Soc Indic Res (2018) 140:975–998
https://doi.org/10.1007/s11205-017-1802-2

Predicting Public Corruption with Neural Networks: An Analysis of Spanish Provinces

Félix J. López-Iturriaga[1,2] · Iván Pastor Sanz[1]

Accepted: 19 November 2017 / Published online: 22 November 2017

Abstract We contend that corruption must be detected as soon as possible so that corrective and preventive measures may be taken. Thus, we develop an early warning system based on a neural network approach, specifically self-organizing maps, to predict public corruption based on economic and political factors. Unlike previous research, which is based on the perception of corruption, we use data on actual cases of corruption. We apply the model to Spanish provinces in which actual cases of corruption were reported by the media or went to court between 2000 and 2012. We find that the taxation of real estate, economic growth, the increase in real estate prices, the growing number of deposit institutions and non-financial firms, and the same political party remaining in power for long periods seem to induce public corruption. Our model provides different profiles of corruption risk depending on the economic conditions of a region conditional on the timing of the prediction. Our model also provides different time frameworks to predict corruption up to 3 years before cases are detected.

Keywords Corruption · Prediction · Early warning system · Neural networks · Self-organizing maps

JEL Classification C45 · D73

✉ Félix J. López-Iturriaga
flopez@eco.uva.es

Iván Pastor Sanz
ivan.pastor@alumnos.uva.es

[1] School of Business and Economics, University of Valladolid, Avda. Valle del Esgueva 6, 47011 Valladolid, Spain

[2] Higher School of Economics, Moscow, Russia

1 Introduction

Although political corruption has been around for a long time, it has attracted considerable attention in recent years, and the literature suggests it is on the increase (Salinas-Jiménez and Salinas-Jiménez 2007; Transparency International 2016). For example, Kaufmann and Bellver (2005) estimate that corruption represented some $1.1 trillion globally. According to a recent estimate from the International Monetary Fund (2016), the annual cost of bribery comes to about $1.5 to $2 trillion (roughly two percent of global GDP). Corruption can have a dramatically negative impact on a country's economic development, which may then in turn lead to more corruption. Spain is a good example of this vicious circle. Between 2007 and 2012, the financial wealth of Spanish households fell by €167 billion, the unemployment rate shot up from 8.8 to 26.2%, and the 2% public surplus turned into a 10.6% public deficit. Furthermore, the risk premium on Spanish treasury bonds reached a worrying 610 point peak in the summer of 2012. At the same time that tough measures to reduce government expenditure and public deficit were enforced, a number of political corruption cases were unearthed, causing alarm all over the country.[1] Moreover, the European Union Anti-Corruption Report issued by the European Commission (2014) highlighted serious concerns about the growth of corruption in certain countries including Spain. Moreover, 95% of Spanish citizens agreed that corruption is rife throughout the country.[2] Spain thus provides a unique framework to study the issue.

The aim of this paper is to provide a neural network prediction model of corruption based on economic factors. We contend that corruption must be detected as soon as possible in order to take corrective and preventive measures. Because public resources for combating corruption are limited, efforts should focus on areas most likely to be involved in corruption cases. We use a unique database that brings together the main cases of political corruption in Spain. We then propose an early warning corruption model to predict whether corruption cases are likely to emerge in Spanish regions given certain macroeconomic and political determinants. We use self-organizing maps (SOMs), a neural network approach, to predict corruption cases in different time horizons. Our model provides different profiles of corruption risk depending on the economic conditions of a region conditional on the timing of the prediction.

This paper contributes to the literature by developing a novel approach with three differential characteristics. First, unlike previous research, which is mainly based on the perception of corruption, we use data on actual cases of corruption. Second, we use the neural network approach, a particularly suitable method since it does not make assumptions about data distribution. Neural networks are quite powerful and flexible modeling devices that do not make restrictive assumptions on the data-generating process or the statistical laws concerning the relevant variables. Third, we report the probability of corruption cases on different time scenarios, so that anti-corruption measures can be tailored depending on the immediacy of such corrupt practices. Consistent with Huysmans et al. (2006), who also use SOMs and support vector machines to forecast changes in the perceived level of corruption, our model allows patterns of corruption to be identified on different time horizons.

[1] A December 2014 survey by the Spanish Center for Sociological Research showed that 63.9% of Spanish citizens cited corruption as the country's major problem.

[2] http://ec.europa.eu/dgs/home-affairs/e-library/documents/policies/organized-crime-and-human-trafficking/corruption/docs/acr_2014_en.pdf.

Our results show that economic factors prove to be relevant predictors of corruption. We find that the taxation of real estate, economic growth, increased house prices, and the growing number of deposit institutions and non-financial firms may induce public corruption. We also find that the same ruling party remaining in power too long is positively related to public corruption. Depending on the characteristics of each region, the probability of corrupt cases emerging over a period of up 3 years can be estimated. We then detect different patterns of corruption antecedents. Whereas in some cases, corruption cases can be predicted well before they occur and thus allow preventive measures to be implemented, in other cases the prediction period is much shorter and urgent corrective political measures are required.

The remainder of this paper is organized as follows. Sections 2 and 3 review the literature on corruption and the foundations of SOMs, respectively. Section 4 explains the empirical characteristics of our early warning system. Section 5 presents the results of our model. In this section is also discussed the implications of our results. Finally, Sect. 6 concludes.

2 Theoretical Background

2.1 The Literature on Corruption

Prior literature reports a widespread consensus that corruption has detrimental effects on the economy (Ortega et al. 2016). Mauro (1998) reports the negative effect of the perception of corruption on investment and GDP growth for a sample of 106 countries. Salinas-Jiménez and Salinas-Jiménez (2007) find a negative relation between corruption, productivity, and economic growth for 22 OECD countries. Transparency International (2009) suggests that the most developed countries have lower levels of corruption. Using a set of micro-data from 67 countries, Pieroni and d'Agostino (2013) show that economic freedom helps to reduce corruption. In the same line, Rajkumar and Swaroop (2008) find that some public expenditure policies perform most poorly in places with high corruption. Cavoli and Wilson (2015) show that corruption imposes an inflationary bias on the optimal monetary policy, and Kunieda et al. (2014) provide evidence that the negative effect of government corruption on economic growth is channeled through higher tax rates and is amplified by capital account constraints. Alternatively, Saha and Gounder (2013) propose a quadratic relation between corruption and economic development. D'Agostino et al. (2016) report that the interaction between corruption and public spending has a strong negative impact on economic growth. Corruption, in general, threatens government legitimacy and economic freedom, leads to regressive taxes, and increases poverty (Nwabuzor 2005).

When developing a model to predict corruption, the causes are as relevant as its consequences (Dong and Torgler 2013; Kong and Volkema 2016). Among the various explanations (i.e., political, historical, social, cultural), economic theory shows that corruption is fuelled when the monetary benefits outweigh the associated penalties. Based on a comprehensive review of the economic determinants of corruption, Aidt (2009) argues that corruption depends on three issues: deterrence measures, bureaucratic discretionary power, and the possibility of generating economic rents. The increased likelihood of being caught coupled with the severity of the penalties reduce the probability of corruption occurring. Similarly, by its enabling corruption to be controlled, freedom of the press is associated with increased real GDP per capita (Ambrey et al. 2016). Education also acts as an important deterrent. Given the link between education and national income, corruption should be lower in richer countries (Treisman 2000).

In addition, the economic rents to be gained from corruption can also act as an incentive. Van Rijckeghem and Weder (2001) show that corruption decreases when official wages increase. If wages are low, the opportunity cost of bureaucrats' accepting bribes decreases. Corruption also occurs more frequently in developing countries due to the lack of commensurate advancements in their legal, political, and social institutions (Kaymak and Bektas 2015). A country's industrial organization also influences corruption levels; specifically, countries with less internal and external competition are more prone to corruption (Gerring and Thacker 2005).

Empirical research is faced with the problem of how to measure corruption (Olken 2009). Thus, prior research often uses surveys of corruption perceptions. For example, Clausen et al. (2011) analyze the relationship between corruption and confidence in public institutions using the Gallup World Poll database. They find that in countries where respondents report a high incidence of personal experiences with corruption, and in which corruption is perceived to be widespread, confidence in public institutions is also low. Pellegata and Memoli (2016) also confirm that corruption negatively affects citizens' confidence among European Union member states. Li et al. (2016) study the factors that explain the variation in people's perceptions of anti-corruption efficacy. Finally, Zheng et al. (2017) found a negative effect of corruption perception on political participation. Other studies measure corruption in a variety of ways, including surveys of bribes that question possible bribe-payers and compare the estimated bribe with the reported costs of public goods, structural equations models, the analysis of noncompliance by public officials as compared to noncompliance within the general population, and the number of crimes against public administration officials (Del Monte and Papagni 2007; Olken 2007; Neiva de Figueiredo 2013).

Nevertheless, as Treisman (2007) admits, perception-based data reflect impressions of corruption intensity rather than actual occurrences of corruption. These perceptions are subjective and can be influenced by respondents' beliefs as well as their social and economic conditions. Similarly, uncorrected measures of the perception of corruption might cause misleading conclusions to be drawn about the comparisons of corruption levels between countries (León et al. 2013). Nevertheless, among the literature, few studies use real data on corruption. Objective data on corruption are difficult and complex to obtain, since crimes are committed in a hidden manner. If available, information is usually found in an unstructured way in the media, court records, etc. Some examples of this research are Dong and Torgler (2013) and Wu and Zhu (2011), who pinpoint the causes of corruption in China, and Stockemer and Calca (2013), who use municipal corruption cases in Portugal and find that highly corrupt areas have a higher turnout in elections than less corrupt areas.

2.2 The Political Framework of Corruption

Interestingly, prior research finds a link between corruption and political decentralization (Fisman and Gatti 2002; Ivanyna and Shah 2011). Nevertheless, these studies, which use subjective indexes of perceived corruption and mostly fiscal indicators of decentralization, report conflicting conclusions. When focusing on political decentralization, most agree that the federal structure is associated with higher perceived corruption (Fisman and Gatti 2002). Diaby and Sylwester (2014) find that in the former communist countries, bribes are higher under a more decentralized bureaucratic structure. In the same vein, Treisman (2002) finds that a larger number of administrative or governmental tiers correlate with higher perceived corruption. Fan et al. (2009) come to a similar conclusion using information on reported bribery. Albornoz and Cabrales (2013) argue that the effect of decentralization on corruption is conditional on the level of political competition. Decentralization is associated with lower levels of corruption if the level of political competition is sufficiently high.

Overall, the results suggest the danger of uncoordinated rent-seeking as government structures become more complex. A greater number of relationships and interactions between public officials and private agents in federal or decentralized states seem to provide increased opportunities for corrupt behavior. In any case, most studies that consider the causes of corruption focus on cross-country comparisons, and only a few employ within-country data (Fisman and Gatti 2002; Leeson and Sobel 2008).

Corruption in Spain has attracted considerable attention, particularly in recent years, when many cases have been unearthed. Spain is a diverse country made up of different regions with varying economic and social structures, as well as different languages and historical, political and cultural traditions. Although Spain is not a federal state, it is a highly decentralized unitary state which endows its regions or autonomous communities (*Comunidades Autónomas*) with high levels of political and economic competences. Autonomous communities is the nomenclature of Territorial Units for Statistics 2 (NUTS-2) regions according to the European Commission classification,[3] with high levels of political and economic competences. The worrying macroeconomic imbalances of the autonomous communities in terms of excessive public deficit, public debt, and sovereign debt problems led the Spanish government to enact a stability program for 2011–2014 to accelerate fiscal consolidation which focused on the autonomous communities. Thus, the political structure of Spain seems to be related to some of this recently growing corruption.

In turn, this analysis of corruption problems among Spanish regional governments is quite timely as is the design of a model of corruption prediction based on macro-economic factors. Our paper contributes to the literature on corruption by using real data and by adopting a different approach which it is hoped will prove useful vis-à-vis understanding the complex process of corruption. Moreover, whereas other studies mainly determine the causes or the specific sign of certain variables related to corruption or predict crimes against the administration, our approach provides an estimation of the probability of new cases of corruption emerging for different time horizons.

3 SOMs

3.1 Unsupervised Self-Organizing Maps

The literature is scarce on corruption from the point of view of data mining techniques. Prior research uses data mining techniques and, specifically, neural networks to predict patterns in some related fields such as crime (Li and Juhola 2014, 2015), credit risk evaluation (Guo et al. 2016; Swiderski et al. 2012), fraud detection (Olszewski 2014), and wellbeing (Carboni and Russu 2015; Rende and Donduran 2013; Lucchini and Assi 2013). We argue that neural networks can be also applied to predict corruption.

SOMs are a kind of artificial neural network that aim to mimic brain functions so as to provide machine learning and pattern recognition (Jagric et al. 2015; Kohonen 1982).

[3] The Nomenclature of Territorial Units for Statistics classification is a hierarchical system for dividing up the economic territory of the European Union.

SOMs have the ability to extract patterns from large data sets without an explicit understanding of the underlying relationships. They convert nonlinear relations among high dimensional data into simple geometric connections among their image points on a low-dimensional display. The most important topological and metrical relations are preserved, as data points with similar properties are placed close to each other within the output (Kohonen 2001). These properties have made SOMs a useful tool to detect patterns and obtain visual representations of large amounts of data. Consequently, predicting corruption is a field in which SOMs can become a powerful tool.

Figure 1 shows the most common version of SOMs. The input layer of neurons represents the original data set and is connected to the output layer of neurons through synaptic weights. The information provided by each neuron of the input layer is transmitted to all the neurons of the output layer. Thus, each neuron in the output layer receives the same set of input layer information. The first and most commonly used version of SOMs is considered an unsupervised network because no objective output occurs. Neurons learn in an unsupervised way to detect and identify data patterns in specific zones in a two-dimensional grid. SOMs are trained by means of an iterative process.

Nour (1994) sums up an SOM learning algorithm in three stages. First, the vector of initial weights $W_i(t)$ in $t = 0$ is set randomly. At this moment, the maximum number or possible iterations in the training phase of the network (T) is defined. Second, an input vector X is added to the network, and the distance (similarity) D is computed using the Euclidean metric to find the closest matching unit c to each input vector as follows:

$$d_{i,j,(t)} = \sqrt{\sum_{h=1}^{k} \left(W_{i,j,h} - X_k\right)^2}.$$

Finally, the weight vector is updated according to the following rule:

$$W_{jik}(t+1) = W_{jik}(t) + \alpha.\left[X_k(t) - W_{jik}(t)\right],$$

where α is the learning ratio, $X_k(t)$ is the input pattern in t and W_{jik} is the synaptic weight that connects the k input with the (j, i) neuron in t. Not only is the winning neuron updated, but also the neighbors following a neighborhood function. The neighborhood ratio

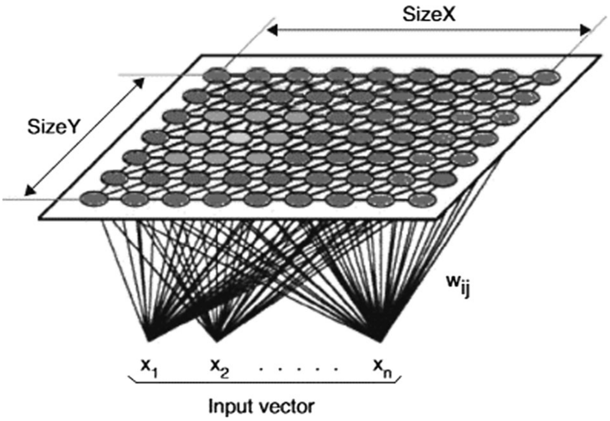

Fig. 1 SOM graphical representation

decreases with the number of iterations of the model so as to achieve a better specialization of each neuron. The process continues an iterative way until *t* reaches the maximum number of iterations T, and then jumps back to step 2. Following Gladyshev's theorem, SOM models almost always reach convergence (Lo and Bavarian 1993).

3.2 Supervised SOMs

Although SOMs are often unsupervised, supervised SOMs can be created by incorporating the desired output in the training phase. The aim is to produce input vectors by linking the (numeric) target vectors with the data label and then training the network in the usual way. Prior research shows that supervised versions of SOMs outperform unsupervised SOMs in predicting problems (Hagenbuchner et al. 2001). In a supervised SOM, two maps can be trained at the same time—one containing the original data (X) without the class information and the other including only the class or target information (Y). A common winning unit for both maps can be determined by calculating a weighted combination of the similarities between the X object and all the units in the X map and the similarities between the corresponding output Y object and all the units in the Y map. A specific weight of X and Y spaces in this combination must be chosen. The training is done in the usual manner and is updated with the information from the winning unit and its neighborhood. As in unsupervised SOMs, the learning rate and size of the neighborhood decrease in each iteration.

This study trains three supervised versions of SOMs. The three models use macroeconomic variables and the region's corruption status (corrupt or not corrupt) one, two and 3 years ahead of the reference point, respectively. Each Spanish region is then classified for each time horizon using the three trained models (i.e., each region is classified three times), which predict the likelihood that a region will have an incident of corruption up to 3 years later. The position of each region in the trained maps is now transformed into a probability of corruption. This probability is then introduced as the last input that renders the early warning corruption system map. This last map is trained using an unsupervised SOM (because no information exists about the possible output). This final map, which classifies Spanish regions according to their corruption risk profile, provides a visual representation of corruption. Figure 2 illustrates the three supervised versions of SOM and the final unsupervised map.

4 Early Warning Corruption Systems

The information we use comes from the corruption database gathered by *El Mundo*,[4] one of the most influential newspapers in Spain. The database contains information about the criminal cases involving a politician or a public official reported in Spain since 2000. The accused can be either already sentenced or awaiting the verdict. Our unit of analysis is the Spanish provinces, which are considered NUTS-2 regions according to the European Commission classification. Spain consists of 52 provinces. A province is considered susceptible to corruption if at least one corruption case occurs in a given year. The dependent variable for each year and province is defined as a dummy variable that equals 1 if at least one corruption case has gone to trial in this province in a given year, and zero otherwise.

[4] http://www.elmundo.es/grafico/espana/2014/11/03/5453d2e6268e3e8d7f8b456c.html.

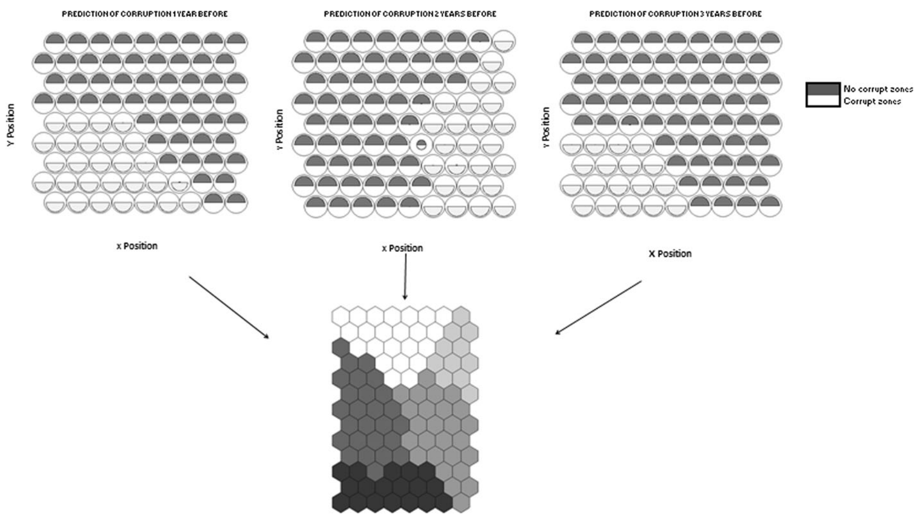
PREDICTION OF CORRUPTION 1 YEAR BEFORE
PREDICTION OF CORRUPTION 2 YEARS BEFORE
PREDICTION OF CORRUPTION 3 YEARS BEFORE
No corrupt zones
Corrupt zones
Y Position
Y Position
Y Position
x Position
x Position
X Position

Fig. 2 Visual representation of our model

The number of corruption cases can be correlated with the population or the economic activity of a given region. Larger regions could be more prone to present more corruption cases because the number of politicians is also higher. Thus, we measure public corruption as the number of registered cases of corruption per 100,000 people each year. This metric was also used by Dong and Torgler (2013) when studying causes of corruption in China.

Although the date on which a corruption case is uncovered can easily be defined, the year in which a bribe occurs proves more difficult to detect. A public official may have been receiving bribes long before he or she was finally charged. Therefore, we use three supervised versions of SOMs to predict corruption one, two, and 3 years before cases were uncovered. The output of each model is a two-dimensional map with two regions: corruption and non-corruption zones. These partial maps allow us to reduce the input data into a position on each map. The SOM maintains the structure of the original data in a two-dimensional map, and therefore similar regions are placed close to one another in each map. Based on these maps, we classify each Spanish province, and, depending on the zone in which they are placed, we define them as corrupt or not corrupt at a given moment. Thus, SOM maps provide an easy way to visualize and compare the level of corruption in each province at the same time.

Finally, the output or probability of corruption of the three previous models is combined into a single final map to create a hybrid model. Results from hybrid models have been used as a way to obtain more accurate models than either of the techniques used separately. The result provides a map on which provinces can be placed according to their profile in terms of the probability of corruption occurring. We use the output of the different provinces in each map as input to train the final map. In this combined map, we provide no output or information about the presence or absence of corruption within a region in the training phase. This map is therefore trained in an unsupervised way.

Once the final map is trained, the *k*-means algorithm is used to display different groups of provinces conditional on their propensity for corruption. The likelihood of a province being considered corrupt in different time horizons is then computed and the main economic factors that cause corruption to occur in a region are identified.

Previous literature suggests that regions (countries) with similar socio-economic characteristics have a similar predisposition towards corruption. Consistent with analogous analyses (Aidt 2003, 2009), the macroeconomic conditions of the Spanish provinces are chosen to develop the maps. Table 1 provides a list and description of nine variables classified into five categories (budget balance, debt levels, economic growth, labor force and political factors). These variables enhance the comparability of our results in the international literature. The variables are discussed below:

- *Real estate taxation* is one of the main sources of revenue for municipalities and provinces. It is levied on the ownership of real estate, whether rural or urban. Local governments can exercise substantial discretion when applying this tax. Spanish law sets a minimum and a maximum tax rate, which is then applied at the discretion of each municipality. We compute the average tax rate in a province and compare this average rate with the maximum and minimum possible tax rates by law. A higher ratio can potentially induce more corruption by incentivizing tax evasion given the relation between tax rates, corruption, and tax evasion as shown by Ivanyna et al. (2010).
- *Debt per capita* is the province's total outstanding public debt relative to its population. The underlying rationale for this variable is the close relation between corruption and public debt. Cooray and Schneider (2013) show that corruption, increased government expenditure, and the size of the shadow economy lead to increased public debt. Grechyna (2012) concludes that corruption causes higher public debt levels for a sample of high income OECD countries, and Nguyen (2006) reaches the same conclusion for a group of emerging economies.
- *Debt service* measures the region's ability to repay its debt by comparing total public debt with total public revenues of each year.
- *Deposit institution growth* is computed by comparing the number of deposit institutions between two consecutive years. The increase in the number of deposit institutions is usually related to economic growth, which implies more interactions between the private and public sector, which can increase the likelihood of corruption (Goel et al. 2012).
- *Population growth* is the rate at which the number of inhabitants in a province increases between two consecutive years. Prior studies find a positive relation between corruption and population (Alt and Lassen 2003; Damania et al. 2004) based on the more frequent interactions between the private and public sectors in regions with a growing population. Knack and Azfar (2003) also find evidence that corruption increases as the population grows.
- *Variation in the number of registered companies* measures the increase or decrease in the number of live registered companies. As with previous indicators, an increase can imply greater economic growth and more investment in the region. In turn, as the number of companies increases, the incentive to pay bribes in order to secure a better position or market share than competitors also increases.
- *House price increase* accounts for the fact that a high proportion of corruption cases in Spain are related to the construction industry (Benito et al. 2015). Between 1997 and 2006, an increase in household savings and population, combined with a reduction in the unemployment rate and the interest rate, caused house prices to rise by nearly 7% annually. The most corrupt areas are the regions in which urban and environmental standards are the least respected.
- *Unemployment rate* is the percentage of the total labor force that is unemployed and actively seeking employment and willing to work. The unemployment rate is usually

Table 1 Initial set of macroeconomic variables

Variable category/code	Variable name	Variable calculation	Literature research
Budget balance			
RE_TAXATION	Real estate taxation	(Real estate tax rate − Legal minimum tax rate)/(Legal maximum tax rate − Legal minimum tax rate) * 100	Ivanyna et al. (2010)
Debt levels			
DEBT_CAPITA	Debt per capita	Government's total debt/province population	Cooray and Schneider (2013); Grechyna (2012); Nguyen and van Dijk (2012)
DEBT_SERVICE	Debt service rate	Outstanding debt/total revenues	
Economy growth			
DEPOT_INST	Deposit institution growth	(Number of deposit institutions year N − Number of deposit institutions N − 1)/Number of deposit institutions N − 1	Goel et al. (2012)
POP_GROWTH	Population growth	(Total population year N − Total population year N − 1)/Total population year N − 1	Alt and Lassen (2003); Damania et al. (2004); Knack and Azfar (2003)
COMPANIES_GROWTH	Variation in the number of registered companies	(Number of active firms year N − Number of active firms year N − 1)/Number of active firms year N − 1	
HOUSE_GROWTH	House price growth	(Average of house prices per m^2 year N − Average of house prices per m^2 year N − 1)/Average of house prices per m^2 year N − 1	Benito et al. (2015)
Labor force			
UNEMPL	Unemployment rate	Number of unemployed people over the age of 16/total labor force	Habib and Leon (2002); Rehman and Naveed (2007); Saha and Gounder (2013); Bouzid (2016)
UNEM_GROWTH	Unemployment rate growth	(Unemployment rate year N − Unemployment rate year N − 1)/Unemployment rate year N − 1	
Political factors			
YEARS_GOVER	Number of years in government	Number of years since the political party came into office	Besley and Case (1995); Ferraz and Finan (2007); Ferejohn (1986)

Table 1 continued

Variable category/code	Variable name	Variable calculation	Literature research
MAJORITY	The ruling party has an overall majority	Tavits (2007)	

related to high informal sectors and corruption. Saha and Gounder (2013) identify the unemployment rate among other determinants that explain differences in corruption between countries. Conversely, Bouzid (2016) finds that corrupt practices tend to increase the unemployment rate, especially in the case of young and educated job seekers.

- *Unemployment rate growth* is the variation of the unemployment rate between two consecutive years. Unemployment has a major impact on corruption. Habib and Leon (2002) and Rehman and Naveed (2007) show that corruption reduces the levels of foreign investment, and results in an increase in the unemployment rate.
- *Number of years in government* the number of years the ruling party has been in office. Given politicians' interest in being re-elected, the ruling government can use the power to set up a network of relationships that help them to get re-elected (Ferejohn 1986; Besley and Case 1995). Ferraz and Finan (2007) find that in municipalities where mayors are in their second and final term, there is significantly more corruption than in similar municipalities where mayors are serving their first term in office.
- *Governments ruling in majority* the winning party enjoys a majority if it obtains at least half plus one of the seats in the last election. In other cases, there will either be a coalition or a minority government. Tavits (2007) reports a negative correlation between majority government and corruption on a cross-section of countries. The main argument is that when there is a majority government, the responsibilities are clearer for citizens. In turn, when political institutions provide high clarity of responsibility, politicians face incentives to pursue good policies and reduce corruption.

Other variables such as the crime rate, the educational level or electoral absenteeism were initially considered but data were not available with sufficient coverage.

Table 2 reports the mean; standard deviation; minimum; maximum; 25th, 50th, and 75th quartile and *p* value of the Shapiro–Wilk normality test. According to this test, not all the variables are normally distributed at the 5% significance level. When the mean values are compared between regions with and without corruption cases, the non-normality of variables makes the nonparametric test (Mann–Whitney U test) more reliable than the parametric test. Thus, Table 3 reports the Mann–Whitney test.

Table 3 compares the mean of the variables between regions with and without corruption cases: 107 out of the 400 province-year observations were corrupt. The table provides the mean for each group in each year. The same analysis is reported one, two, and 3 years before cases were disclosed. Provinces in the sample are weighted considering the recorded cases of corruption per 100,000 people so as to avoid biases by population or government size. The last group of columns reports the *p* value of the Mann–Whitney test of means equality. The lower the *p* value, the more likely the means are to be different. The

Table 2 Descriptive statistics

Variable code	# Obs.	Mean	Std.	Min	Max	Q25	Q50	Q75	SW sig.
RE_TAXATION	400	0.185	0.102	0.000	0.433	0.117	0.173	0.262	0.000
DEBT_CAPITA	250	0.480	0.221	0.116	1.449	0.337	0.446	0.546	0.925
DEBT_SERVICE	250	0.063	0.030	0.017	0.192	0.044	0.057	0.078	0.633
DEPOT_INST	400	0.002	0.041	− 0.204	0.102	− 0.026	0.009	0.030	0.000
POP_GROWTH	400	0.010	0.012	− 0.012	0.061	0.002	0.006	0.014	0.000
COMPANIES_GROWTH	400	0.013	0.035	− 0.140	0.206	− 0.015	0.012	0.035	0.281
HOUSE_GROWTH	400	0.002	0.081	− 0.206	0.206	− 0.058	− 0.012	0.066	0.769
UNEM_GROWTH	400	0.144	0.180	− 0.102	1.016	0.023	0.109	0.200	0.002
UNEMPL	400	0.090	0.030	0.020	0.210	0.050	0.080	0.110	0.000
YEARS_GOVER	400	17	7.81	1	24	5	17	20	0.000
MAJORITY	400	0.571	0.495	0	1	0	1	1	–

SW sig. is the *p* value to reject the null hypothesis of normal distribution of the variable according to the Shapiro–Wilk test

means of these variables are significantly different between regions with and without corruption cases, which confirms the choice of explanatory variables. The time framework is also relevant for the comparison. Tax range, population growth, unemployment rates, and number of years in office are significantly different between both groups of provinces one, two, and 3 years before corruption is uncovered, so that they are likely to play an important role in predicting corruption. One year before the corrupt acts are uncovered, the increase in unemployment and in the number of deposit institutions are also different (i.e., both are higher in corrupt regions). In contrast, 3 years before corruption is uncovered, the Mann–Whitney test shows significant differences in all the variables except debt service and governing with an absolute majority. Thus, as a preliminary conclusion, the differences between corrupt and non-corrupt regions diminish when we examine the moment at which corruption is reported.

In other kinds of prediction models such as bankruptcy models, the accuracy of the model and the predicting power of the individual variables increase as the bankruptcy date approaches (Ivanyna et al. 2010; Grechyna 2012). In our case, the trend is the opposite, with more significant differences for longer prediction periods. This finding suggests that corruption could have been uncovered some time before had a reliable method of prediction been available. Conversely, in the very short term, the differences between corrupt and non-corrupt regions diminish.

Public debt is not very different between corrupt and non-corrupt provinces. Less data exists about public debt because this information is only available from 2008 (see Table 3). Thus, due to missing data and lower predictive power, we drop DEBT_SERVICE and DEBT_CAPITA and retain the remaining nine variables for the remainder of our analysis. We do not find significant differences if a government is ruling with a majority or not. Consequently, we remove this variable for subsequent analysis.

Table 3 Test of means comparison

	Mean in $t-1$		Mean in $t-2$		Mean in $t-3$		Mann–Whitney U test		
	N	Y	N	Y	N	Y	U ($t-1$)	U ($t-2$)	U ($t-3$)
RE_TAXATION	0.169	0.218	0.173	0.216	0.176	0.214	0.000	0.000	0.002
DEBT_SERVICE	0.062	0.066	0.064	0.063	0.064	0.063	0.633	0.499	0.630
DEBT_CAPITA	0.480	0.481	0.490	0.453	0.488	0.437	0.925	0.080	0.030
COMPANIES_GROWTH	0.010	0.011	0.006	0.021	0.002	0.033	0.945	0.000	0.000
HOUSE_GROWTH	− 0.012	0.000	− 0.017	0.014	− 0.026	0.042	0.968	0.041	0.000
POP_GROWTH	0.008	0.013	0.007	0.015	0.007	0.016	0.000	0.000	0.000
DEPOT_INST	− 0.004	0.015	− 0.009	0.028	− 0.009	0.033	0.000	0.000	0.000
UNEMPL	0.093	0.099	0.099	0.086	0.102	0.075	0.052	0.009	0.000
UNEM_GROWTH	0.125	0.183	0.122	0.195	0.159	0.105	0.002	0.013	0.003
YEARS_GOVER	13.14	17.83	13.50	16.51	13.58	15.27	0.000	0.003	0.045
MAJORITY	0.58	0.56	0.58	0.57	0.56	0.54	0.112	0.250	0.320

N indicates no reports of corruption, and Y represents provinces in which corruption cases have gone to trial

5 Empirical Results

5.1 Time Horizon Prediction Models

Once we test the ability of our variables to predict corruption in Spanish provinces, we create three different models of supervised SOMs. To validate each SOM, we divide the sample into training and validation subsets. Selected randomly, the training data in each SOM accounts for 70% of the sample.

SOMs are usually implemented in an unsupervised way, and the network does not receive any output information provided in the training phase. We improve the model by providing the network with the output (corrupt or not corrupt region) so as to train the map, converting the model into a supervised version. The distance of the input to a unit is defined as the sum of the separate distances for X (macroeconomic variables) and Y (region situation) spaces. The prediction is carried out using only the X space. Introducing class membership information into the learning process increases the performance relative to traditional SOMs (Hagenbuchner and Tsoi 2005). As part of the preprocessing, all the variables are linearly scaled to have a zero mean and unit variance. Each model uses the same macroeconomic variables but a different independent variable, which is a binary variable depending on whether any corruption cases have been reported in the province in year $t-1$, $t-2$, and $t-3$. The size of each two-dimensional map is fixed following the recommendations of Kohonen (1993) and Kaski and Kohonen (1994) to maintain a balance between quantification and topological errors. The quantification error is calculated as the average distance between each data vector and its best matching unit or final position in the map. The topological error measures the topology preservation and is calculated as the proportion of all the data vectors for which the first and second best matching units are not adjacent. Different SOM sizes (5×5, 6×6, 7×7, 8×8, 9×9, 10×10, 11×11, and 12×12) and different learning parameters are tested to determine the parameters that render the lowest error rate in classification. Test results show that the best parameters are similar for each of the three supervised maps. Specifically, the optimal size is a 9×9 cell grid, and the weight assigned to the X data is 0.5. Learning rate and decay are initially set at 0.6 and 0.1, respectively. The chosen neighborhood function is Gaussian. Each map is also trained in two phases: a rough training phase with a large initial neighborhood width and a fine-tuning phase with a small initial neighborhood width.

In these prediction models, two types of error can occur: predicting as corrupt a province that is not involved in corruption cases and not predicting as corrupt a province that is. Thus, to assess the results of our model, we compare predicted cases with actual observed cases for both corrupt and non-corrupt provinces.

Table 4 provides the results. We report the classification results of the training and validation sample for the three models. As previously stated, the performance of the model improves as long as corruption is predicted on a longer term basis. The adjustment of the training sample is 86.74% 1 year before corruption comes to light and 88.49% when we use information 3 years in advance. Similarly, the proportion of accuracy for the test sample is 74.17% 1 year in advance and 84.30% 3 years in advance.

We compare the results of our supervised SOM models with two of the most widely used artificial neural network approaches within the field of task classification: multi-layer perceptron and the radial basis function network. Table 5 shows the correct classification rates of the three methods. There are no major differences among the three methods, although it is the SOM approach which provides the best results. Although the multi-layer

Table 4 Results of the classification

Observed\predicted	Corruption in 3 years			Corruption in 2 years			Corruption in 1 year		
	0	1	Overall (%)	0	1	Overall (%)	0	1	Overall (%)
Training sample									
0	190	14	93.14	180	13	93.26	168	16	91.30
1	18	56	75.68	18	68	79.07	21	74	77.89
Total	208	70	88.49	81	279	88.89	189	90	86.74
Test sample									
0	78	10	88.64	62	21	74.70	60	19	75.95
1	9	24	72.73	9	28	75.68	12	29	70.73
Total	87	34	84.30	71	49	75.00	72	48	74.17

Table 5 Correct classification percentage calculated with data from the test sample

	Years before corruption		
	3	2	1
Non-corrupt cases			
SOM	88.64	74.70	75.95
MLP	**89.77**	**77.31**	**78.48**
RBF	86.36	72.29	74.68
Corrupt cases			
SOM	**72.73**	**75.68**	**70.73**
MLP	69.70	67.57	**70.73**
RBF	63.63	70.27	65.85
Total accuracy			
SOM	**84.30**	**75.00**	74.17
MLP	76.67	74.17	**75.83**
RBF	73.33	71.67	71.67

Bold indicates Models with the highest classification rates

perceptron approach predicts non-corrupt cases slightly better, the SOM predicts corrupt cases more accurately. These results confirm the ability of the supervised SOM to predict corruption at least as well as most supervised models. However, SOM also provides a visual representation of provinces at the same time, which provides a quick snapshot of the situation in provinces and the risk of corruption.

5.2 SOM Early Corruption Warning System

In the final step of the analysis, the output of the three previous SOMs is combined to obtain a final map or early warning corruption system. The resulting probability of each region on each map is trained as a standard unsupervised SOM. Up to this point, the trained models produce three different maps and the probability of a given province having corruption cases. In this step, the different corruption profiles from the different time horizons are combined to create a visual tool. The input of this final map is the probability of a given province having corruption cases in the last three trained models, as previously explained.

Following the same previously discussed method to determine the size of the map, a trained map of 13 × 8 cells is obtained. Once the model is trained, all the regions are classified inside this new unsupervised map. In the final step, the optimal separations or clusters in the map are identified.

A key decision concerns the number of groups to be formed. If the number is too low, the groups will be too heterogeneous. However, too many groups can result in characteristics common to the regions being insufficiently identified. The ideal number of groups is one that maximizes intra-group homogeneity and inter-group heterogeneity. Following previous literature, we apply the *K*-means non-hierarchical clustering function to find an initial partitioning (Moreno et al. 2006; Kuo et al. 2002). The clustering process has no predefined classes, so the number of groups must be set a priori. Prior studies propose several measures to check the validity and quality of the results: the Silhouette index, the homogeneity and separation index, the weighted inter-intra index, and the Davies-Bouldin index, among others.

The Davies–Bouldin index (1979) is one of the most widely used algorithms (Kang et al. 2016). It is a function of the ratio within-cluster variation to between-cluster variations. The smaller the index, the better the partition. According to this index, the optimal number of groups in our sample is five. Figure 3 provides the final map.

Groups are labeled according to the proportion of detected corruption cases, with group 1 (group 5) having the fewest (most) cases of corruption. In turn, different profiles of corruption are established using this approach. In the next section, the main characteristics of the corruption profiles are described and the implications and uses of this approach are discussed.

5.3 Discussion

Table 6 shows the different incidence of corruption across groups. Groups 5 and 4 reported corruption cases in 70.22 and 44.24% of included regions, respectively. Groups 2 and 3 are intermediate groups, with a corruption rate of 31.58 and 34.74%, respectively. Group 1

Fig. 3 Early corruption warning system final map

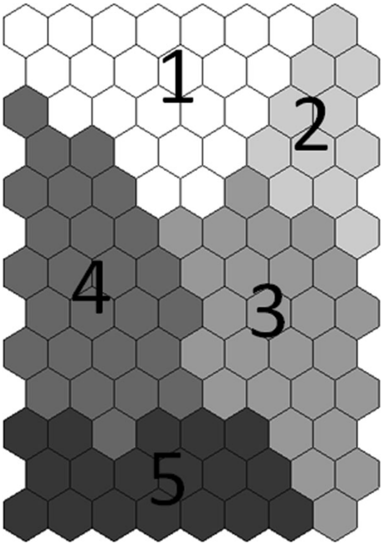

Table 6 Percentage of corrupt provinces by group and time before detection

Group identification	Number of provinces by group	Corrupt provinces 3 years before (%)	Corrupt provinces 2 years before (%)	Corrupt provinces 1 year before (%)	Corrupt cases to total regions ratio (%)
1	179	6.70	5.59	11.73	8.01
2	19	63.16	15.79	15.79	31.58
3	71	19.72	19.72	64.79	34.74
4	56	56.36	63.64	12.73	44.24
5	75	50.67	81.33	78.67	70.22
Total	399[a]	26.82	30.83	34.09	100.00

[a] Although we analyze 400 region-year observations, the information for the Balearic Islands in 2005 was not available

comprises provinces with the lowest corruption rate, with only 8.01% of the provinces involved in corruption cases.

Table 7 provides the mean of the explanatory variables for each group. The last two columns report the cluster-wise comparison using analysis of variance (F test and p value). The result from the analysis of variance shows significant differences in the economic situation among groups. Taken together, Tables 6 and 7 provide some interesting insights into corruption in Spain. The variables indicating economic growth are highly correlated with corruption. In other words, although prior literature reports the negative effect of corruption on growth and investment, the most corrupt regions in our sample grew rapidly in the time before corrupt cases were identified. Table 6 highlights the two extreme situations on our map: group 1, which includes the least corrupt regions, and groups 4 and 5, which comprise the most corrupt regions.

Our model suggests a link between corrupt regions and the real estate bubble that occurred in Spain during our sample period (2005–2012). As a consequence of the high liquidity in the financial market and low interest rates, increased demand for houses fueled the construction industry and led to an increase in real estate prices, which were financed

Table 7 Means distribution by group

	Gr. 1 (%)	Gr. 2 (%)	Gr. 3 (%)	Gr. 4 (%)	Gr. 5 (%)	Mean (%)	F. Stat.	Sig.
RE_TAXATION	15.55	17.35	22.73	17.99	22.70	18.60	11.041	0.000
COMPANIES_GROWTH	0.28	3.00	− 0.08	2.42	2.30	1.02	12.457	0.000
HOUSE_GROWTH	− 2.92	5.72	− 1.14	1.66	1.29	− 0.77	3.559	0.007
POP_GROWTH	0.54	1.45	0.86	1.50	1.56	0.96	15.936	0.000
DEPOT_INST	− 1.74	2.26	− 0.33	2.40	3.37	0.23	35.894	0.000
UNEMPL	9.84	6.48	11.32	8.03	8.75	9.49	10.668	0.000
UNEM_GROWTH	11.27	10.67	15.21	14.27	22.39	14.45	5.501	0.000
YEARS_GOVER	8.23	9.76	15.82	22.14	19.11	16.35	8.349	0.000

by new bank branches and deposit institutions. The spillover effects of the construction industry laid the foundation for Spain's rapid economic growth in the late 1990s and early years of the twenty-first century.

However, this accelerated growth came at a cost. In the most corrupt regions (i.e., group 5 and, to some extent, group 4), the number of deposit institutions grew much faster than the population. The number of non-financial firms also increased in these regions, which resulted in strong competition and may have led to bribery as a way to obtain a better position in such competitive markets. Thus, our model shows that regions with real estate prices growing faster than the average and both the number of deposit institutions and non-financial firms growing faster than the population are among the most prone to generate corruption cases. The mass media and the courts, which have proved that some companies paid bribes to officials to strengthen their position in the market and obtain public licenses, especially in the construction sector, support this model. These bribes could have been received by officials who had been in office for a long time. Actually, in groups 4 and 5 the ruling party has been in power longer than in less corrupt regions. Thus, these officials are likely to have had enough time to create a network of corrupt practices. Conversely, the characteristics of the least corrupt group, group 1, are quite dissimilar from most corrupt groups 4 and 5. Group 1 provinces have the lowest tax pressure and the lowest population growth. Public authorities still have the power to increase some taxes should there be a future economic recession or macroeconomic difficulties. In addition, the growth in the number of firms is below the mean. This group of provinces has the lowest increase both in real estate prices and in the number of deposit institutions. Finally, it also has the highest turnover in power, such that the ruling party has not been in power for too long.

In addition to the economic and political factors related to corruption, the time horizon of corruption is quite relevant, especially when resources for fighting corruption are limited and scarce. Thus, if different temporal patterns were detected, resources could be allocated in the different regions more efficiently. For example, in regions where a high likelihood of corruption exists in the short term, more resources and efforts should be assigned. It also allows the establishment of action plans to avoid corruption in the medium or long term in regions at least risk.

Two main patterns have been identified in terms of the time dimension. These patterns are attributable to the different economic and political models in each region. As explained, Spanish regions have followed quite a diverse historical and social evolution, which has resulted in distinct economic and political situations. The first model concerns groups 3 and 5, and shows a clear increasing trend before corruption is detected. For instance, in group 3 the proportion of provinces identified as corrupt 3 years before the corruption was discovered is 19.72%, whereas this proportion soars to 64.79% 1 year before (see Table 6). Group 5 shows similar results. This trend may imply that, in these regions, efforts to discover and fight corruption must be short term because it is difficult to identify and prevent corruption several years before it happens. Second, the opposite trend holds in groups 2 and 4: the closer to the moment of detection, the lower the corruption rate. In group 2, for instance, the proportion of corrupt provinces 3 years before detection is 63.16%, whereas this percentage is only 15.79% the year before. Consequently, efforts to fight corruption should be long term, and public authorities should be aware that corrupt behaviors arise in these regions in the long run.

For regions located in the groups more prone to corruption in the short term (Groups 3 and 5), resources for investigating potential public crimes should be increased as soon as possible. However, early detection of corruption patterns proves insufficient unless supported by legal measures. Such measures should ensure a quick response so as to prevent

Table 8 Average annual crimes against public administration per 100,000 inhabitants

Group	Mean number of crimes
1	0.926
2	1.269
3	1.498
4	1.355
5	1.562

corrupt behavior from persisting. For example, when there are signs of corruption, an investigation should begin. The effectiveness of the investigation depends to a large extent on the availability of more judicial resources to fight corruption as well as on having independent courts. The work of these courts should be carried out by public intervention personnel or by inspectors with unquestionable independence and objectivity, elected by majority support and not only by the ruling government. Furthermore, should a politician be formally charged, he/she should be expelled, albeit as a precautionary measure, given the social repercussion of such behavior.

The situation may differ for regions in which corruption cases might arise in the medium or long term (groups 2 and 4). The investigation should be initiated as well, but the priority is different. In these cases, measures should promote a strategic plan against corruption with more complex or deep-rooted laws. The best way to reduce or prevent corruption in the future is through prevention. Some examples of these more deep-rooted proposed measures could be:[5] reform of the Law on Political Party Financing and the Law on Conflicts of Interest; a modification of the Criminal Procedure Act to increase the penalties for tax fraud and public corruption; to ensure the independence of the Judicial Authority and the Prosecutor General's Office; digitization of public information and transparency; the creation of a specialized public agency or tax office to recover the money defrauded by corruption; or the review of incompatibility of political positions in the public sector. Citizens are also responsible for demanding a battery of actions aimed at pursuing and punishing corruption. This applies particularly to public officers. Any civil servant who has witnessed or has proof of public crimes should be encouraged to report said crimes.

To conclude, we address the question concerning the reliability of our results by comparing our predictions with previous literature. Given the innovative nature of our research, the main question is the way in which we measure corruption and the importance we attach to economic factors as causes of corruption. Habib and Leon (2002) use the number of crimes against public administration as a proxy for political corruption and show that this metric performs as well as other corruption indexes. Therefore, we analyze whether the crimes against public administration rate differs across the groups as we have defined them. Table 8 provides results which corroborate our model.

Data come from judicial statistics from the Spanish National Statistics Institute. The number of crimes against public administration is scaled by the population of each region to calculate the mean value for each group. Groups 1 and 5 have the lowest and highest rate of crimes against public administration, respectively. Furthermore, the trend toward more crimes against public administration is almost uniform as corruption increases (with the exception of group 3, which has more crimes than group 4). The conclusion is that the model, as outlined in this study, provides an accurate forecast of the likelihood of

[5] Some of these measures are under study or are in the process of being implemented.

corruption cases, and reliability is validated by comparing the results with those of similar studies.

6 Concluding Remarks

We develop a model of neural networks to predict public corruption based on economic and political factors. We apply this model to the Spanish provinces in which corrupt cases have been uncovered by the media or have gone to trial. Unlike previous research, which is based on the perception of corruption, we use data on actual cases of corruption. The output of our model is a set of SOMs, which allow us to predict corruption in different time scenarios before corruption cases are detected.

Our model provides two main insights. First, we identify some underlying economic and political factors that can result in public corruption. Taxation of real estate, economic growth, and an increase in real estate prices, in the number of deposit institutions, and the same party remaining in office for a long time seem to induce public corruption. Second, our model provides different time frameworks to predict corruption. In some regions, we are able to detect latent corruption long before it emerges (up to 3 years), and in other regions our model provides short-term alerts, and suggests the need to take urgent preventive or corrective measures.

Given the connection we find between economic and political factors and public corruption, some caveats must be applied to our results. Our model does not mean that economic growth or a given party remaining in power causes public corruption but that the fastest growing regions or the ones ruled by the same party for a long time are the most likely to be involved in corruption cases. Economic growth per se is not a sign of corruption, but rather it increases the interactions between economic agents and public officers. Similarly, being in office too long might prove to be an incentive for creating a network of unfair relations between politicians and economic agents. In addition, more competitive markets may induce some agents to pay bribes in order to obtain public concessions or a better competitive position. These results are consistent with some research exploring the relation between economic growth and corruption (Kuo et al. 2002; Kaufman and Rousseeuw 2009; Chen et al. 2002).

Since corruption remains a widespread global concern, a key issue in our research is the generalizability of our model and the proposed actions. We have used fairly common macroeconomic and political variables that are widely available from public sources in many countries. In turn, our model can be applied to other regions and countries as well. Of course, the model could be improved if country or region-specific factors were taken into account.

Our approach is interesting both for academia and public authorities. For academia, we provide an innovative way to predict public corruption using neural networks. These methods have often been used to predict corporate financial distress and other economic events, but, as far as we are aware, no studies have yet attempted to use neural networks to predict public corruption. Consequently, we extend the domain of neural network application. For public authorities, we provide a model that improves the efficiency of the measures aimed at fighting corruption. Because the resources available to combat corruption are limited, authorities can use the early corruption warning system, which categorizes each province according to its corruption profile, in order to narrow their focus and better implement preventive and corrective policies. In addition, our model predicts corruption cases long before they are discovered, which enhances anticipatory measures. Our model can be especially relevant in countries suffering the severest corruption problems. In fact, European Union authorities are highly concerned about widespread corruption in certain countries.

The study of new methodologies based on neural networks is a fertile field to be applied to a number of legal and economic issues. One possible direction for future research is to extend our model to the international framework and to take into account country-specific factors. Another application may be the detection of patterns of corruption and money laundering across different countries in the European Union.

Acknowledgements We are grateful to Alisa Larson and Philip Jaggs for their comments on previous versions. All the remaining errors are the authors' sole responsibility. We acknowledge the Spanish Ministry of Economy and Competitiveness for financial support (Project ECO2014-56102-P). This paper was prepared also within the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) and supported within the framework of a subsidy granted to the HSE by the Government of the Russian Federation for the implementation of the Global Competitiveness Program.

References

Aidt, T. S. (2003). Economic analysis of corruption: A survey. *The Economic Journal, 113*(491), F632–F652.

Aidt, T. S. (2009). Corruption, institutions, and economic development. *Oxford Review of Economic Policy, 25*(2), 271–291.

Albornoz, F., & Cabrales, A. (2013). Decentralization, political competition and corruption. *Journal of Development Economics, 105,* 103–111.

Alt, J. E., & Lassen, D. D. (2003). The political economy of institutions and corruption in American states. *Journal of Theoretical Politics, 15*(3), 341–365.

Ambrey, C. L., Fleming, C. M., Manning, M., & Smith, C. (2016). On the confluence of freedom of the press, control of corruption and societal welfare. *Social Indicators Research, 128*(2), 859–880. https://doi.org/10.1007/s11205-015-1060-0.

Benito, B., Guillamón, M.-D., & Bastida, F. (2015). Determinants of urban political corruption in local governments. *Crime, Law and Social Change, 63*(3–4), 191–210.

Besley, T., & Case, A. (1995). Does electoral accountability affect economic policy choices? Evidence from gubernatorial term limits. *The Quarterly Journal of Economics, 110*(3), 769–798.

Bouzid, B. N. (2016). Dynamic relationship between corruption and youth unemployment: Empirical evidences from a system GMM approach. *World Bank Policy Research Working Paper*: World Bank.

Carboni, O. A., & Russu, P. (2015). Assessing regional wellbeing in Italy: An application of Malmquist–DEA and self-organizing map neural clustering. *Social Indicators Research, 122*(3), 677–700.

Cavoli, T., & Wilson, J. K. (2015). Corruption, central bank (in)dependence and optimal monetary policy in a simple model. *Journal of Policy Modeling, 37*(3), 501–509. https://doi.org/10.1016/j.jpolmod.2015.03.012.

Chen, G., Jaradat, S. A., Banerjee, N., Tanaka, T. S., Ko, M. S., & Zhang, M. Q. (2002). Evaluation and comparison of clustering algorithms in analyzing ES cell gene expression data. *Statistica Sinica, 12*(1), 241–262.

Clausen, B., Kraay, A., & Nyiri, Z. (2011). Corruption and confidence in public institutions: Evidence from a global survey. *The World Bank Economic Review, 25*(2), 212–249.

Cooray, A., & Schneider, F. (2013). How does corruption affect public debt? An empirical analysis. *Working Paper*: Department of Economics, Johannes Kepler University of Linz.

D'Agostino, G., Dunne, J. P., & Pieroni, L. (2016). Government spending, corruption and economic growth. *World Development, 84,* 190–205.

Damania, R., Fredriksson, P. G., & Mani, M. (2004). The persistence of corruption and regulatory compliance failures: Theory and evidence. *Public Choice, 121*(3–4), 363–390.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 1*(2), 224–227.

de Figueiredo, J. N. (2013). Are corruption levels accurately identified? The case of U.S. states. *Journal of Policy Modeling, 35*(1), 134–149. https://doi.org/10.1016/j.jpolmod.2012.01.006.

Del Monte, A., & Papagni, E. (2007). The determinants of corruption in Italy: Regional panel data analysis. *European Journal of Political Economy, 23*(2), 379–396.

Diaby, A., & Sylwester, K. (2014). Bureaucratic competition and public corruption: Evidence from transition countries. *European Journal of Political Economy, 35,* 75–87.

Dong, B., & Torgler, B. (2013). Causes of corruption: Evidence from China. *China Economic Review, 26,* 152–169. https://doi.org/10.1016/j.chieco.2012.09.005.

European Commission (2014). *EU Anti-corruption Report*. Brussels: European Commission.

Fan, C. S., Lin, C., & Treisman, D. (2009). Political decentralization and corruption: Evidence from around the world. *Journal of Public Economics, 93*(1–2), 14–34. https://doi.org/10.1016/j.jpubeco.2008.09.001.

Ferejohn, J. (1986). Incumbent performance and electoral control. *Public Choice, 50*(1), 5–25.

Ferraz, C., & Finan, F. (2007). Electoral accountability and political corruption in local governments: Evidence from audit reports," *Working Paper*: IZA.

Fisman, R., & Gatti, R. (2002). Decentralization and corruption: Evidence across countries. *Journal of Public Economics, 83*(3), 325–345. https://doi.org/10.1016/S0047-2727(00)00158-4.

Gerring, J., & Thacker, S. C. (2005). Do neoliberal policies deter political corruption? *International Organization, 59*(1), 233–254.

Goel, R. K., Nelson, M. A., & Naretta, M. A. (2012). The internet as an indicator of corruption awareness. *European Journal of Political Economy, 28*(1), 64–75.

Grechyna, D. (2012). Public corruption and public debt: Some empirical evidence.

Guo, Y., Zhou, W., Luo, C., Liu, C., & Xiong, H. (2016). Instance-based credit risk assessment for investment decisions in P2P lending. *European Journal of Operational Research, 249*(2), 417–426. https://doi.org/10.1016/j.ejor.2015.05.050.

Habib, M., & Leon, Z. (2002). Corruption and foreign direct investment. *Journal of International Business Studies, 33*(2), 291–307. https://doi.org/10.2307/3069545.

Hagenbuchner, M., & Tsoi, A. C. (2005). A supervised training algorithm for self-organizing maps for structures. *Pattern Recognition Letters, 26*(12), 1874–1884.

Hagenbuchner, M., Tsoi, A. C., & Sperduti, A. (2001). A supervised self-organizing map for structured data. In N. Allinson, H. Yin, L. Allinson, & J. Slack (Eds.), *Advances in self organising maps* (pp. 21–28). London: Springer.

Huysmans, J., Martens, D., Baesens, B., Vanthienen, J., & Van Gestel, T. (2006). Country corruption analysis with self organizing maps and support vector machines. In H. Chen, F.-Y. Wang, C. Yang, D. Zeng, M. Chau, & K. Chang (Eds.), *Intelligence and security informatics* (Vol. 3917, pp. 103–114)., Lecture notes in computer science Berlin: Springer.

International Monetary Fund. (2016). Corruption: Costs and mitigating strategies. *Discussion note*: IMF Fiscal affairs department and legal department.

Ivanyna, M., Mourmouras, A., & Rangazas, P. (2010). The culture of corruption, tax evasion, and optimal tax policy. *Working paper*: International monetary fund.

Ivanyna, M., & Shah, A. (2011). Decentralization and corruption: New cross-country evidence. *Environment and Planning C: Government and Policy, 29*(2), 344–362.

Jagric, T., Bojnec, S., & Jagric, V. (2015). Optimized spiral spherical self-organizing map approach to sector analysis—the case of banking. *Expert Systems with Applications, 42*(13), 5531–5540. https://doi.org/10.1016/j.eswa.2015.03.002.

Kang, Q., Liu, S., Zhou, M., & Li, S. (2016). A weight-incorporated similarity-based clustering ensemble method based on swarm intelligence. *Knowledge-Based Systems, 104,* 156–164. https://doi.org/10.1016/j.knosys.2016.04.021.

Kaski, S., & Kohonen, T. (1994). Winner-take-all networks for physiological models of competitive learning. *Neural Networks, 7*(6), 973–984.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: An introduction to cluster analysis*. New York: Wiley.

Kaufmann, D., & Bellver, A. (2005). Transparency: Initial empirics and policy applications. *Policy research working paper*. Washington: World Bank.

Kaymak, T., & Bektas, E. (2015). Corruption in emerging markets: A multidimensional study. *Social Indicators Research, 124*(3), 785–805.

Knack, S., & Azfar, O. (2003). Trade intensity, country size and corruption. *Economics of Governance, 4*(1), 1–18.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics, 43*(1), 59–69.

Kohonen, T. (1993). Physiological interpretation of the self-organizing map algorithm. *Neural Networks, 6*(6), 895–905.

Kohonen, T. (2001). *Self-organizing maps*. Berlin: Springer.

Kong, D. T., & Volkema, R. (2016). Cultural endorsement of broad leadership prototypes and wealth as predictors of corruption. *Social Indicators Research, 127*(1), 139–152.

Kunieda, T., Okada, K., & Shibata, A. (2014). Corruption, capital account liberalization, and economic growth: Theory and evidence. *International Economics, 139,* 80–108.

Kuo, R. J., Ho, L. M., & Hu, C. M. (2002). Integration of self-organizing feature map and K-means algorithm for market segmentation. *Computers & Operations Research, 29*(11), 1475–1493.

Leeson, P. T., & Sobel, R. S. (2008). Weathering corruption. *Journal of Law and Economics, 51*(4), 667–681.

León, C. J., Araña, J. E., & de León, J. (2013). Correcting for scale perception bias in measuring corruption: An application to Chile and Spain. *Social Indicators Research, 114*(3), 977–995.

Li, H., Gong, T., & Xiao, H. (2016). The perception of anti-corruption efficacy in China: An empirical analysis. *Social Indicators Research, 125*(3), 885–903.

Li, X., & Juhola, M. (2014). Country crime analysis using the self-organizing map, with special regard to demographic factors. *AI & Society, 29*(1), 53–68.

Li, X., & Juhola, M. (2015). Country crime analysis using the self–organising map, with special regard to economic factors. *International Journal of Data Mining, Modelling and Management, 7*(2), 130–153.

Lo, Z., & Bavarian, B. (1993). Analysis of convergence properties of topology preserving neural networks. *IEEE Transactions on Neural Networks, 11,* 207–220.

Lucchini, M., & Assi, J. (2013). Mapping patterns of multiple deprivation and well-being using self-organizing maps: An application to swiss household panel data. *Social Indicators Research, 112*(1), 129–149.

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

Moreno, D., Marco, P., & Olmeda, I. (2006). Self-organizing maps could improve the classification of Spanish mutual funds. *European Journal of Operational Research, 174*(2), 1039–1054.

Nguyen, T. H. (2006). *Tax corruption, public debt and the policy interaction in emerging economies*. Paper presented at the Vietnam Development Forum (VDF), Tokyo, Japan.

Nguyen, T. T., & van Dijk, M. A. (2012). Corruption, growth, and governance: Private vs. state-owned firms in Vietnam. *Journal of Banking & Finance, 36*(11), 2935–2948. https://doi.org/10.1016/j.jbankfin.2012.03.027.

Nour, M. A. (1994). Improved clustering and classification algorithms for the Kohonen self-organizing neural network. *Unpublished Ph.D. dissertation*: Kent State University.

Nwabuzor, A. (2005). Corruption and development: New initiatives in economic openness and strengthened rule of law. *Journal of Business Ethics, 59*(1–2), 121–138.

Olken, B. A. (2007). Monitoring corruption: Evidence from a field experiment in Indonesia. *Journal of Political Economy, 115*(2), 200–249. https://doi.org/10.1016/j.jpubeco.2009.03.001.

Olken, B. A. (2009). Corruption perceptions vs. corruption reality. *Journal of Public Economics, 93*(7–8), 950–964. https://doi.org/10.1016/j.jpubeco.2009.03.001.

Olszewski, D. (2014). Fraud detection using self-organizing map visualizing the user profiles. *Knowledge-Based Systems, 70,* 324–334. https://doi.org/10.1016/j.knosys.2014.07.008.

Ortega, B., Casquero, A., & Sanjuán, J. (2016). Corruption and convergence in human development: Evidence from 69 countries during 1990–2012. *Social Indicators Research, 127*(2), 691–719. https://doi.org/10.1007/s11205-015-0968-8.

Pellegata, A., & Memoli, V. (2016). Can corruption erode confidence in political institutions among European countries? Comparing the effects of different measures of perceived corruption. *Social Indicators Research, 128*(1), 391–412.

Pieroni, L., & d’Agostino, G. (2013). Corruption and the effects of economic freedom. *European Journal of Political Economy, 29,* 54–72.

Rajkumar, A. S., & Swaroop, V. (2008). Public spending and outcomes: Does governance matter? *Journal of Development Economics, 86*(1), 91–111.

Rehman, H. U., & Naveed, A. (2007). Determinants of corruption and its relation to GDP (A panel study). *Journal of Political Studies, 12*(2), 27–59.

Rende, S., & Donduran, M. (2013). Neighborhoods in development: Human development index and self-organizing maps. *Social Indicators Research, 110*(2), 721–734.

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

Salinas-Jiménez, M. M., & Salinas-Jiménez, J. (2007). Corruption, efficiency and productivity in OECD countries. *Journal of Policy Modeling, 29*(6), 903–915.

Stockemer, D., & Calca, P. (2013). Corruption and turnout in Portugal—a municipal level study. *Crime, Law and Social Change, 60*(5), 535–548.

Swiderski, B., Kurek, J., & Osowski, S. (2012). Multistage classification by using logistic regression and neural networks for assessment of financial condition of company. *Decision Support Systems, 52*(2), 539–547. https://doi.org/10.1016/j.dss.2011.10.018.

Tavits, M. (2007). Clarity of responsibility and corruption. *American Journal of Political Science, 51*(1), 218–229.

Transparency International (2009). *Global corruption report. Corruption and the private sector*. Cambridge: Cambridge University Press. Transparency International. Ernst & Young.

Transparency International (2016). *Global Corruption Barometer 2015/16*: Transparency International.

Treisman, D. (2000). The causes of corruption: A cross-national study. *Journal of Public Economics, 76*(3), 399–457.

Treisman, D. (2002). Decentralization and the Quality of Government. *UCLA manuscript*.

Treisman, D. (2007). What have we learned about the causes of corruption from ten years of cross-national empirical research? *Annual Review of Political Science, 10,* 211–214.

Van Rijckeghem, C., & Weder, B. (2001). Bureaucratic corruption and the rate of temptation: Do wages in the civil service affect corruption, and by how much? *Journal of Development Economics, 65*(2), 307–331.

Wu, Y., & Zhu, J. (2011). Corruption, anti-corruption, and inter-county income disparity in China. *The Social Science Journal, 48*(3), 435–448.

Zheng, W.-W., Liu, L., Huang, Z.-W., & Tan, X.-Y. (2017). Life satisfaction as a buffer of the relationship between corruption perception and political participation. *Social Indicators Research, 132*(2), 907–923.