Artificial Intelligence and Anti-Corruption

Fabrizio Sbicca[1, *]

[1] Autorità Nazionale Anticorruzione (ANAC), Via Marco Minghetti 10, 00187 Rome

The opinions expressed in this paper are the author's own and do not reflect the view of ANAC.

Abstract

The article presents recent developments undertaken by ANAC in the understanding of corruption and suggests possible avenues for further analysis of the phenomenon using machine learning techniques.

Keywords

corruption, public procurement, big data, machine learning

1. Introduction

Although corruption represents one of the main obstacles to economic, political, and social development, it is a latent phenomenon and, therefore, difficult to measure. Indeed, the corruptive phenomenon can be compared to an iceberg of which only the tip is visible, despite the submerged part being much larger than it appears. The cases of corruption that are learned about, for example, through court rulings, constitute the visible part, but they leave us ignorant regarding the size and characteristics of the phenomenon that remains largely hidden. Not surprisingly, there is an extreme shortage of structured scientific data on the corruptive phenomenon internationally that goes beyond the measurement of the so-called "perception" or of ad hoc studies, certainly very interesting and rich in insights, but whose contents and results are difficult to generalize.

2. ANAC's experience in measuring corruption

A significant step forward in the understanding of this phenomenon was made by the Italian Anti-Corruption Authority (ANAC), which in July 2022 presented to the public a section of its portal called "Measure Corruption" (https://www.anticorruzione.it/il-progetto). Seventy indicators are made available to the community capable of measuring the risk of corruption in the territory (https://www.anticorruzione.it/gli-indicatori).

These indicators can be considered as warning bells signaling potentially anomalous situations. They allow to have a picture of territorial contexts more or less exposed to corruptive phenomena on which to invest in terms of prevention and/or investigation. They can also direct the attention of civil society and increase civic participation. From this point of view, this system of indicators could represent a useful contribution to the country for the construction and implementation of further and more targeted tools for the prevention, monitoring, and control of corruption, with the ultimate aim of better managing the future use of public financial resources. The perspective pursued in developing the website has been to highlight the importance of strengthening collective awareness on the serious social consequences resulting from corruption. Prevention and repression are in fact necessary but not sufficient, to fight the phenomenon in a more profound way we need an increase in social capital. For this reason, the dashboards in the website are "easy", behind them

Ital-IA 2024: 4th National Conference on Artificial Intelligence, organized by CINI, May 29-30, 2024, Naples, Italy

* f.sbicca@anticorruzione.it
0009-0000-9372-2325

there are complex data, algorithms and IT structures but the result that ANAC tried to achieve is that they would be understandable to everyone and captivating, especially for young people, in order to engage more easily in questions, reflections and awareness.

In particular, the so-called "context indicators" provide an idea of the complex social and economic context of the territory in which a risk of corruption is more or less likely to manifest. This analysis indeed took into consideration 18 indicators, collected in four thematic domains (education, economy, crime, social capital). Other 25 indicators were then added. These indicators are useful for evaluating the conditions of the territorial context (for a total of 43 simple indicators), all related to the main hypotheses identified in the literature regarding factors associated with corruption. The analysis of the external context, in fact, aims to identify the cultural, economic, and social characteristics of the provincial territory in which the administrations operate, which can favor, or conversely hinder, the occurrence of corruptive phenomena.

Each thematic domain is summarized by a composite index to simplify the reading of complexity due to the many dimensions considered. The four thematic composite indicators are in turn synthesized, by combining them, into a further "composite of composites" index that therefore provides a highly informative synthetic measure on some characteristics of the entire phenomenon. Thus, the "context dashboard" makes available to the community a total of 48 indicators, of which 5 are composite.

The risk indicators for corruption in the public procurement, on the other hand, provide information related to the purchases of administrations located in the province to which they refer and are particularly important both because of the unique weight of the corruptive phenomenon in the public procurement market and the institutional purposes of ANAC. The source of the information is in fact the National Database of Public Contracts (BDNCP), a great value asset that, for the quantity and detail of the data contained, relating to about 70 million contracts, represents a unique experience at the European level. The availability of this database allows for the computation of corruption risk indicators with an extreme degree of territorial, sectoral, and temporal detail.

Based on an increasingly important and substantial body of scientific studies, ANAC has identified 17 indicators that, in various ways, identify aspects highlighting potential corruptive phenomena in the context of public procurement, thus signaling the risk of corruption in every Italian province.

An example of corruption risk indicators in public procurement is the use of discretionary procedures [3] or tenders with very few bidders [4], but also delays and cost overruns [9]. The literature identifies that low competition in tenders associated with more discretion is typically a signal of corruption risk [8]. Other examples of contract-level red flags for corruption can be found in [10, 11, 12].

The portal allows for the calculation of synthesis indicators according to different risk thresholds, obtained by condensing the information coming from all or part of the 17 indicators. For each of the selected indicators, in fact, it is possible to highlight the provinces whose value exceeds a given percentage of the provinces with a less risky value. The threshold value can be freely chosen from the 75th to the 99th percentile.

Finally, five indicators were calculated at the level of single administration, in this case, the 745 Italian municipalities with a population equal to or greater than 15,000 inhabitants. These indicators were calculated based on the statistical analysis of the relationships between variables potentially related to corruption and episodes that occurred at the level of single administration.

3. Artificial Intelligence and Anti-Corruption

What are the potential future developments and opportunities opened by technological innovation that is evolving with unprecedented speed, particularly regarding artificial intelligence? First, an opportunity arises from the increasing availability of information in large public databases of various kinds which, if properly used, allow for the extraction of potentially very useful indicators. The joint use of separate databases is very advantageous, based on the principle that the value of data tends to grow more than proportionally with the combination of different sources. In the Italian case, though, several databases are often owned by distinct public administrations. Their joint use is hindered by several factors, including concerns about privacy protection. The need to overcome such impediments is particularly urgent today, with the spread of tools and techniques for analyzing so-called "big data," which the Italian public administration generates in increasing measure. They can unleash their potential to support a public debate that is anchored in the evidence of facts and can help policymakers to take more informed decisions.

Another important aspect is certainly the digitalization of the procurement lifecycle, an important and difficult transformation process that is occurring worldwide. In Italy, digitalization has been expressively envisaged by the new Public Contract Code. First of all, digitalization could in itself constitute an effective measure for the prevention of corruption as it is likely to bring a higher degree of transparency, traceability, participation, control of activities, potentially suitable to ensure compliance with legality [1, 2, 13, 18]. With the full implementation of the digitalization of the contract lifecycle, data should be "natively digital," which could improve not only the quality and completeness of information but also allow for the acquisition of additional data not previously detected by the mentioned BDNCP or acquired in a very deficient manner. The informative bases held by ANAC could therefore have in the future a role of great importance and greater centrality also in the prevention and combat of corruption and other phenomena (such as fraud, collusion, conflict of interest) strongly detrimental to the correct functioning of the market and the effective and efficient allocation of resources in the context of public procurement, including those funded with EU funds. Recent experiences of full digitalization of the public procurement process analyzed in the literature can be found in Ukraine [6, 5] and Georgia[21].

On the other hand, both Regulation (EU) 2021/241 of February 12 2021 establishing the Recovery and Resilience Facility and Regulation (EU) 2021/1060 of 24 June 2021 laying down common provisions for different European funds , provide that Member States implement effective control mechanisms on procurement based as much as possible on methodologies and tools for collecting and analyzing large volumes of information available in computerized databases, emphasizing the centrality of risk indicators as a fundamental tool for the prevention and combat of serious irregularities in such market, such as fraud, corruption, and conflicts of interest. And “Notice on tools to fight collusion in public procurement and on guidance on how to apply the related exclusion ground” (2021/C 91/01 of 18 March 2021) , emphasizes, with specific reference to collusion, the importance of indicators as a tool to combat distortive phenomena of competition, reaffirming the need for central authorities in Member States to increasingly and effectively collaborate in the analysis of procurement data, developing methodologies and tools that are simple and easy to apply to collect and analyze large volumes of information available in computerized databases.

The ever-greater availability of large data sets has also increasingly shifted attention to the potential for developing advanced algorithms, using big data analytics and artificial intelligence in addition to traditional statistical analyses [16]. Machine learning can help identifying further and more targeted red flags that concerning both the individual transaction and the purchasing activity of a certain administration or the set of administrations in a certain territorial area. For instance, [17] studies a particular red flag for corruption, which is the degree of political connection of firms.

In this regard, AI anti-corruption tools can be defined as "data processing systems driven by tasks or problems designed to, with a degree of autonomy, identify, predict, summarize, and/or communicate actions related to the misuse of position, information and/or resources aimed at private gain at the expense of the collective good" [19].

Thanks to the processing of large volumes of data with the current processing speed, artificial intelligence can indeed contribute to uncovering patterns of corruption and identifying warning signs. Research on the potential of such tools in the field of corruption prevention and combat is still in its initial phase [14], and so far, there are not many concrete examples of application to this theme, among these are cited the case of Brazil (Anti-corruption tools based on artificial intelligence to monitor public spending, for example cartel practices); the Chinese "Zero Trust" system to predict the risk that public officials are involved in corrupt practices; the "SyRI" algorithm used by the Dutch authorities to identify fraud in the social security sector, however dismantled in 2020 due to often discriminatory and biased results; the Ukrainian "ProZorro" system to detect violations from public procurement data and prevent the misuse of public funds. Moreover, some authors used data science techniques to construct networks of firms bidding in the same auctions in the Georgian public procurement market to find possible networks of firms that collude to win public contracts [21], other researchers use a neural network approach to detect corruption in the Spanish provinces [15], or methods from network science to analyze the corruption risk at the EU level [20].

From this point of view, the indicators of the ANAC portal need to be “valid” in order to be used in the future in a targeted way and with a solid scientific basis of reference also for preventive purposes. This validation can be obtained thanks to techniques that go beyond the deductive reasoning that led to their

identification in the first place. In this regard, any further future developments exploring this line of research, already practiced in the case of the above-mentioned municipal risk indicators already present in the ANAC portal, could be based on a validation methodology that is based on the distinction between:
a. "relevant events," which are summarized by the risk indicators, for example, those related to public procurement, calculated thanks to the BDNCP;
b. "phenomena of possible corruption," as indicated by other types of sufficiently structured and numerous data, among these: judicial convictions for corruption crimes or, more generally, for crimes against the PA; reports received by ANAC; news articles related to episodes of corruption; dissolution of municipal councils for mafia infiltration, etc.

Validating the risk indicators (which summarize the "relevant events") means evaluating their capacity to "predict" the "phenomena of possible corruption". Regarding this, the procedure that could be experimented is based on two areas of statistical techniques that can be used. On one hand, there are the so-called traditional statistical models, both parametric (such as, for example, regression models) and non-parametric. On the other hand, there are various types of machine learning techniques. In both cases, the analysis is only carried out in a subset of the available data, to then be able to consider the predictive capacity of the estimated relationship "outside the sample" (the part of the data not used). This allows, among other things, to evaluate which indicators, among the alternatives considered, have the best predictive capacity. Finally, the potential of this approach should be considered where the programs used for the construction of the algorithms were made available to the public in order to prevent such measures from being perceived (or actually are) as "black boxes". This is an important issue today, and will gain even greater prominence in the future, as the use of artificial intelligence techniques that can be particularly opaque spreads.

A first example of statistical validation has already been carried out within the project on measuring corruption and concerns the 5 risk indicators at the municipal level mentioned earlier, which are in fact significantly associated with the occurrence of corruption episodes of a single administration. Unlike the 48 context indicators and the 17 procurement indicators, which were calculated at the aggregated territorial level of the province, in this case, the unit of analysis is indeed the single Municipality intended as an entity.

Based on the results achieved, it is possible to identify various possible lines of development of predictive indicators, using also machine learning techniques. First, the development of a cluster analysis on "infected" Municipalities (i.e., characterized by at least one episode of corruption in the time period examined) with the aim of identifying some subgroups of municipalities that present recurring organizational, governance, and managerial characteristics. The analyses conducted have indeed allowed identifying among the medium-large Italian Municipalities those in which episodes of corruption occurred in the five-year period 2015-2019. It might be interesting, within this group, to therefore conduct a cluster analysis to be able to identify the "similarity" between the Municipalities in which episodes of corruption were detected, proceeding with a classification of the same based on: 1) organizational variables; 2) governance variables; 3) risk indicators in public procurement; 4) accounting variables. The development of this type of investigation could allow identifying within the "infected" municipalities, subgroups characterized by a high internal homogeneity with respect to some features. This could make it possible to identify some organizational, governance, and managerial characteristics that are recurrent among the Italian Municipalities characterized by corruptive episodes.

Another possible deepening could concern the extension of the analysis to other sectors of public administration (e.g., health units) in order to identify potential corruptive risk indicators linked to organizational, managerial, and accounting variables. Indeed, it is known in the literature that Public Administrations constitute a varied and heterogeneous set of entities profoundly different in terms of institutional, organizational, managerial, accounting arrangements that operate in significantly different normative and regulatory contexts. In other words, the corruptive risk indicators could vary from sector to sector of Public Administrations, due to the specificities and differentiations that characterize them.

Finally, a further interesting development could concern the use of the findings on corruption cases in municipalities to support the development of predictive techniques of corruption risk based on artificial intelligence. Municipalities are one of the areas of the PA where there is more need to strengthen the analysis of context variables affecting the corruptive risk. The magnitude of this risk is presumably destined to grow with the use of EU funds to finance the numerous projects approved in the various territorial areas. In the most recent economic literature, the analysis of the corruptive risk in Italian municipalities has been conducted by some researchers through the application of Artificial Intelligence techniques, such as machine learning in order also to build predictive models of corruption in Italian municipalities, using as predictors a series of socio-economic, demographic, geographic, and biophysical variables, drawn from the sector literature [7]. From this point of view, the analyses conducted within the ANAC project have led to the development of a database of corruptive events for medium-large Italian Municipalities in the period 2015-2019, which has allowed to detect also numerous organizational, governance, accounting, and risk variables in public procurement. The variables available in the Datasets prepared during the project could therefore be used for the construction of new algorithms that can learn from this set of data for predictive purposes.

4. References

[1] P. Aarvik, Artificial intelligence – a promising anti-corruption tool in development settings? U4 Report Insights, Chr. Michelsen Institute (2019).

[2] I. Adam, M.Fazekas, Are emerging technologies helping in the fight against corruption? A review of the state of evidence, Information Economics and Policy 57 (2021).

[3] A. Abdou, A. Ágnes Czibik, B. Tóth, M. Fazekas, COVID-19 emergency public procurement in Romania: corruption risks and market behavior, Budapest, GTI-WP/2021:03, 2021.

[4] E. Auriol, S. Straub, T. Flochel, Public procurement and rent-seeking: the case of Paraguay, World Development, 77 (2016) 395–407.

[5] B. Baránek, V. Titl, L. Musolff, Detection of collusive networks in e-procurement,2021.

[6] S. Baumann, M. Klymak, Paying over the odds at the end of the fiscal year. Evidence from Ukraine (2022).

[7] G. de Blasio, A. D'Ignazio, M. Letta, Gotham city. Predicting 'corrupted' municipalities with machine learning, Technological Forecasting & Social Change, 184 (2022).

[8] F. Decarolis, R. Fisman, P. Pinotti, S. Vannutelli, Rules, discretion, and corruption in procurement: evidence from italian government contracting, NBER Working Paper 28209 (2020).;

[9] F. Decarolis, C. Giorgiantonio, Corruption red flags in public procurement: new evidence from italian calls for tenders, EPJ Data Science 11 (2022).

[10] M. Fazekas, I. J. Tóth, L. P. King, An objective corruption risk index using public procurement data, European Journal of Criminal Policy and Research, 22(2016) 369–397.

[11] M. Fazekas, L. Cingolani, B. Tóth, A comprehensive review of objective corruption proxies in public procurement: risky actors, transactions, and vehicles of rent extraction, Budapest, GTI-WP/2016:03, 2017.

[12] M. Fazekas, G. Kocsis, Uncovering high-level corruption: cross-national objective corruption risk indicators using public procurement data, British Journal of Political Science 50 No.1 (2020) 155-164.

[13] J. Ferwerda, I. Deleanu, B. Unger, Corruption in public procurement: finding the right indicators, European Journal on Criminal Policy and Research, 23 (2017) 245–267.

[14] J. Li, W. H. Chen, Q. Xu, N. Shah, J.C. Kohler, T.K. Mackey, Detection of self-reported experiences with corruption on twitter using unsupervised machine learning, Social Sciences & Humanities Open, 2 (2020).

[15] F.J. Lopez-Iturriaga, I.P. Sanz, Predicting public corruption with neural networks: an analysis of spanish provinces, Social Indicators Research 140 (2018) 975-998.

[16] N. Kobis, C. Starke, I. Rahwan, Artificial intelligence as an anti-corruption tool (AI-ACT) -- potentials and pitfalls for top-down and bottom-up approaches. arXiv, 2102.11567, 2021.

[17] D. Mazrekaj, F. Schiltz, V. Titl, Identifying politically connected firms: a machine learning approach, OECD Global Anti-Corruption & Integrity Forum, March 20-21,2019.

[18] F. Merenda, Legalità, algoritmi e corruzione: le tecniche di intelligenza artificiale potrebbero essere utilizzate nel e per il sistema di prevenzione della corruzione? Rivista italiana di informatica e diritto, 4 No.2 (2022) 23-38.

[19] F. Odilla, Bots against corruption: exploring the benefits and limitations of AI-based anti-corruption technology, Crime, Law and Social Change 80 (2023) 353-396.

[20] J. Wachs, M. Fazekas, J. Kertesz, Corruption risk in contracting markets: a network

science perspective, International Journal of Data Science and Analytics 12 (2021) 45–60.

[21] J. Wachs, J. Kertesz, A network approach to cartel detection in public auction markets, Scientific reports 9 No.1 (2021).