting against corruption.

The CGU Random Audits Anti-Corruption Program (*Programa de Fiscalização por Sorteios Públicos*) has contributed to overcome data scarcity and has been widely used by researchers in Economics and Political Sciences.

- Nevertheless, the use of reports from this program is sub-optimal: most researchers read and manually classify just a small sample to construct corruption indicators.

Motivation

Ferraz and Finan (2008): found evidence that public disclosure of audit reports before local elections contributed to reduce in 17% the expected probability of reelection of mayors in municipalities where corruption was detected.

Ferraz, Finan and Moreira (2012): found evidence that schooling outcomes are significantly worse in municipalities where corruption in education was detected.

Lichand, Lopes and Medeiros (2016): found evidence that, although the program contributed to reduce irregularities associated to corruption, it was not followed by an improvement in municipal health basic services.

Motivation

Research goals

1. To propose a framework for working with text data in inferential models, using irregularities detected during the CGU anti-corruption program audits as a particular case.

2. To construct and publicly release a dataset of irregularities categorized through a supervised learning task.

CGU Random Audits Anti-Corruption Program

- Randomly selects small and medium municipalities for auditing of federal transfers.
- Inspects federal transfers associated to ministries of Education, Health, Social Development, National Integration, Cities, Agricultural Development and others, depending on the edition.
- Available in pdf: 2,241 reports of audits conducted between 2003-2015 (40 editions).
- Dataset of irregularities (obtained through *LAI*): 81,715 irregularities detected in the 1,223 audits conducted between 2006-2015 (20 editions).

CGU Random Audits 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.

Figure: Example of Service Order in Education executed in Boninal (BA)

CGU Random Audits Anti-Corruption Program

Severe irregularities:

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

Moderate irregularity:

- "O Conselho do Fundeb não atua no acompanhamento da execução do Programa Nacional de Transporte Escolar - Pnate."

CGU Random Audits Anti-Corruption Program

And its conclusion was:

"Com base nos exames realizados, conclui-se que a aplicação dos recursos federais recebidos está inadequada aos normativos referentes ao objeto fiscalizado. A Prefeitura de Boninal/BA contratou e pagou irregularmente uma cooperativa para a prestação do serviço de transporte escolar. O desvio de recursos, apesar de efetivo, é de difícil mensuração com as informações disponibilizadas à Fiscalização."

Methodology

Automate Text Classification

1. Manual classification of irregularities
2. Supervised learning for classification of irregularities

Model estimation with Measurement Error

1. Construction of a municipal indicator of corruption
2. Model estimation adjusting for measurement error

Case Study: Does corruption in educational funds affect schooling outcomes?

Automate Text Classification

Manual classification of irregularities

We randomly selected 4,200 irregularities (5% of the sample) to classify into 12 well-delimited classes, based on Ferraz and Finan (2008) and Lichand, Lopes and Medeiros (2012):

Associated to public corruption	Other
Procurement related flaws (FL)	Councils and public secretaries (IM)
Over-invoicing (SF)	Mismanagement practices in the municipal government (MM)
Resource Diversion (DR)	Poorly evaluated public services and facilities (IF)
	Unfinished projects and goals in public construction (OB)
	Frauds in public programs (FD)
	Human resources anomalies (RH)
	Incomplete documentation (DOC)
	Lack or divergence on published information (INFO)
	Absence of State's government participation (RE)

Automate Text Classification

Related to public corruption:

- [FL] "Exigências indevidas com repercussão na competitividade do certame e direcionamento em licitação do transporte escolar (Pregão n.º 012/2013)"
- [DR] "Falta de comprovação documental das despesas realizadas em 2013 e 2014."
- [DR] "Pagamento por serviços não realizados, no montante de R$13.577,27, referente à despesa do transporte escolar do mês de junho de 2013."

Not related to public corruption:

- [IF] "Intermediação irregular na prestação do serviço de transporte escolar."
- [IM] "O Conselho do Fundeb não atua no acompanhamento da execução do Programa Nacional de Transporte Escolar - Pnate."

Automate Text Classification

Manual classification of irregularities

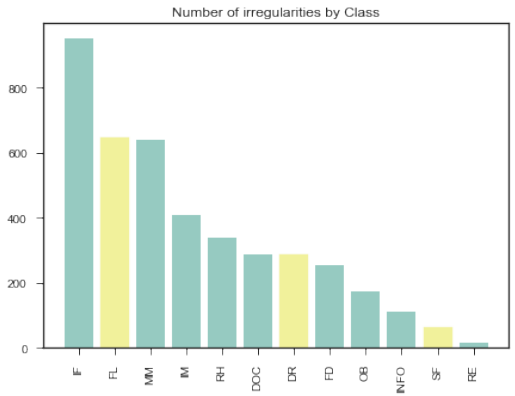

Figure: Number of irregularities manually assigned to each class

Automate Text Classification

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)

Table: Three most frequent (stemmized) tokens in each class and share of irregularities that contain it

Automate Text Classification

Supervised Learning for Text Classification

Considering five classifiers - Naive Bayes, Multinomial Logistic Regression, SVM, Random Forest and a Voting ensemble method - we perform the following steps:

1. Text mining: text pre-processing, exploratory analysis, feature generation
2. 10-fold cross-validation for choosing each algorithm's 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. 10-fold cross-validation to compute accuracy, precision and recall for each model and choose one as our final classifier.

Automate Text Classification

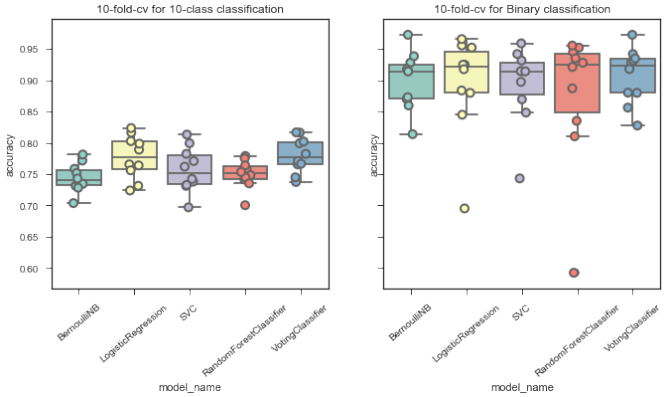

Figure: 10-fold cross-validation for Supervised Learning tasks

Automate Text Classification

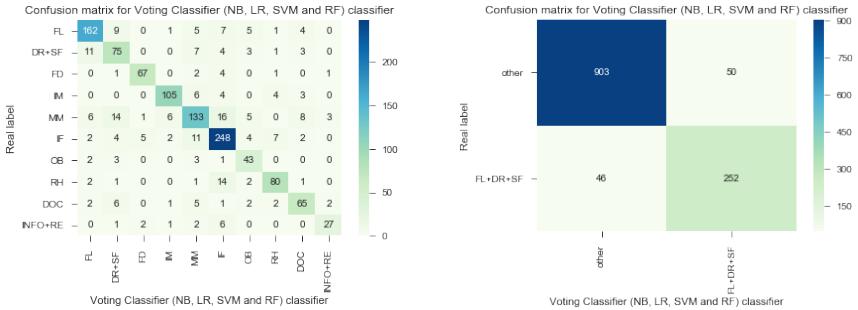

Figure: Confusion matrices on the test set using Voting classifier

Automate Text Classification

Supervised Learning for Text Classification

For the Model estimation step, it is convenient to write the joint probabilities taken from the normalized confusion matrix as:

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

Model estimation with Measurement Error

Constructing a Municipal Indicator of Corruption

For each municipality i we have N_i irregularities D_j classified as 1 (associated to corrupt practices) or 0. An indicator variable of corruption for municipality i is:

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

Since we want an indicator of corruption in Education, we observe three conditions:

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 criteria, in approximately 41% of the audits, at least one irregularity associated to corruption in education was detected.

Model estimation with Measurement Error

We want to use this indicator as explanatory variable in a regression model. Note that it was constructed with a measurement error:

$$C = Z + U_C$$

Thus, given $X = [Z \quad : \quad W]$ we have the following model:

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

We can show that the OLS estimate in this case is asymptotically biased and, using extra-sample information, we can take an MLS approach Johnson (1997) and Savoca (2000) to correct for measurement error:

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

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

Model estimation with Measurement Error

$$\hat{\beta}_{MLS} = \begin{bmatrix} \hat{\alpha}_{MLS} \\ \hat{\theta}_{MLS} \end{bmatrix}, \quad \hat{\beta}_{OLS} = \begin{bmatrix} \hat{\alpha}_{OLS} \\ \hat{\theta}_{OLS} \end{bmatrix},$$

$$\Sigma_{X_M}^{-1} = \begin{bmatrix} \sigma_C^2 & \sigma_{C,W} \\ \sigma_{W,C} & \Sigma_W \end{bmatrix}^{-1}, \quad \Sigma_{X_M U} = \begin{bmatrix} \sigma_{C,U_C} & \sigma_{C,0} \\ \sigma_{W,U_C} & \Sigma_{W,0} \end{bmatrix}.$$

Assuming that variables in W have no measurement error: $\sigma_{C,0} = 0$ and $\sigma_{W,0} = 0$. Since the error U_C is a consequence of the text classification, it is not correlated with municipal variables in W: $\sigma_{W,U_C} = 0$. Thus:

$$\hat{\alpha}_{MLS} = \left(1 - \frac{\sigma_{Z,U_c}\sigma_{U_c}^2}{\sigma_C^2 - \sigma_{C,W}\Sigma_W^{-1}\sigma_{W,C}}\right)^{-1} \cdot \hat{\alpha}_{OLS},$$

where σ_{Z,U_c} and $\sigma_{U_c}^2$ can be estimated with a closed-form using P, r_{01}, r_{10} and the denominator term can be estimated using the inverse of the sample covariance matrix.

Model estimation with Measurement Error

Besides constructing a municipal indicator, we need an estimate for its measurement error.

Particularly, the conditional probability of municipality i being misclassified as "corrupt" when it is not (*false positive* rate), can be calculated through:

$$r_{10,i} = \mathbb{P}(\hat{C}_i = 1 | C_i = 0, N_i) = \mathbb{P}(\exists\{\hat{D}_{ij} = 1\} | \nexists\{D_{ij} = 1\})$$

The case is analogous for the *false negative* rate and taking the expected values on N_i, we have:

$$r_{10} = 1 - \mathbb{E}\left[\left(\frac{\pi_{00}}{\pi_{00} + \pi_{10}}\right)^N\right] \text{ and } \quad r_{01} = 1 - \mathbb{E}\left[\left(\frac{\pi_{11}}{\pi_{11} + \pi_{01}}\right)^N\right].$$

Model estimation with Measurement Error

We propose two approaches for estimating this expected value:

- Assuming an underlying distribution for N: $NB(\alpha, p)$, with $\hat{\alpha}$ and $\hat{p}$ estimated from data.
- Calculating it directly from the confusion matrix, grouping irregularities by municipality for cross-validation.

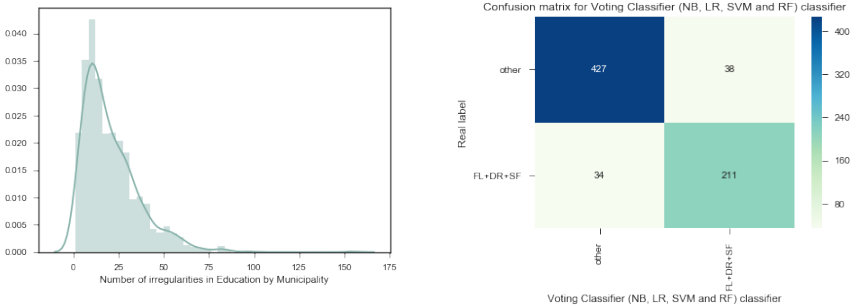

Figure: Number of Irregularities in Education Frequency Distribution and Confusion matrix for the classification of municipalities

Case Study: Does Corruption affect Educational Outcomes?

As an application of this framework we propose replicating the empirical strategy proposed by Ferraz, Finan and Moreira (2012), that estimates the relationship between corruption in educational transfers and schooling outcomes:

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

- $A_{s,m,4}$ is a vector of average student achievements for each school s in the 4th grade,
- C_m is a measure of corruption in education,
- Z_m is a vector of municipal characteristics,
- $X_{s,m}$ is a vector of students characteristics, aggregated by school.

Case Study

Since we have irregularities detected in audits taken in three political terms (2005-2008, 2009-2012, 2013-2016), we gathered data for available years within this range.

Educational outcomes and school characteristics:

- Math and Portuguese scores, selected Students and Principals surveys answers from Prova Brasil/SAEB, INEP.
- Failure and Dropout rates, existence of computer lab, science lab and sanitation from School Census, INEP.

Municipal characteristics:

- Gini and Population from 2010 Census, IBGE.
- GDP per-capita from Gross Domestic Product of Municipalities, IBGE.
- Municipal spending per pupil in primary education from SIOPE, MEC.

Case Study

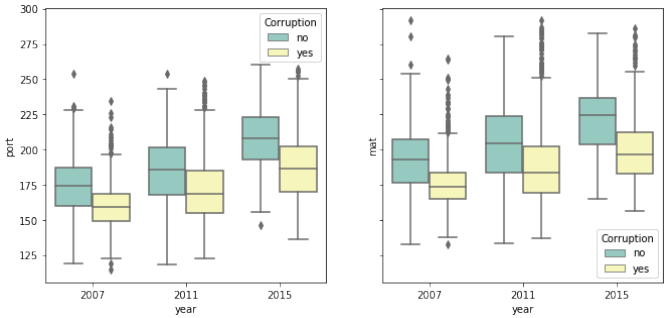

Figure: Portuguese and Math ccores on Prova Brasil, split by term and incidence of corruption in Education

Note: Among 4,670 schools, approximately 49% are in municipalities where exists corruption in Education, according to our indicator.

Case Study

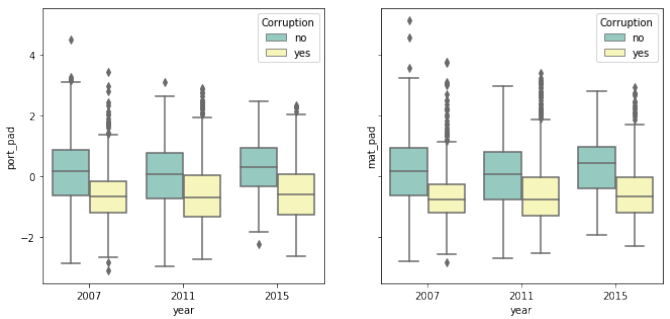

Figure: Portuguese and Math standardized scores on Prova Brasil, split by term and incidence of corruption in Education

Case Study

Table: 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

Case Study

Table: Effects of Corruption on Mathematics Standardized Test Scores

	OLS Estimation	**MLS Estimation** (1)	(2)
Corruption	-0.1098***	-0.183***	-0.152***
in Education	(0.025)	(0.042)	(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. $N = 4670$

Case Study

Table: Effects of Corruption on Portuguese Standardized Test Scores

	OLS Estimation	**MLS Estimation** (1)	(2)
Corruption	-0.111***	-0.185***	-0.153***
in Education	(0.024)	(0.04)	(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. $N = 4670$

Case Study

Table: Effects of Corruption on Dropout Rates

	OLS Estimation	**MLS Estimation** (1)	(2)
Corruption	0.003**	0.005***	0.004***
in Education	(0.001)	(0.002)	(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. $N = 4670$

Case Study

Table: Effects of Corruption on Failure Rates

	OLS Estimation	**MLS Estimation**	
		(1)	(2)
Corruption in Education	0.006*	0.009**	0.008**
	(0.003)	(0.005)	(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. $N = 4670$

References I

Conclusion

In this study we proposed a framework to guide the efficient use of text data in inferential models, taking Audit reports from the CGU Anti-Corruption program as a particular application.

Overall, we obtained satisfactory results using the proposed framework, both on the machine learning classification step and on the regression model estimation.

Nevertheless, we understand that it has some limitations to 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.
- Although we proposed a particular use case, we expect this framework to be useful with other sources of text data, accounting for differences in structure and vocabulary.