UNIVERSITAT JAUME I

VNIVERSITAT DE VALÈNCIA

Master’s degree in Economics
Master’s thesis

Developing corruption indicators at the municipal level: prosecution mined data for Spain

Supervisor:
Jesús Peiró-Palomino

Author:
Jakub Jelinek

Academic year 2020-2021

ABSTRACT

After wide evidence of government quality effects in the economic performance at national, and recently at the regional levels, analyzes at the municipal level count with few examples due to severe data limitations. This paper reviews the corruption indicators used in the literature with different purposes and develops a data mining model able to identify municipal corruption cases from a Spanish judicial database. Our model identified 773 municipal corruption cases prosecuted in Spain during the period 2000-2020, and after validation tests, we found an accuracy of 92% in the data extraction and significant correlations between the number of cases and the European Quality of Government Index (EQI) at the NUTS-2 level. Nevertheless, the application of this data for further analyzes is still limited by a lack of statistics with municipal detail.

KEYWORDS

Corruption indicators, data mining, municipalities, quality of government.

JEL O43, R11, R5

1. Introduction

The relevance of the institutional framework in determining many aspects of economic development has been largely studied and the evidence strongly supports the importance of good governance (e.g. Acemoglu & Johnson 2005). This quality of government is measured in the literature as a combination of the different channels through which the institutions influence economic activity and social wellness.

Between the main channels usually studied we find; the legal framework, positively evaluating laws that defend the property rights and business; the effectiveness of the government management; or the presence of corruption, which limits the effective share of rents, acting as an illegal tax. Additional different factors can be found in the literature as health, education, or environment protection indicators (Charron & Lapuente 2013)

The institutional quality research has mainly focused on the country level, analyzing the evolution over years or comparing between different countries. In recent years the study of the institutional factors has extended to the regional level, adding new evidence for the relevance of the institutional quality in determining between regions in; economic growth (Palomino et al. 2019) welfare (Palomino 2019) or economic resilience (Rios & Gianmoena 2020).

Whereas, a scarce number of studies are able to analyze this phenomenon at the municipality or local[1] level. One of these few examples is Rodríguez-Pose & Zhang (2019) which found a significant effect of the governmental efficiency and fight against corruption in the urban growth of 283 Chinese cities, after obtaining isolated indicators for these urban areas.

The main reason behind the low number of studies about institutional determinants at the local level is the difficulty to obtain representative indicators, a problem that is even more pronounced for corruption indicators. However, new technologies offer a potential solution to overcome the local indicators limitation with high detail and segregated data.

Big data, artificial intelligence and data mining techniques provide new indicators for different fields which amplify, complement or substitute traditional data. These types of indicators are especially interesting to fill gaps where the data cannot be provided by traditional methods or is not reliable.[2]

In this paper, we build a database that includes the number of corruption cases prosecuted by municipality in Spain for the period 2000-2020, which might suppose the first local corruption indicator being able to cover the totality of the Spanish territory. For this purpose, we propose a data mining method, consisting of a combination of data scraping and natural language processing techniques capable of identifying corruption cases prosecuted related with the local government, including information about the municipality and being able to disaggregate by corruption type. The result is the identification of 773 corruption cases prosecuted, corresponding to 446 municipalities.

[1] Understanding by local the smaller possible administrative level. Several articles refer to local studies when considering regional levels, relative to NUTS-2 or NUTS-3 levels.

[2] Two examples are Cavallo (2013) who used web scrapping techniques in online stores in order to build prices indicators, which allow for example inflation monitoring, showing differences with official inflation reports from Argentina; and Hnderson et al. (2012), the authors used satellite data to perform indicators of development and welfare in African rural areas.

Our method presents two main contributions to the existing literature. First, it offers an objective corruption indicator at the local level capable of covering the full national territory, with a high level of detail, which might contribute to new studies of institutional quality at the lowest administrative level. Second, we provide an example about how the big data techniques are a great cost saving alternative for the collection of prosecution data, which traditionally has supposed a huge time-consuming task. Decreasing the cost of obtention of the data facilitates covering larger periods of time in the studies, none of the previous works reviewed more than 12 years.

Additionally, this methodology can be easily extended to all types of crimes, which would be of particular interest for other economic crimes with difficulties to obtain reliable indicators, such as money laundering, drug trade, tax evasion etc.

In the next pages, we will first review the literature about corruption and the types of indicators used, with a critical comparison between indicators. Following, we will present our methodology for the database construction, present the information extracted and their correlation with institutional quality indicators and economic resilience. Finally, we will discuss the limitations, benefits and potential of these methods and present our conclusions.

2. Literature review

2.1 Does the quality of local governments impact economic performance?

Even as the number of contributions are few, several arguments are pointing at the role of municipal government quality as a determinant of different economic dimensions. The negative effects in the economy of bad governments at the national and regional levels, largely evidenced in the literature, are expected to be in line with those at the local level (Balaguer-Coll et al. 2021).

At the national level, low quality governments are related with less transparent labor markets (Di Cataldo & Rodríguez-Pose 2017), lower trade and growth (Dollar & Kraay 2003), or lower investments (Buchanan et al. 2012). Particularly, corruption is linked with inefficient public spending and education (Mauro 1998), greater income inequality (Gupta et al. 2002) and worse environmental outcomes (Welsch 2004).

Studies at the regional level, mainly performed in Europe, found significant effects of institutional quality on economic growth (Palomino et al. 2019) welfare (Palomino 2019), economic resilience (Rios & Gianmoena 2020), (Ezcurra & Rios 2019), returns of public investments (Rodríguez-Pose & Garcilazo 2015) or innovation (Rodríguez-Pose & Di Cataldo 2015)

The channels in which government institutions influence economic performance at the local level might be for one side, directly through the efficiency of the public spending and their consequences in public services such as education or infrastructures. On the other side, inefficient governments increase transaction costs and dissuade economic activity (Rodríguez-Pose & Zhang 2019).

Recent studies are adding evidence to the unexploited field of the effects of local institutional quality on economic performance. Rodríguez-Pose and Zhang (2019) combined public management efficiency and corruption prosecution measures of local institutional quality in a sample of 283 Chinese cities, finding a positive relationship between government efficiency and the fight against corruption with urban growth.

This paper will focus on Spain, for which Balaguer-Coll et al. (2021) measured government effectiveness through government spending in Spanish municipalities with a population between 1.000 and 50.000 inhabitants, the results prove a positive effect of government efficiency on disposable income per capita growth.

Hortas-Rico and Rios (2019) analyzed the determinants of local income inequality in Spain with a sample of 977 municipalities, concluding that corruption, among other local political factors, has distributional consequences. The authors measured corruption with an index capturing the number of corruption scandals both in each municipality and the region it belonged[3], obtained from press sources during the years 2000-2006.

2.2 Measuring corruption.

The complexity of the corruption phenomenon explains that there does not even exist a consensus for its definition. For this paper, we will consider the corruption definition of Transparency International, according to which, corruption represents the abuse of entrusted power for private gain. An extended segregation of its main forms in the literature can be classified in; bribery, embezzlement, fraud, and extortion (Andvig et al. 2000, 14ff.)

Obtaining valid corruption indicators is a struggle by itself, the reason is obvious and clearly synthesized by Kaufmann et al. 2007: "Since corruption is clandestine, it is virtually impossible to come up with precise objective measures of it." (p. 3).

We can summarize the corruption indicators used in the literature in the following 4 groups:

(1) Perception indicators based on surveys or interviews;
(2) Individual detailed and specific studies (mainly audits);
(3) Proxy indicators, and;
(4) Official reports from institutions and collections of data from prosecuted cases or press.

(1) The usage of survey indicators is the most extended and widely discussed in the literature. The main indicators of this type are the European Union's (EU) European Quality of Government Index (EQI), the Transparency International's (TI) Corruption Perceptions Index (CPI), and the World Bank's Worldwide Governance Indicators (WGI). For an extended review of the methodology, main findings and problems of those indicators see (Rohwer, 2009), (Knack, 2006), (Razafindrakoto, & Roubaud 2010).

(2) Less generalized in academia, and with much more specific applications we can find the indicators based on individual studies or audits. Among the notable examples; (Olken, 2007) collected data on spending reports for road projects and compared them with engineer valuations for each project of the real costs. The findings support that higher controls reduce corruption risk and spend efficiency, where increasing the audits from 4% to 100% of 600 road projects led to a cost reduction of 8%.

[3] The index $CI = \frac{C_i - C_{min}}{C_{max} - C_{min}}$ where $C_i = 0.5CC_i + 0.5CR_i$ with CC_i a dummy variable indication the exitence of a corruption scandal in the municipality and CR a continuous variable with the number of scandals in the region

Colonnelli & Prem (2020) evidenced an increase of the activity in the local Brazilian economies after campaigns of random anti-corruption audits. Avis et al., (2018) pointed out that the random audits in Brazil were successful in reducing corruption by increasing their perceived nonelectoral costs of corruption, which represent additional support for the link between corruption reduction and increase in economic development. These articles benefited from the reports of the audits that offered highly detailed data of the irregularities found after weeks of fieldwork in each selected locality.

(3) Another important group of indicators proxy the presence or the risk of corruption, mainly using public administration data. One example is Hsieh & Moretti (2006), who compared market oil prices with the official settlement prices of Iraq sales under the UN embargo relief for a humanitarian program, arguing that Iraq sold oil below market price to collect bribes and political favors from the oil buyers.

The increasing number of administrative data available in conjunction with the usage of big data has made possible new indicators and research. The leading examples are the indicators based on public procurement electronic data. Procurement is one of the fields of public administration more prone to corrupt procedures. Public procurement represents approximately one third of the total public expenditures in the OECD countries (OECD 2016); the combination of conflicts of interest with large sums contracts and the lack of transparency in some processes creates a high risk for corruption.

Several studies using electronic procurement data to develop proxy indicators for corruption stand out in this type of new indicators. These indicators of public spending have proved a significant association with GDP per capita and the perception of institutional quality in regional surveys (Fazekas et al. 2016).

(4) The last group is the least used in economic studies. The phenomenon of corruption has been largely studied in Spain, gaining considerable attention after the 2007 crisis, with good examples for the application of prosecution indicators. Most of the studies are related to political accountability, with individually studied data in different levels of detail.

Jiménez & García (2016) built a database of cases from new sources available electronically, considering the cases they could match in at least 2 "reliable" sources. They identified 234, 18, and 8 corruption cases at the respective, local, provincial, or regional levels during the period 2000-2011. The same database was used later by Lopez-Valcarcel et al. (2018) to test for the presence of geographical contagion in the Spanish corruption cases.

Darias et al. (2012) focused on urbanistic crimes, building a database of the urbanistic corruption cases that they could find in the news. This work identified 676 municipalities affected by urbanistic corruption cases which were covered by electronic newspapers during the years 2000-2010, which represents that 8,3% of the total Spanish municipalities were affected by this type of corruption.

None of the reviewed databases built could represent the totality of the Spanish territory for a major part of corruption crimes, with detail of the municipality level. This paper intends to fill this gap in the literature, proposing a method for obtaining a representative indicator for corruption at the minimum administrative level, covering the totality of the national territory.

3. A critical review of the main corruption indicators

	Cost	Detail level	Comparable accross countries	Comparable in time	Disagragate corruption type	Accuracy	Examples in the literature
(1) Perception	Low	Low	Yes	Yes	No	Low	(Palomino 2019), (Rios 2020)
(2) Audits	Very high	Very high	No	No	Yes	Very high	(Olken, 2007), (Colonnelli & Prem, 2020)
(3) Proxies	Low	Low	Yes	Yes	No	Low	(Hsieh & Moretti, 2006), (Fazekas et al. 2016)
(4) Prosecution	High	High	No	Yes	Yes	High	(Jimenez & Garcia, 2016), (Darias et al. 2012)

Table 1: Comparative between corruption indicators. Source: Own elaboration.

In the previous section, we presented the main types of corruption indicators with remarkable examples found in the literature. In this part we will extend this analysis, with the pretension of, on one hand, clarifying the strengths and weaknesses of each category and on the other hand help to identify the best indicator type depending on the characteristics and the goal of the study. In the table 1 we present a summary of the categorization for the different types of indicators.

Prior to that, we need to add another extended distinction of corruption, which is between small daily cases, more related with informality in the economy or petty corruption, and grand corruption, which stands for structured and high-level forms of corruption more involved at the political level (Rohwer, 2009). We find this distinction important as might affect considerably on how effectively can be measured by each indicator type.

(1) We begin this comparison with the most used type of indicator, the perception of corruption indexes. The perception indicators had received several critiques in the literature. The main critiques respond to the usage of surveys, which are subject to biases such as the economic context or political orientation.

Another important problem that presents this type of indicator is that the methodology changes from year to year, mainly in the sources of the survey, which rest validity for using the time series data, or to make year-to-year comparisons.

The main advantage is how easily they can be integrated into a study, with the data publicly available, with some of those indexes arriving at the regional level. Often these indicators are part or integrated of other institutional quality indicators, which facilitates to determine the impact of different institutional framework categories at once.

Additionally, perception indicators might be a strong option to monitor petty corruption. Corruption in the administration, police, or any public servants in contact with the citizens can be monitored with surveys about bribes paid, or corruption observed.

(2) For the second type of indicator, the individual specialized studies, we can find the most reliable way of determining the presence of corruption with enough detail to add intensity or any form of specific characteristics.

The main limitations of this indicator category are first, the high cost needed for its elaboration and second, the level of specification which makes it difficult to extract generalized conclusions or compare between regions. The level of accuracy obtained makes this indicator ideal for testing strategies against corruption or studies in regions with low reliability in the official data or prosecution.

(3) Considering the proxy indicators, they have important limitations. This was advanced by Kaufmann et al. in 2007 way before the start of the large open data initiatives "One can also obtain objective data on institutional features such as procurement practices or budget procedures that may create opportunities for corruption (…) But these will only be imperfect proxies for actual corruption, not least because the "on the ground" application of these rules and procedures might be very different." (p. 2).

Even not considering this source of biasness, particularly the indicators based on e-procurement data are limited first by the number of tenders available (the obligation to follow a tender process is required for contracts above a certain amount) and only covers certain types of corruption related to the government spending. This excludes all the types of corruption related to urbanism, fraud, or extortion.

Nevertheless, we can find several advantages; they are objective, quantifiable, and traceable in a time series period. These indicators stand as a good alternative or complementary to perception or prosecution indicators. The cost of elaboration is particularly low if we consider that these indicators come as a by-product of already available data.

(4) Finally, the indicators based on prosecuted cases represent a reliable identification of corruption that allows for an extended analysis across the national territory. The source of this data can be directly from the judicial data or through press sources. The use of press sources for the data might be beneficial for studying the effects in the field of political accountability, as it supposes an indicator of the public exposure of corruption.

The main disadvantages are that the judicial power or press has limited capacity and cannot cover the totality of the cases and this information would be poorly represented in the countries where corruption is widespread.

In terms of elaboration, the cost for obtaining this type of indicator was relatively high compared to proxies or survey data, as it required for the individual analysis of cases. An idea of this cost can be found in (Darias et al. 2012), translated from the paper, the database is the result of "*a tedious and incessant search of news*".

Big data and ML learning techniques represent a way for avoiding this high cost disadvantage, using data mining to automate the case collection, with the possibility to module the information intended to extract, as we will present in the following sections.

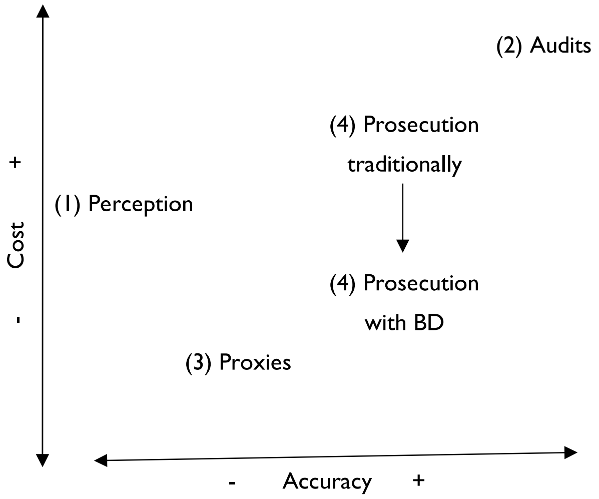


Figure 1: Comparative between corruption indicators. Source: Own elaboration

4. The database

Our sample covers the corruption prosecution data in Spain, including the islands, for the period 2000-2020. In contrast with similar works, our method considers the totality of the territory, without discriminating by the number of population. The final database contains data about 8.136 municipality codes of Spain, of which 446 municipalities presented at least one prosecuted case of corruption.

In the following lines, we will present the technology used, data sources, and steps in building this database that allows for the identification of cases related to the municipal government and identifying the municipality. Finally, we will present the characteristics and a summary of the results.

4.1 Technology used

A. Data mining (Python)

Web scraping
-Selenium: Originally intended to perform automated tests in web environments, allow us to program the interactions that we perform when surfing a website.
-BeutifulSoup: This python package brings the possibility to extract the information from HTML documents and with a few command lines extract the pieces of information targeted.

Natural Language Processing (NLP)
-Spacy: Using this library simplifies many NLP tasks such as tokenization, NER, lemmatization… It includes the tools to build a "matcher" to detect the presence and/or extract information with personalized rules. Additionally, it contains pre-trained models for 64 languages, including Spanish, which can be used as a base for training custom NLP models.

Geoparsing
-Geopy: Georeference the data, with the input of an address, city, or country it retrieves the coordinates with accuracy. In our script, this function allows us to georeference each case once we identify the municipality name.

B. Data visualization (R)

-Leaflet: Creates web maps with geolocated tags. It will be used for the spot representation of cases in the map.
-rgdal & ggplot2: Allow to transform "shapefiles" for R and plot. These tools are useful for a representation of the data considering the territories.

4.2 Data sources

In Spain exist several databases of jurisprudence, most of them private and commercialized by law editorials, despite of the public database of the Spanish judicial system or C.G.P.J for the initials in Spanish for General Conseil of the Judicial Power. The public database however is not available for the purpose of this project as it explicitly forbidden massive downloads.

Among the different private databases that we could access, we selected the database of the editorial Tirant, a private database that incorporates a big data analysis system that eases the selection of cases. This engine allows us to search all the available cases related to specific law and therefore disaggregate the corruption cases by a specific type of the following corruption crimes: Influence traffic, Urbanism, Embezzlement, Fraud, and Bribery.[4]

In the following table, we summarize the type of corruption crime with relation to the specific law that is related, according to the classification of the C.G.P.J.

Corruption type	Spanish law related
Urbanism	Arts. 320 & 322 CP
Bribery	Arts. 419, 420, 421 & 422 CP
Influence trafic	Arts. 428, 429 & 430 CP
Embezzlement	Arts. 432, 433, 434 & 435 CP
Fraud	Arts. 436, 437 & 438 CP

Table 2: Corruption type by law. Source: Own elaboration from C.G.P.J

Additionally, to the corruption cases, we included demographic data about population to all the geographical levels, obtained from the Spanish statistical office or INE and the EU European Quality of Government (EQI) index at the NUTS-2 level[5]. Finally, we included data about unemployment, at the province or NUTS-3 level, directly provided by the INE and we calculated the unemployment rates at the municipal levels.

National data with the municipal detail is scarce, at least in the case of Spain, and so is the case of the unemployment rates. The Spanish national workers insurance system provides data about the number of people in situation of unemployment on monthly basis, this data averaged together with an extrapolation of the active population between the municipal population average and the percentage of the active population in the province, both obtained from INE, gave as an approximation to the unemployment rates at the municipal level, the details of this calculation will be presented in section 6.

This complementary data will serve to perform tests about the validity of the sample, testing the correlation with institutional quality indicators which we might suppose that should present a significant relation. Additionally, we want to use this dataset for an approximation to their potential in new research areas about institutional quality with a local approach, more specifically analyzing the relation between the number of prosecuted corruption cases and economic resilience at the municipal level.

4.3 Elaboration

The design of the strategy for the data extraction was inspired by the works of Espinosa (2019) and Hernández-Díaz (2017), both developed data mining codes of legal information which served as a base to adapt for our specific needs of the firsts steps of corruption cases identification.

[4] This segregation allows us to make a distinction similar to the main corruption types usually found in the literature: bribery, embezzlement, fraud and extortion (Andvig et al. 2000, 14ff.); apart from this distinction we find the necessity for including crimes related with urbanism, which as we will show in the results represent a high amount of the total cases prosecuted.

[5] Available at: https://www.gu.se/en/quality-government/qog-data/data-downloads/european-quality-of-government-index

The first step is the selection of the cases, by filtering all the prosecuted cases related with the laws of the previous table. As the database does not have the option for massive downloads, we used the script B1 in Appendix B, which consist in a crawler that given a list of laws, downloads the summary data and the ID of each of the cases.

Most of the corruption cases prosecuted were related to more than one of the corruption types analyzed, after eliminating the duplicates the total number of cases identified after our first step was 2.585. At this level, we only could have information of the court and the corruption types to which each case corresponds.

Once we have identified all the judicial cases related to corruption in the period, the next step was the identification of the cases at the local level. This is the purpose of the script B2 in Appendix B, which consists of a crawler that with the input of a list of ID's, checks each case summary of the facts and identifies the mentions to a major or to a council, for which extract the name of the municipality that follows the mention.

4.4 Description of the data

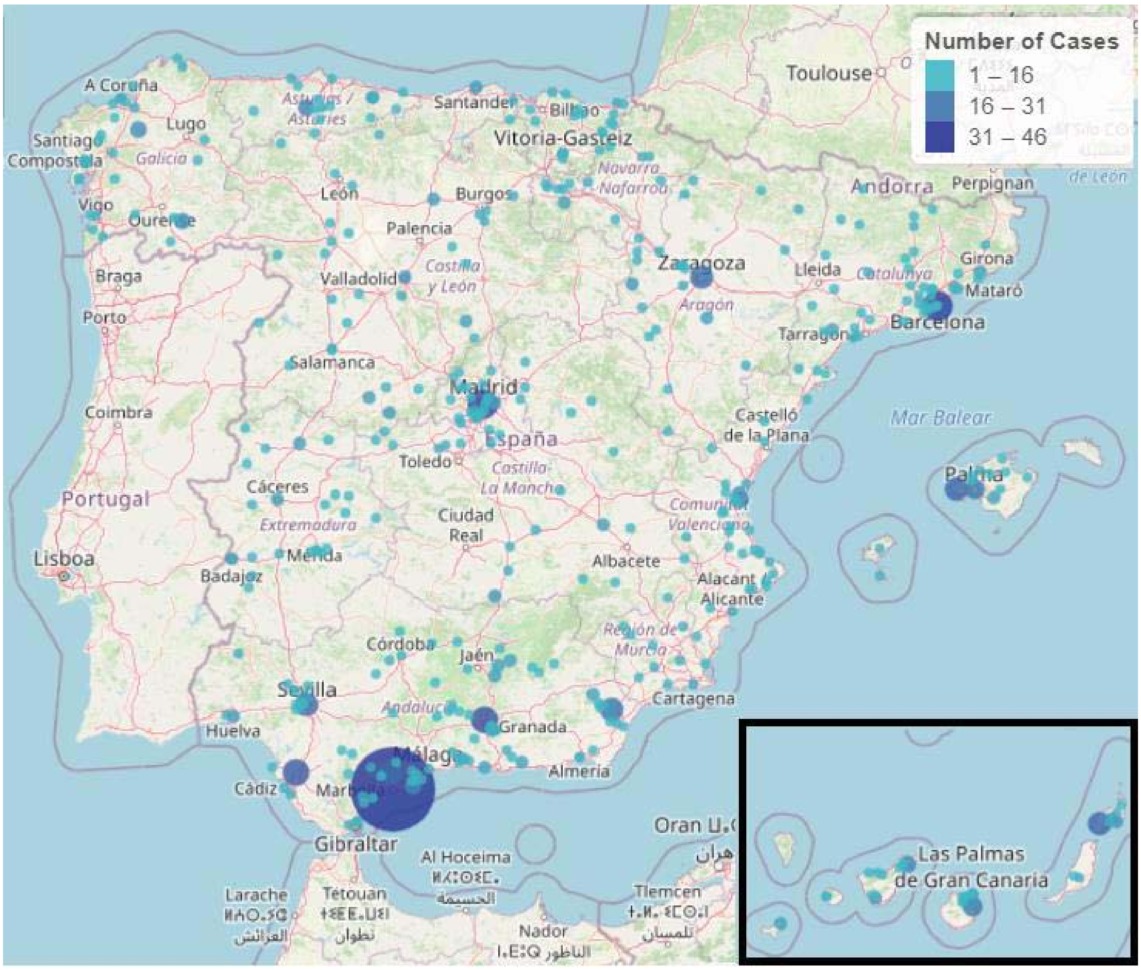


Figure 2: Number of local corruption cases prosecuted by municipality with sentence during the years 2000-2020. Source: Own elaboration.

Our resulting database contains 773 corruption cases[6], which corresponding to the type of corruption are: fraud (134), traffic of influences (93), bribery (124), urbanism (232), and embezzlement (357).

We must differentiate between the results according to the total number of cases and the results in relative population terms, as there are significant differences in the population between regions.

In our first geographical level, the regions corresponding to the NUTS-2 level, Andalucia concentrates an important part of the total number of cases (236), followed by Catalonia (73) and the Canary Islands (71). On the opposite side of the chart, the northern regions present a low number of cases Asturias (13), Cantabria (11), La Rioja (9), and Navarra (6).

However, when considering the population in each region, the outlook changes significantly. Andalucia moves to 4th place (30 cases by million citizens), Canary Islands (36) and Balear Islands (35) lead the chart, together with La Rioja (31) which contrast with their position in the total number of cases.

Region (NUTS-2)	Number of cases	Cases by million citizens	Fraud	Influence Trafic	Bribery	Urbanism	Embezzlement
Andalucía	236	30	56	14	34	99	99
Cataluña	73	10	4	15	11	14	34
Canarias	71	36	8	12	13	29	31
Castilla y León	62	25	9	3	8	13	37
Madrid, Comunidad de	54	9	4	11	17	13	16
Comunitat Valenciana	41	9	5	7	9	12	18
Balears, Illes	35	35	12	5	4	15	12
Aragón	34	27	10	7	6	2	24
Extremadura	32	30	9	6	0	7	19
Galicia	32	12	4	6	1	6	18
Castilla - La Mancha	31	16	2	3	3	8	17
País Vasco	19	9	2	1	4	4	8
Murcia, Región de	14	11	4	1	4	2	5
Asturias, Principado de	13	12	3	0	3	3	5
Cantabria	11	20	0	0	2	5	5
Rioja, La	9	31	0	2	5	0	4
Navarra, Comunidad Foral de	6	10	2	0	0	0	5
Total	**773**	**332**	**134**	**93**	**124**	**232**	**357**

Table 3: Corruption cases summary by NUTS-2 region. Source: Own elaboration

In the next geographical level, the provinces, corresponding to a NUTS-3, the highest numbers of cases corresponds to Malaga (88), Barcelona (54), and Madrid (54) while the lowest numbers are observed in Soria (0), Guadalajara (2) and Lleida (3). Similar to before, when considering the population, the balance change, Malaga maintains the first position (5,94 cases by million citizens), but Avila (5,58) and Zamora (5,35) appear in the next positions of the top. Main provinces as Madrid (0,9), Barcelona (1,03) and Valencia (1,04) place in the bottom quarter of the table.

The granularity of the data allows us to deepen this analysis to the lowest administrative level, the municipalities. Marbella presents the highest number of corruption cases (46), more than 50% of the Malaga region.

This level of the specification allows us to find the clusters, taking into example the cases in the Canary Islands, with tops both in the total number of cases and relative population terms, the 80% of the cases are found in 15 municipalities (of a total of 88), outstanding Yaiza (10 cases, 15% of the total), Telde (7 cases, 10% of the total) and Santa Cruz de Tenerife (6 cases, 9% of the total).

[6] The dataset that support the findings of this study is available upon request to the author.

Analyzing the corruption cases by type of delict, we will focus on urbanism (30% of the cases) and embezzlement (46% of the cases).

We find that almost a half of the total number of municipal corruption cases related to urbanism (99 o 232) are in the region of Andalucía, distributed inside the region as follows: Malaga (40), Granada (26), Almeria (20), Jaen (5), Sevilla (4). The municipalities that outstand in this category are Marbella, Malaga (14), and Atarfe, Granada (12).

The island follows at the top of urbanism cases, with Canary Islands (29 cases) and the Balear region (15 cases). Navarra and La Rioja did not presented cases related to urbanism in this analysis.

Andalucía is the first region again in the number of embezzlement cases (99) almost a third of the total cases (357), followed by Castilla y Leon (37) and Catalonia (34). At the provincial level Malaga (40), Barcelona (19), and Zaragoza (18) present the higher number of cases. The higher municipalities are Marbella (30), Jerez de la frontera (7) and Zaragoza (6).

Finally, the chart corresponding to the Figure 3 evidence an increase of the number of judicial sentences related with local corruption cases after the year 2011, with a steady increase until maximum levels in 2018. The year of the sentence does not reflect the time were the corruption activities occurred, however this data suggest an increase of the fight against corruption after the financial crisis.

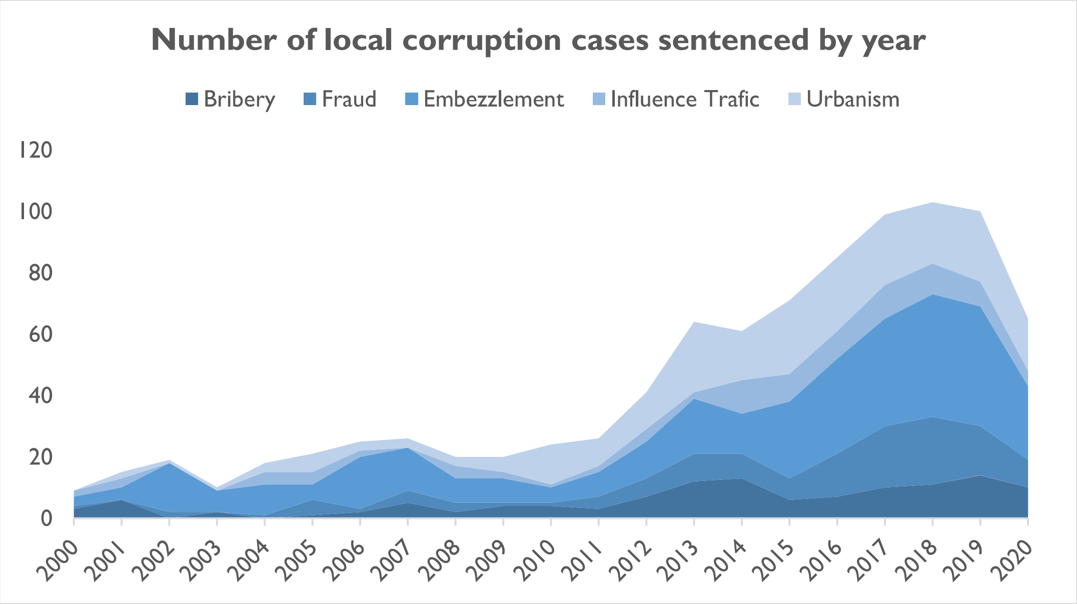


Figure 3: Local corruption sentences over time. Source: Own elaboration.

5. Sensitivity analysis

5.1 Accuracy of the data mining model

Contrary to more sophisticated machine learning methods which allow for the computation of accuracy and efficiency measures, our data mining method requires a manual verification of the outcome.

We designed a verification strategy to test the performance of the method, and control for the possible mismatches or potential sources of biasness of the data extraction. Out from the final sample of 773 corruption cases related to municipal corruption, we selected a subsample of 63[7] aleatory cases, and individually revised each case file to verify the following questions with a yes/no answer:

(1) Does the case respond to an abuse of entrusted power for private gain[8]?
(2) Does the case imply higher geographical powers, i.e. regional servants?
(3) The municipality identified by the script is correct?
(4) Are more than one municipalities implied in this case?

The report of the validation is presented in appendix A table A2, out of this subsample the results indicate that 92% (58/63) of the identified cases were successfully identified, and the case responds to local corruption cases, in which no additional municipalities were involved and the municipality affected was correctly captured.

A 5% (3/63) of the cases were related to public servants but the crime did not correspond to an abuse of their entrusted power, therefore we do not consider those cases as corruption cases. Regarding the other possible sources of biasness, 2% (2/63) corresponded to higher levels of power, one provincial and one regional while similarly a 2% of the cases were related with a higher number of municipalities that were not captured.

5.2 Correlation with the European Quality of Government Index

As mentioned in the previous section, our database includes the EU European Quality of Government EQI indexes, for the years 2010, 2013, 2017, and the average of the three, for the NUTS-2 level.

The EQI is a perception indicator elaborated from Quality of Government (QoG) surveys that include questions categorized in the following categories of institutional quality; the corruption level, the impartiality, and the quality of the public services. The data available corresponds to the publications of the years 2010, 2013, 2017, and 2021, all of them available with both NUTS-1 and NUTS-2 levels. One important remark is that the index ranges from 2.5 to -2.5, with mean 0 as the average EU level, therefore is a relative indicator, where values below 0 represent results below the EU average, and lower index represent lower levels of governmental quality.

[7] Which corresponds to a 90% confidence level with a 10% error possibility subsample.

[8] In other words, we want to differentiate if the case responds to a corruption case as per the International Transparency definition.

The corruption-related questions have an important weight in the QoG survey questions[9], consequently, we might expect to find a significant correlation between the number of prosecuted cases and the perception of the corruption represented in the regional EQI indexes.

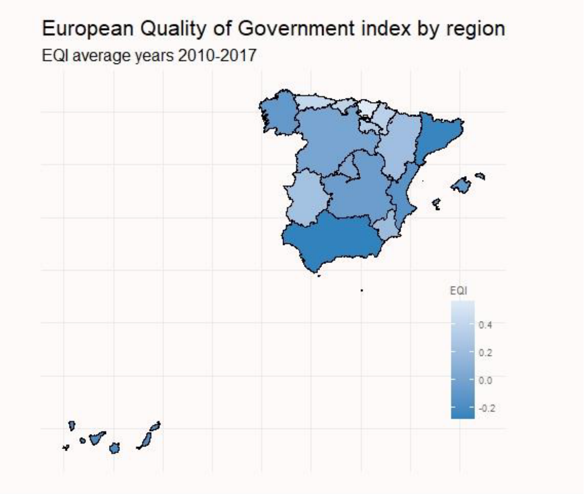

European Quality of Government index by region
EQI average years 2010-2017
EQI
0.4
0.2
0.0
-0.2

Figure 4: EQI by NUTS-2 region. Source: Own elaboration from EQI.

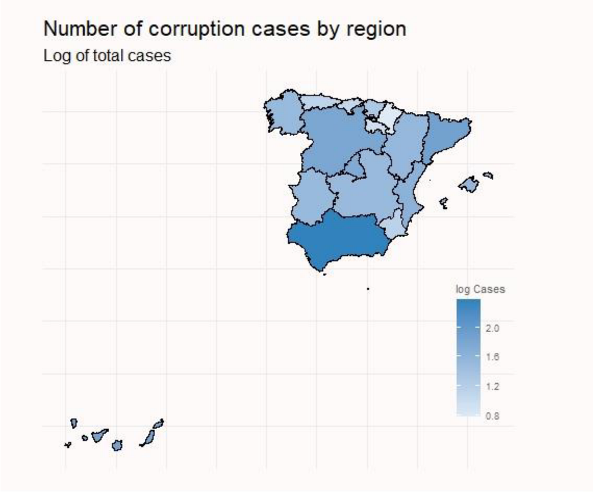


Figure 5: Log of corruption cases by NUTS-2 region. Source: Own elaboration.

We tested the relation between the number of cases prosecuted and the EQI, in the chart corresponding to the figure 6 we present the result. We find a high Pearson correlation index between the number of cases in total or logarithm, with the average of the EQI. This relation is negative as a higher number of cases is related to a lower level of EQI. In our opinion, the log version of cases is more appropriate to compare with the EQI rather than the total number of cases due to the small range and relative nature of the EQI (-2.5 to 2.5), the correlation index when comparing with the EQI average 2010-2017 with the total number of cases is still significative (-0.64).

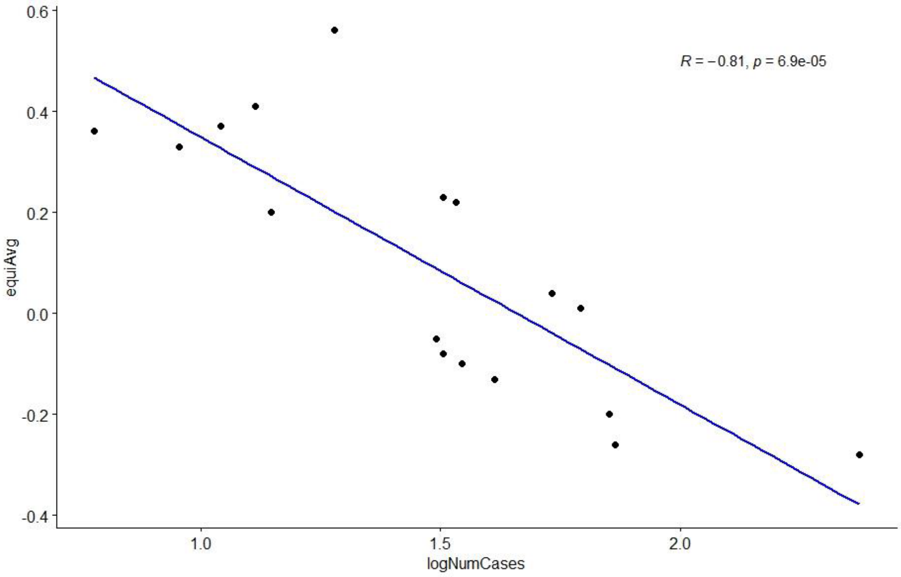


Figure 6: EQI and log of Corruption cases correlation. Source: Own elaboration with EQI data.

[9] The questions related with corruption in each survey year are; 8 out of 20 in the EQI 2017 (Charron et al. 2019), 5 out of 16 in 2013 (Charron et al. 2015), and 6 out of 16 in 2010 (Charron et al. 2014).

However, we do not find such a strong correlation between the cases and EQI when considering the population. One of the possible hypotheses behind this finding is that larger regions present a higher number of corruption cases that have a higher exposition in the media, inducing a higher perception of corruption. This idea would be supported by the relation between population and EQI, with a correlation index of 66% for the EQI average.

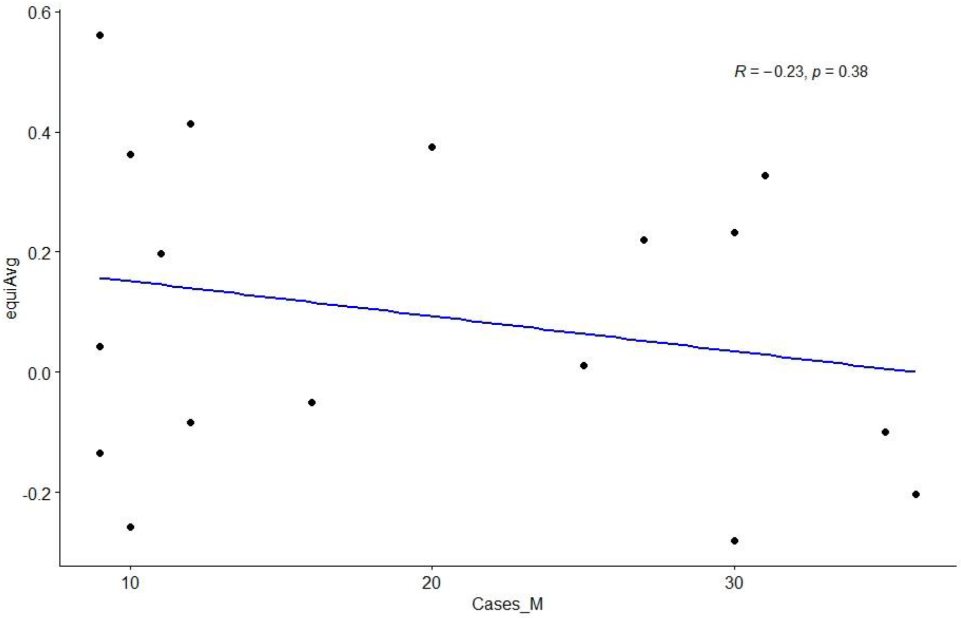
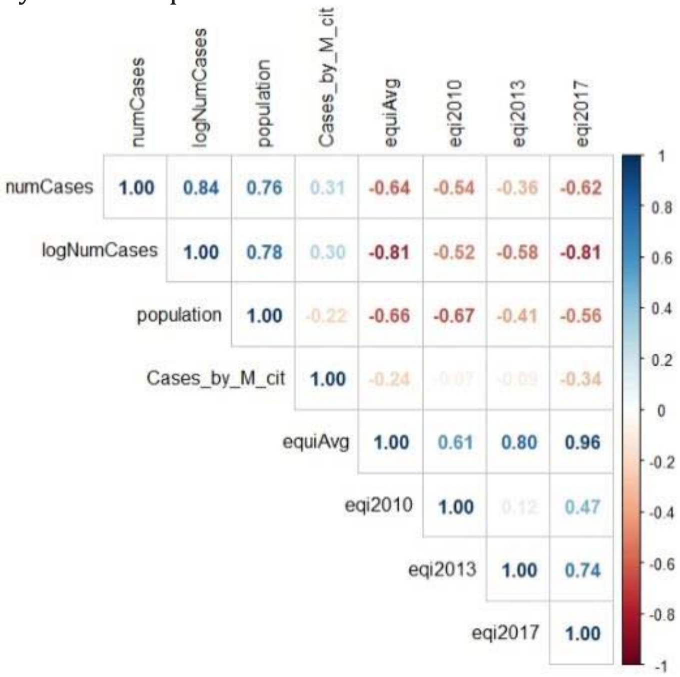


Figure 7: EQI number of corruption cases by million citizens correlation. Source: Own elaboration with EQI data.

Another possibility behind this lower correlation index would be that the population corrects the real correlation between variables, and the remaining pillars of the EQI, impartiality, and quality of the public services, accounts for the rest of the differences. Nevertheless, the presence of correlations supports the performance of our data mining model and the validity of the sample.



	numCases	logNumCases	population	Cases_by_M_cit	equiAvg	eqi2010	eqi2013	eqi2017
numCases	1.00	0.84	0.76	0.31	-0.64	-0.54	-0.36	-0.62
logNumCases		1.00	0.78	0.30	-0.81	-0.52	-0.58	-0.81
population			1.00	-0.22	-0.66	-0.67	-0.41	-0.56
Cases_by_M_cit				1.00	-0.24	-0.07	-0.09	-0.34
equiAvg					1.00	0.61	0.80	0.96
eqi2010						1.00	0.12	0.47
eqi2013							1.00	0.74
eqi2017								1.00

Figure 8: Correlation matrix from NUTS-2 level variables. Source: Own elaboration with EQI and INE data.

6. Corruption cases and local economic resilience

Spain stands above the EU averages both in corruption perception[10] and unemployment[11] levels. Our novel database supposes an opportunity to explore the effects of corruption on economic resilience at the local level and will intend as an example of the potential that similar constructs offer for new developments in institutional quality research.

Rios & Gianmoena (2020) situate the institutional quality as one of the most robust drivers of economic resilience at the regional NUTS-2 level. Behind this relation, further analysis suggests that QoG may increase policy responses, minimize barriers of entry, improve inefficient companies' replacement mechanism or reduce the likelihood for capital inflows stops among others.

The method for measuring economic resilience is by the change in the unemployment rates between 2007 and 2013[12]. We obtain this indicator at both province or NUTS-3 and municipal level, for the province indicator we compare the regional variation with the average change of the country. For the municipal level consequently, we compare the change in unemployment level with the average change of the province.

$$RES_i = \frac{\Delta E_i - \Delta E_{avg}}{\left|\Delta E_{avg}\right|}$$

$$\textit{where } RES_i \textit{ is the Resilence index; } \Delta E_i \textit{is the unemployment variation } 2007-2013;$$
$$\textit{and } \Delta E_{avg} \textit{is the verage unemployment variation of the above geographical level}$$

Prior to obtaining the municipal resilience indicator, we had to face the lack of official unemployment statistics at the local level. In order to sort this obstacle, we estimated the unemployment rates for each municipality and each year out from the data available; municipal total population, municipal total unemployment, and regional average active population, as follows:

$$UR_i = \frac{U_i}{N_i \cdot LF_j} \cdot 100$$

where UR_i *is unemployment rate of municipality* i; U_i *is unemployed people of municipality* i;
N_i *is population of municipality* i *and* LF_j *is the labour force average of the region* j,
to which belongs the municipality i.

When analyzing at the province level, the results do not present significant correlations for the total corruption cases or by population, neither with the resilience indexes nor with the unemployment levels.

[10] According to the 2017 QoG surveys almost all the Spanish NUTS-2 regions show results below the EU average with the only exceptions of Navarra and the Basque country, (Charron et al. 2019).

[11] For example, in May 2015 the Spanish unemployment rate was of 22,6%, more than double of the EU average of 11,1%. Source: EUROSTAT.

[12] The years 2007 and 2013 were selected by the fact that represent the minimum and maximum employment levels respectively, which difference captures the relative impact of the 2008 financial crisis.

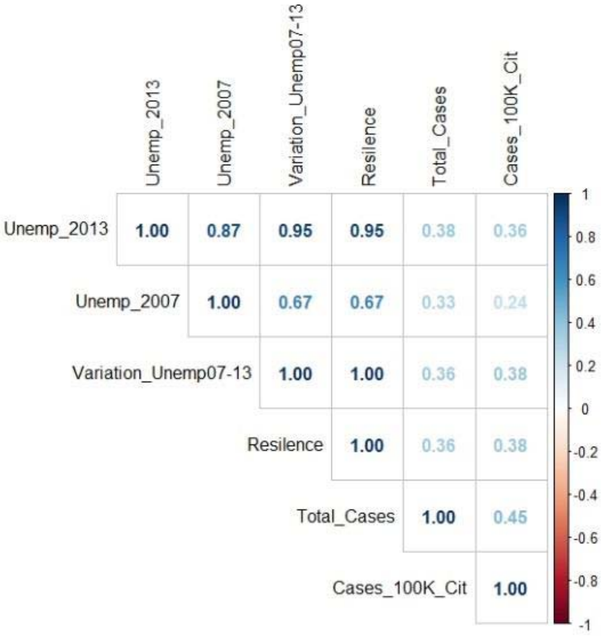


Figure 9: Correlation matrix from NUTS-3 level variables. Source: Own elaboration from INE and SS.

Similarly, if we test the correlation between the economic resilience at the municipal level, with the number of corruption cases, we do not find any significant relation.

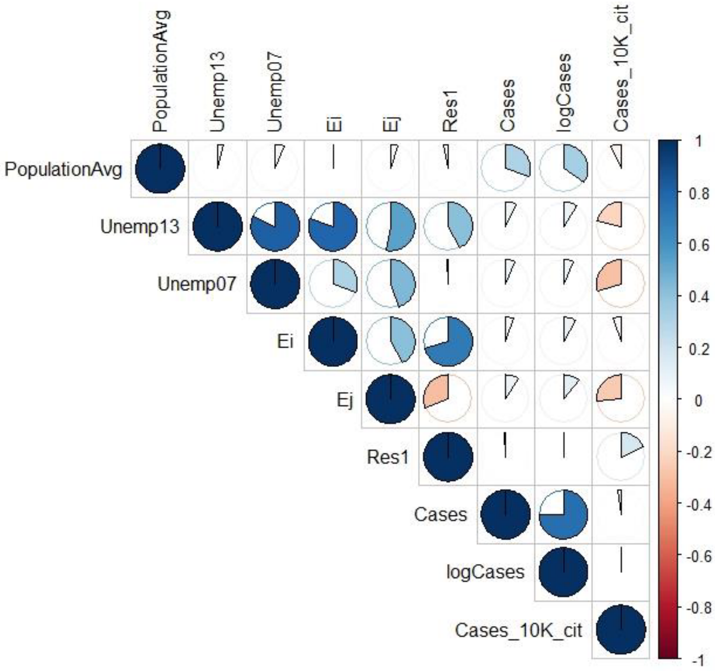


Figure 10: Correlation matrix from variables at the local level. Source: Own elaboration from INE and SS.

Several arguments can explain these results, the first one is that economic resilience is related to different aspects of the quality of government different from corruption, as the government effectiveness or the labor legal framework. Second, the economic resilience might be affected only by decisions at higher administrative levels, covered by NUTS-2 regions but not at the NUTS-3 or local level.

Another factor that could explain this lack of relation is that, as Rios & Gianmoena (2020) evidenced, additional factors related to knowledge and innovation reinforce the effects of the institutional framework on the economy. However, this is a preliminary approximation to the study of local corruption's effects on economic resilience and might be affected by the lack of unemployment statists.

7. Discussion

The potential of this method is greater than the examples presented, we did not include in our analysis the sentence of the case, and changing the code of the data mining scripts can adapt the extraction to other prosecution databases or tune different outputs.

This method is also extensible to different types of crimes, such as drug dealing, money laundering, tax evasion… and can contribute with high detail data at a small cost to perform all sorts of analysis or studies in the related fields.

The relation between these methods and the individual analysis of cases could be seen as a trade-off between the accuracy of the data and amplitude of the sample together with the reduction of the time and cost for the elaboration of the database. However, the accuracy problems can be reduced with the development and the sophistication of the data extraction and evaluation techniques, specifically with the usage of adapted NLP models.

More specifically, the data mining model developed presents limitations in the identification of several municipalities affected by the same case, residual mismatches of the municipalities, the geographical level of the corruption or even capturing cases that do not correspond to the definition of corruption. Additionally, the difference between the number of cases identified in the first step (2.585) of the extraction and the final sample (773), together with the number of urbanism corruption crimes (676) found by (Darias et al. 2012) strongly suggest that there are still a high number of corruption cases not captured by our model. All these problems are possible to overcome with the improvement of the data mining model.

Important to remark the difficulty to capture specifically the time horizon of the corruption phenomenon and the intensity or magnitude of the cases, a common problem with the perception and most of the traditional prosecution indicators. A problem that might be potentially addressed as well with the usage of efficient NLP techniques.

Still, the main limitation of the presented method, and all sorts of indicators built from prosecuted cases is that it bases on the performance of the judicial system. In highly corrupted states or with inefficient judicial systems, the prosecuted cases would not capture the reality of the corruption phenomenon and therefore be of low accuracy.

This method might be combined with other measures of corruption to obtain the strongest indicators. For example, a combination of the prosecuted cases, with perception or proxy indicators might cover a great part of the areas susceptible to corruption. Indeed, the combination with recent developments in other fields of local government quality indicators, such as government efficiency through spending (Balaguer-Coll et al. 2021) has the potential to offer multidimensional institutional quality indicators, similar to the EQI, with the advantage of using objective data.

8. Conclusions

The impact of the institutional quality in general, and corruption in particular, on the economic or welfare aspects has been largely proved at the national level, still, the studies at the regional level are recent and virtually non-existent when we check the municipal level.

In this paper, we focused on the indicators used to represent corruption in the literature, which we categorize into four groups: perception, audits, proxy, and prosecution. The high costs of the audits make their usage impossible at large scale studies but represent the most accurate indicators. Proxy indicators instead, can be applied to large territories and even in cross-country studies, even so, are strongly limited in their capacity to capture only certain types of corruption. Perception indexes stand as the most commonly used method due to their simplicity, but the usage of surveys has important biasness, and obtaining this information at the municipal level for a wide territory would have a high cost.

Prosecution data offer indicators with high accuracy and detail, this approach was traditionally limited by the fact that obtaining this type of data supposed a tedious task that required reading a high number of cases in the press or judicial databases. We propose the usage of data mining, with a model composed of data scrapping and natural language processing tools to reduce the cost of collection prosecution data, generating an objective indicator valid for countries with efficient judicial systems.

We have applied this model to a Spanish jurisprudence database, aiming at certain laws that capture different types of corruption. The result is a sample composed of 773 corruption cases related to municipal public servants that were prosecuted in the complete Spanish territory during the years 2000-2020.

This particular method includes geographical information until the municipal level and allows to differentiate between the main corruption types, which can be used for further development of studies about the effects of institutional quality at the lower administrative level. Our model presents important limitations, most of them shared with similar indicators, as the complexity to identify exactly the years where the corruption occurred and quantifying the magnitude or intensity of the cases. However, refining the model might overcome most of those limitations.

Finally, we found that the efforts in developing our sample are not sufficient to perform analyzes of the effects of corruption or governmental quality at the municipal level, as there are important difficulties to obtain statistical data at this level, particularly in our case local unemployment rates were not available, which we had to approximate.

The access to new big data and artificial intelligence tools makes it possible to analyze massive amounts of information, which allows amplifying the granularity of the data, both in content or detail and geographical. However, when deepening in the municipal level the information available is still very limited, constraining the research at this level.

As synthesized by Mullainathan& Spiess (2017), the arrival of new empirical tools has expanded the kind of problems we work on, and solving the data constraints is becoming a problem. Access to new data with higher detail should be accomplished in order to make valuable the available analyzes. Thus, emphasis on making municipal data available should be made to extract the complete potential of those techniques.

References.

Cavallo, A. (2013). Online and official price indexes: Measuring Argentina's inflation. *Journal of Monetary Economics*, 60(2), 152-165.

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

Balaguer-Coll, M.T., Narbón-Perpiñá, I., Peiró-Palomino, J., Tortosa-Ausina, E. (2021). Quality of government and growth at the municipal level: Evidence for Spain. *Journal of Regional Science*, forthcoming.

Henderson, V.; Storeygard, A.; Weil, D.N. (2012). Measuring Economic Growth from Outer Space. *The American Economic Review*, 102, (2), 994-1028. http://dx.doi=10.1257/aer.102.2.994

Razafindrakoto, M., & Roubaud, F. (2010). Are international databases on corruption reliable? A comparison of expert opinion surveys and household surveys in sub-Saharan *Africa. World development*, 38(8), 1057-1069.

Andvig, J. C., Fjeldstad, O. H., Amundsen, I., & Søreide, T. (2000). Research on Corruption A policy oriented survey.

Rohwer, A. (2009). Measuring corruption: a comparison between the transparency international's corruption perceptions index and the World Bank's worldwide governance indicators. CESifo DICE Report, 7(3), 42-52.

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

Hsieh, C. T., & Moretti, E. (2006). Did Iraq cheat the United Nations? Underpricing, bribes, and the oil for food program. *The Quarterly Journal of Economics*, *121*(4), 1211-1248.

Lopez-Valcarcel, B. G., Jiménez, J. L., & Perdiguero, J. (2017). Danger: local corruption is contagious!. *Journal of Policy Modeling*, *39*(5), 790-808.

Jiménez, J. L., & García, C. (2018). Does local public corruption generate partisan effects on polls?. *Crime, Law and Social Change*, *69*(1), 3-23.

Darias, L. M. J., Martín, V. O. M., & González, R. P. (2012). Aproximación a una geografía de la corrupción urbanística en España. *Ería*, (87), 5-18.

Colonnelli, E., & Prem, M. (2020). Corruption and firms. Available at SSRN 2931602.

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

Fazekas, M., Tóth, I.J. & King, L.P. (2016). An Objective Corruption Risk Index Using Public Procurement Data. *European Journal on Criminal Policy and Research* 22(3), 369–397.

Kaufmann, D., Kraay, A., & Mastruzzi, M. (2007). Measuring corruption: myths and realities.

Charron, N., & Lapuente, V. (2013). Why do some regions in Europe have a higher quality of government?. *The Journal of Politics*, 75(3), 567-582.

Charron, N., V. Lapuente & P. Annoni (2019). 'Measuring Quality of Government in EU Regions Across Space and Time.' *Papers in Regional Science*. DOI: 10.1111/pirs.12437

Charron, N., Dijkstra, L., & Lapuente, V. (2015). Mapping the regional divide in Europe: A measure for assessing quality of government in 206 European regions. *Social Indicators Research*, 122(2), 315-346.

Charron, Nicholas, Lewis Dijkstra & Victor Lapuente. 2014. Regional Governance Matters: Quality of Government within European Union Member States*,* *Regional Studies*, 48 (1): 68-90. DOI:10.1080/00343404.2013.770141

Hortas-Rico, M., & Rios, V. (2019). The drivers of local income inequality: a spatial Bayesian model-averaging approach. *Regional Studies*, 53(8), 1207-1220.

Gupta, S; Davoodi, H. and Alonso-Terme, R. (2002). Does corruption affect income inequality and poverty? *Economics of Governance*, 3(1), 23-45.

Dollar, D., & Kraay, A. (2003). Institutions, trade, and growth. *Journal of monetary economics*, 50(1), 133-162.

Buchanan, B. G., Le, Q. V., & Rishi, M. (2012). Foreign direct investment and institutional quality: Some empirical evidence. *International review of financial analysis*, 21, 81-89.

Welsch, H. (2004). Corruption, growth and environment: A cross-country analysis. *Environment and Development Economics*, 9(5), 663-693.

Rodríguez-Pose, A., & Garcilazo, E. (2015). Quality of government and the returns of investment: Examining the impact of cohesion expenditure in European regions. *Regional Studies*, 49(8), 1274-1290.

Rodríguez-Pose, A., & Di Cataldo, M. (2015). Quality of government and innovative performance in the regions of Europe. Journal of Economic Geography, 15(4), 673-706.

Hernández Díaz, C. (2017). Extracción de información de textos legales y notas de prensa. Unpublished manuscript.

Peiró-Palomino, J., Picazo-Tadeo, A. J., & Rios, V. (2020). Well-being in European regions: Does government quality matter?. Papers in Regional Science, 99(3), 555-582.

Peiró Palomino, J. (2019). Government quality and regional growth in the enlarged European Union: Components, evolution and spatial spillovers.

Rios, V., & Gianmoena, L. (2020). The link between quality of government and regional resilience in Europe. *Journal of Policy Modeling*, 42(5), 1064-1084.

Acemoglu, D., & Johnson, S. (2005). Unbundling institutions. *Journal of political Economy*, 113(5), 949-995.

Rodríguez-Pose, A., & Zhang, M. (2019). Government institutions and the dynamics of urban growth in China. *Journal of Regional Science*, 59(4), 633-668.

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

Knack, S. F. (2006). Measuring corruption in Eastern Europe and Central Asia: A critique of the cross-country indicators (Vol. 3968). World Bank Publications.

Espinosa Villar, M. (2019). Legal Data mining: análisis y predicción de sentencias judiciales. Unpublished manuscript.

Ezcurra, R., & Rios, V. (2019). Quality of government and regional resilience in the European Union. Evidence from the Great Recession. *Papers in Regional Science*, 98(3), 1267–1290.

OECD. (2017). *Government at a Glance. 2017*. Paris: OECD.

Appendix A.

Table A1: prosecuted corruption cases 2000-2020 by province.

Region (NUTS-2)	Number of cases	Cases by 100K citizens	Fraud	Embezzlement	Influence Trafic	Bribery	Urbanism
Málaga	88	5,94	29	40	4	16	40
Barcelona	54	1,03	4	19	15	8	13
Madrid	54	0,90	4	16	11	17	13
Palmas, Las	48	4,73	6	14	9	13	24
Granada	40	4,63	1	11	6	4	26
Balears, Illes	35	3,47	12	12	5	4	15
Almería	29	4,65	2	7	1	2	20
Zaragoza	26	2,86	8	18	6	6	1
Cádiz	25	2,11	12	13	1	3	3
Valencia/València	25	1,04	3	12	5	3	8
Santa Cruz de Tenerife	23	2,47	2	17	3	0	5
Sevilla	21	1,14	7	8	0	6	4
Cáceres	17	4,25	6	8	4	0	5
Badajoz	15	2,25	3	11	2	0	2
Coruña, A	15	1,35	1	10	4	0	0
Jaén	14	2,19	2	6	1	1	5
Murcia	14	1,05	4	5	1	4	2
Asturias	13	1,24	3	5	0	3	3
Valladolid	12	2,36	2	4	2	5	2
Tarragona	12	1,70	0	11	0	1	0
Alicante/Alacant	12	0,72	1	4	1	5	3
Ciudad Real	11	2,26	1	4	1	0	5
Cantabria	11	1,97	0	5	0	2	5
Córdoba	11	1,42	2	8	1	0	1
Zamora	10	5,35	1	7	0	0	2
Ávila	9	5,58	1	9	0	0	0
Rioja, La	9	3,08	0	4	2	5	0
Toledo	9	1,46	1	7	0	1	1
Gipuzkoa	9	1,28	0	5	1	3	0
Ourense	8	2,45	0	4	1	1	3
Burgos	8	2,27	2	5	1	2	2
Huelva	8	1,63	1	6	0	2	0
Salamanca	7	2,06	2	3	0	1	1
León	7	1,46	0	2	0	0	5
Navarra	6	1,00	2	5	0	0	0
Segovia	5	3,33	0	4	0	0	1
Cuenca	5	2,52	0	1	2	1	2
Araba/Álava	5	1,61	0	2	0	1	2
Pontevedra	5	0,54	2	2	0	0	2
Bizkaia	5	0,44	2	1	0	0	2
Teruel	4	2,96	0	2	1	0	1
Palencia	4	2,36	1	3	0	0	0
Huesca	4	1,87	2	4	0	0	0
Lugo	4	1,15	1	2	1	0	1
Albacete	4	1,06	0	4	0	0	0
Castellón/Castelló	4	0,75	1	2	1	1	1
Girona	4	0,59	0	2	0	2	0
Lleida	3	0,75	0	2	0	0	1
Guadalajara	2	0,94	0	1	0	1	0
Total	**773**	**106,23**	**134**	**357**	**93**	**124**	**232**

Table A2: accuracy measure with aleatory subsample.

Case_ID	Municipality_Identified	Corruption_case(1)	Higher_than_local_level(2)	Correct_Municipality(3)	More_municipalities(4)
4356184	Formentera	YES	YES	YES	YES
6920273	Baena	YES	NO	YES	NO
3723900	Valverde	YES	NO	YES	NO
6252961	Coria	YES	NO	YES	NO
2659452	Sóller	YES	NO	YES	NO
4837270	Plasencia	YES	NO	YES	NO
5684071	Roda	YES	NO	YES	NO
2248017	Monachil	YES	NO	YES	NO
6920352	Santaliestra	YES	NO	YES	NO
7449034	Cambrils	YES	NO	YES	NO
6003657	San Bartolomé	YES	NO	YES	NO
7976182	Torremontalbo	YES	NO	YES	NO
448609	Corvera	YES	NO	YES	NO
7884243	San Bartolomé	YES	NO	YES	NO
272176	Badajoz	YES	YES	YES	NO
2717400	Cabrales	YES	NO	YES	NO
8213301	Valverde	YES	NO	NO	NO
2020541	Cáceres	YES	NO	YES	NO
6956890	Sabadell	YES	NO	YES	NO
5914080	Artà	YES	NO	YES	NO
6512579	Casares	YES	NO	YES	NO
7513343	Lousame	YES	NO	YES	NO
6161992	Cartagena	NO	NULL	NULL	NULL
6939242	Trasmoz	YES	NO	YES	NO
6526543	Arona	YES	NO	YES	NO
174353	Villaveza	YES	NO	YES	NO
7664297	Castro Urdiales	YES	NO	YES	NO
2542574	La Línea de la Concepción	YES	NO	YES	NO
103303	Badajoz	YES	NO	YES	NO
6590192	Oria	YES	NO	YES	NO
509829	San Bartolomé	YES	NO	YES	NO
2207648	Barcelona	YES	NO	YES	NO
2248017	Monachil	YES	NO	YES	NO
5500645	Gondomar	YES	NO	YES	NO
6919536	Yaiza	YES	NO	YES	NO
5345116	La Línea de la Concepción	YES	NO	YES	NO
963696	Marbella	YES	NO	YES	NO
7141207	Santa Cruz	YES	NO	YES	NO
6028248	Jerez de la frontera	YES	NO	YES	NO
6108618	Jerez de la frontera	YES	NO	YES	NO
7920282	Mahide	YES	NO	YES	NO
7270035	Santa Cruz de Tenerife	YES	NO	YES	NO
3867192	Cantoria	YES	NO	YES	NO
1292783	Sanlúcar	YES	NO	YES	YES
6161992	Cartagena	NO	NULL	NULL	NULL
1770653	Madrid	YES	NO	YES	NO
5563733	Telde	YES	NO	YES	NO
8279677	Villaconejos	YES	NO	YES	NO
5701084	Viveiro	YES	NO	YES	NO
5898052	Humanes	YES	NO	YES	NO
4129186	Andratx	YES	NO	YES	NO
4840018	Valderas	YES	NO	YES	NO
2516555	Aranjuez	YES	NO	YES	NO
8091432	Santa Brígida	YES	NO	YES	NO
6110656	Oria	YES	NO	YES	NO
7730030	Cáceres	NO	NULL	NULL	NULL
7725941	La Haba	YES	NO	YES	NO
3758351	Nulles	YES	NO	YES	NO
7912065	Puerto	YES	NO	YES	NO
7902699	Castellón	YES	NO	YES	NO
6668323	Serranillos	YES	NO	YES	NO
5820012	MIRANDA	YES	NO	YES	NO
7515708	Lousame	YES	NO	YES	NO

Appendix B.

Script B1

```
#LYBRARIES
from selenium import webdriver
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions
from selenium.webdriver.common.by import By
from bs4 import BeautifulSoup

from lwlist import CorruptionLawsList as lawList #List of laws to extract cases

import time
import pandas as pd
import re
import csv

#Extracts str between two texts
def extractStr(org, ini, fin):
    plus = int(len(ini))
    try:
        ini_tt = org.index(ini)
        fn_tt = org.index(fin)
        dest = org[(ini_tt + plus):(fn_tt)]
    except:
        dest = 'ERROR'
    return dest

chrome_options=webdriver.ChromeOptions()
driver=webdriver.Chrome(executable_path='./chromedriver.exe',chrome_options=chrome_options)

csvFile = open("2prev.csv", "w")
writer = csv.writer(csvFile)

tableList = []
loopCount = 0
for law in lawList:
    driver.get('https://analytics.tirant.com/analytics/login.do?&user=UVAL&password=XXXXXX')
#Login
    driver.get('https://analytics.tirant.com/analytics/busquedaJurisprudencia/index') #Load
webpage

    #Close cookies advert, only in the first loop
    if loopCount == 0:
        content = driver.find_element_by_id('cookie_accept_button').click()

    #Select "PENAL"
    content = driver.find_element_by_xpath('//*[@id="boxJuris"]/label[2]').click()

    #Select resolution tipe = SENTENCIA
    content = driver.find_element_by_id('tiporesolucion').send_keys("Sentencia")

    #Select date range
    content = driver.find_element_by_xpath('//*[@id="fecha2"]').send_keys("31/12/2020") #End
date
    content = driver.find_element_by_xpath('//*[@id="fecha1"]').send_keys("01/01/2000") #Initial
date

    #Select Law
    content =
driver.find_element_by_xpath('/html/body/div[2]/div[2]/div[2]/form/div[3]/div/div[8]/div/span/i'
).click()
    content = driver.find_element_by_xpath('//*[@id="titulonorma"]').send_keys(law) ##law
looped##
    content = driver.find_element_by_xpath('//*[@id="buscarnorma"]').click()
    time.sleep(3)
    content =
driver.find_element_by_xpath('/html/body/div[8]/div[2]/div[2]/div[2]/ul/li/a/ins[1]').click()
    content =
driver.find_element_by_xpath('/html/body/div[8]/div[11]/div/button[1]/span').click()

    #Launch search
    content = driver.find_element_by_xpath('//*[@id="buscar"]').click()

    #Obtain number of results
    numResultsFIRST = driver.find_element_by_class_name('masDatosBusq')
    numResultsHtml = numResultsFIRST.get_attribute('innerHTML')
    soupResultList=str(BeautifulSoup(numResultsHtml,'html.parser'))
    numResultsList = re.findall(r'\d+', soupResultList)
    numResults = int(numResultsList[0]) + 1

    for i in range(0,numResults): #Scrolls needed to load the full list, depending on the number
of results
        driver.execute_script('window.scrollBy(0, 900)')
        time.sleep(0.5)

    itemList = driver.find_elements_by_class_name('docItemList')

    csvRowList = []
    for item in itemList:
        try:
            itemList_html = item.get_attribute('innerHTML')
            soupItemList=BeautifulSoup(itemList_html,'html.parser')

            csvRow = []

            for enlace in soupItemList.find_all('a',href=True):
                id = extractStr(str(enlace), '/show/', "?")
                csvRow.append(id)

            count = 0 #count for detecting the position of the court data
            for span_tag in soupItemList.find_all('span'):
                if (span_tag.text != ' | ' or count == 2): #after second bar we find the court
                    tempText = str(span_tag.next_sibling.rstrip(' '))
                    a = re.sub('\s+',' ', tempText)
                    csvRow.append(a)
                    if count == 2:
                        count = count + 1
                else:
                    count = count + 1

        except:
            csvRow.append('ERROR')


        csvRow.append(law)
        csvRowList.append(csvRow) #Stores each row generated in the setences list
        tableList.append(csvRow)


    loopCount = loopCount + 1
    print(loopCount/len(lawList)*100, '%')

writer.writerows(tableList) #Transfers the list of sentences to a CSV file
csvFile.close()
```

Script B2

```
#LYBRARIES
from selenium import webdriver
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions
from selenium.webdriver.common.by import By
from bs4 import BeautifulSoup
import urllib3

import time
import pandas as pd
import re
import csv

import spacy
import collections
from spacy.matcher import Matcher
from spacy.tokens import Span

from geopy.geocoders import Nominatim

from CajadeHerramientas import getGeoLoc
from CajadeHerramientas import sorterOfLists
from IDlist import listId1 as ids               #List of case IDs to crawl

##### SET SPACY #####

nlp = spacy.load("es_core_news_sm", exclude=["ner", "lemmatizer"])
matcher = Matcher(nlp.vocab)

pattern1 = [{'LOWER': 'ayuntamiento'},
           {'LOWER': 'de'},
           {'IS_LOWER': False}]

pattern2 = [{'LOWER': 'alcalde'},
           {'LOWER': 'de'},
           {'IS_LOWER': False}]

matcher.add('ayto', [pattern1] )
matcher.add('acde', [pattern2] )

######################

def getMunicipality(data): #Obtains municipality name from text

    doc = nlp(data)
    matches = matcher(doc)

    matchesList = []
    sortedList = []
    for match_id, start, end in matches:
        span = Span(doc, start, end, label=match_id)

        #temp = (doc[start:end].text)
        tokens = [token.text for token in doc[start:end]]

        if (tokens[2] == 'La') or (tokens[2] == 'El') or (tokens[2] == 'San') or (tokens[2] ==
'Santa') or (tokens[2] == 'Sant') or (tokens[2] == 'Las') or (tokens[2] == 'Puerto'):
            temp_municipality = (str(str(tokens[2]) + ' ' + str(doc[span.start + 3])))

        else:
            temp_municipality = tokens[2]

        matchesList.append(temp_municipality)

    sortedList = sorterOfLists(matchesList)
    sortedList.append('')
    return(sortedList[0])

csvFile = open("IDsPart1a,mbk.csv", "w")
writer = csv.writer(csvFile)

urllib3.disable_warnings()
http = urllib3.PoolManager()

r=http.request('GET','https://analytics.tirant.com/analytics/login.do?&user=UVAL&password=XXXXXX
',redirect=False)
headers={'cookie': r.getheader('set-cookie')}

tableList = []
loopCount = 0

for id in ids: #crawls list of IDs
    csvRow = []
    municipality = ''
    geoLoc = ''

    try:
        id_web = id.replace('\n','')

        r = http.request('GET',"http://www.tirantonline.com/tol/documento/show/"
+str(id_web),headers=headers)

        html = r.data
        soup = BeautifulSoup(html,'html.parser')

        content = soup.find_all("span", id=re.compile('^an-')) #Obtains list of text from the
case summary

        results = []
        for item in content:
            parrafo = (item['rel'])

            texto = soup.find("div",
id=parrafo).text.replace('[siguiente]','').replace('[anterior]','').replace('[Contextualizar]','
').replace('\n',' ') #Cleans texts before tokenize
            municipality_prf = getMunicipality(texto) #Extracts municipality name of the case

            if municipality_prf != '':
                results.append(municipality_prf)

        if results == []:
            municipality = 0
            geoLoc = 0
        else:
            municipality = (sorterOfLists(results))[0]
            geoLoc = (getGeoLoc(municipality))          #Extracts coordinates from text

            csvRow.append(id)
            csvRow.append(municipality)
            csvRow.append(geoLoc)

    except:
        csvRow.append(id)
        csvRow.append('ERROR')

    tableList.append(csvRow)

    loopCount = loopCount + 1
    progress = loopCount/len(ids)*100
    print("%.2f" % progress, '%')

writer.writerows(tableList) #Transfers the list of sentences to a CSV file
csvFile.close()
```