Organisation for Economic Co-operation and Development

DAF/COMP/WP3/WD(2022)33

Unclassified

English - Or. English

3 November 2022

DIRECTORATE FOR FINANCIAL AND ENTERPRISE AFFAIRS
COMPETITION COMMITTEE

Working Party No. 3 on Co-operation and Enforcement

Data Screening Tools for Competition Investigations – Note by Spain

28 November 2022

This document reproduces a written contribution from Spain submitted for Item 3 of the 136th OECD Working Party 3 meeting on 28 November 2022.

More documents related to this discussion can be found at www.oecd.org/daf/competition/data-screening-tools-for-competition-investigations.htm

Ms Despina PACHNOU
[Email: Despina.PACHNOU@oecd.org]

JT03506566

Spain

1. Introduction

1. A feature common to developed economies in recent decades is that all of them have experienced a quantitative, (but not always qualitative correlated), explosion in the volume of available open data. In general terms, it is increasingly difficult to find sectors that are not undergoing digitisation and traceability of their actions. Therefore, Competition Authorities (hereinafter "CAs") should be familiar with these concepts and know how best to take advantage of this data availability.

2. Combining information and new tools opens a new era for competition enforcers, as detecting collusion becomes easier, for instance, in public procurement. Thus, CA's should reshape their organisation to include units capable of using, developing, studying, and advancing new tools such as the automatization of tasks through algorithms, or the extraction and processing of information through data analysis, machine learning and graph databases.

3. This new approach is the basis of modern ex officio detection**,** which constitutes a complementary and proactive alternative to leniency. On the one hand, it allows the enforcer to select the markets under analysis to act on more stable cartels (whose members are less prone to defection). On the other hand, it requires considerable investment in hiring human resources, and in continuous training. Surely, while the information obtained is most often less targeted than the one derived from leniency applications, it confers more autonomy and independence from external factors to the CA.

4. Data availability is key to ex officio detection. In Spain, the quality and quantity of the data housed in public sources has improved notably thanks, among other, to various regulatory changes. Nevertheless, there are still some difficulties for its treatment such as the plurality of sources, the existence of errors or inconsistencies, or the lack of crucial information.

5. In order to address those pitfalls, the CNMC has built its own public procurement database based on the selected download of certain information from the original sources. The data retained are then automatically filtered and cleaned of obvious errors. In this new database the data are categorised according to quality levels, so that they are highly reliable.

6. Although data concerning losing bids and bidders are vital information to detect anticompetitive behaviours, until very recently CNMC could not obtain them in an aggregated and automatic way. Therefore, the CNMC's Economic Intelligence Unit (EIU) developed its own algorithm inhouse that reads all documents attached to each tender, obtaining this important information as structured data.

7. However, lacking those data in the meanwhile did not prevent the CNMC from investigating the tenders by resorting to different techniques useful in that scenario to identify indicia of collusion. Tools such as indicators on business intelligence platforms, statistics analysis or certain algorithms can prove very successful on a case-by-case basis. Now the availability of the non-winning bids opens the way to a myriad of possibilities, including machine learning and other artificial intelligence developments.

8. This CNMC's contribution to the OECD roundtable on data screening tools in competition investigations presents in a nutshell the CNMC's approach to the screening of public procurement data in Spain, focused on bid rigging detection.

2. Public Procurement Sources in Spain and CNMC Approach

9. Among the different data sources available, those relating to public procurement acquire a particularly prominent role in terms of competition. It is not for nothing that bid rigging is considered one of the most damaging behaviours for developed economies, both due to the volume of funds involved in public procurement (in general terms not less than 10/15% of GDP) and to the sensitivity of the affected sectors (education, health, defence...).

10. This section provides a brief description of the public procurement sources available in Spain and how the CNMC uses them.

2.1. Public Procurement sources in spain

11. In Spain, public procurement data are stored in different repositories:

12. Firstly, the Public Sector Contracts Register, includes basic information[1] on the contracts awarded by all contracting bodies (central, regional, autonomous bodies, other entities governed by public law) including later amendments.

13. In addition to it, Public Sector Procurement Platform (PSPP) is a completely up-to-date electronic platform publishing all the calls for tenders and their outcomes. Even if according to law, all central contracting bodies are obliged to use this platform to manage their purchasing processes, regional and local contracting bodies can choose to create their own ones, provided that the platforms of the different government authorities and public entities are interconnected to establish a single platform that centralises the publication of public sector procurement.

14. Then, Ministry of Finance Centralised Procurement Portal publishes calls for tenders in processes using framework agreements, centralised contracts and other centralised procedures. Likewise, this is a compulsory method for State-wide contracting bodies but optional for regional and local ones.

15. Finally, autonomous regions and local entities have established their own **autonomous regions platforms**[2].

16. Such complexity reflects in the exploitation of the data in an aggregated manner. Taking this into account and bearing in mind the importance of the public procurement/competition relationship, in 2015 the **CNMC** decided to undertake the construction of its **own public procurement database** by **downloading certain information from the PPSP**. Thereafter this repository is completed with data from the procurement bodies which are not integrated within PPSP, such as the centralized contracts and with the non-standardized documents associated with the tenders, such as notice of award or tender amendments.

[1] Appendix I of Royal Decree 817/2009, «the following information must be reported to the Public Sector Contracts Register: type, contract year, contracting government, contracting authority, contract identification code, place of execution, subject-matter of the contract, CPV code for the subject of the contract, procurement by lots, whether the contract is a mixed contract or a framework agreement, publicity of the contract, chosen procedure, contract prices (bidding, award), performance period, whether the time span of the contract is multiannual, price review mechanisms, contractor, award date and performance date »

[2] Autonomous regions can develop their own platforms for the publication of tender data and for e-procurement. In the first case, they must send all information about tendering processes to the PSPP. In the second one, they are not obliged to send e-procurement information to PSPP.

17. Data from the PPSP and the Ministry of Finance are downloaded using web feeds based on Atom[3], so that each entry is in XML format with all the structured data fields. For the documents attached to tenders, we either insert a direct link as a structured field on the Atom XML entry, or we developed several browser automations process (bots) to download them browsing inside the URL that we have available. Considering that each contractor's profile has different websites, structures and formats, this automatization is highly complex.

18. Therefore, **this CNMC's public procurement database could be considered one of the most complete at national level**, with structured data about tenders from all public administration levels and types, but also the related documents (non-structured data).

19. The following table shows the current total volume of tenders by source included in the CNMC's database of Spanish public tenders:

Table 1. Total volume of tenders – CNMC Database (oct 2022)

Current Total Volume of Tenders by Source - CNMC Database		
SOURCE	**Number of different tenders**	**Number of different lots**
PPSP	568.494	757.724
PPSC Regions	246.638	342.619
PSPP Minor Tenders	1.464.619	1.470.705
Centralized Contracts	45.333	45.333
TOTAL	**2.325.084**	**2.616.381**

20. Regarding the documents, the ad-hoc browser automations (bots) developed have downloaded millions of documents related with the tenders in different formats, measuring the total amount of hard disk space in Terabytes.

21. This download is carried out periodically for the continuous updating of data and always with the CNMC's own resources.

2.2. Quality and quantity issues in public procurement data in Spain: The CNMC's approach

22. The above-mentioned platforms include an extensive amount of information on a large volume of tenders. However, their usefulness could be seriously undermined by the lack of data, and by quality problems of existing data.

23. First, data **quality** is essential to obtain valid results when carrying out a globalized treatment of the information. In Spain, data are introduced manually by the contracting authorities. Considering the multiplicity of these bodies and their different levels of technical development, it is not difficult to find basic typographical errors in these databases, such as impossible dates, incongruous bids, or flipped numbers at identification data and bids.

24. To avoid these problems, the **CNMC runs a convenient automated process of filtering and cleaning manifest errors, as well as categorizing the data by quality levels.** Once there's a raw data backup copy, a cleaning process is launched amending most

[3] XML based web syndication format.

obvious mistakes[4] and unifying company identifiers and names. For the cases that cannot be easily corrected, we add a metadata tag noting the register quality level[5]. The last step is loading all corrected data on a new database. This will be repository on which all screening methods will be applied to detect bid rigging.

25. Secondly, data **quantity**. Despite national and international procurement regulation, often enough the compulsory publication of data is not effectively developed. For instance, until 2018 there was a lack of data integration for certain local and regional platforms.

26. It should be noted that, although the Public Sector Procurement Platform provides open datasets with a lot of public procurement information, it is only recently that CNMC gets structured information on loser bids and bidders. These data are essential to produce bid-rigging risk indicators (including statistical graphs) as well as for developing an automated and complete screening system. Nevertheless, these non-winning bids come only from 2018 onwards and solely for e-procurement procedures developed on the PSPP, which reduces data available to an estimated 40% of the total bids of Spain during that period.

27. To address this problem, the CNMC's Economic Intelligence Unit (**EIU) developed its own Natural Language Processing (NLP) algorithms and regular expressions** that process the documents attached to tenders and identify the losing bids and bidders, adding this information as new variables inside our public procurement database. This process is run for all tenders from 2012 to 2018 and for the ones that do not use the PSPP e-procurement platform.

28. Finally, - although not directly related to public procurement -, it is worth mentioning the possibility for Competition Authorities of taking advantage of other sources of information from public bodies and exploiting their potential. It is noteworthy the agreement signed between the CNMC and the Property, Mercantile and Real Estate Registers College (CORPME). The inclusion of CORPME's information about companies, owners, and shareholders, at CNMC's database will enable a more complete view.

3. CNMC's Unit Responsible for Data Treatment: The Economic Intelligence Unit

29. The correct and effective use of these data requires the creation of a differentiated and specific unit in charge of their treatment within CAs.

30. Thus, the increasing volume of data in open sources, together with the development of the new tools, pushes the search for profiles beyond those traditionally present in the CAs, such as lawyers or economists.

31. This realisation drove the creation of the Economic Intelligence Unit (EIU) within the CNMC's Competition Directorate in 2018.

32. In order to fulfil its task of applying advanced statistical techniques and incipient artificial intelligence to improve ex-officio detection of anti-competitive practices, especially in bid rigging cases, its team consists of a multidisciplinary group with

[4] Some of these mistakes can be easily solved such as misspelled company names or dates.

[5] This step creates an intermediate database and allows us decide on case basis about the convenience or not of using those records, in situations, for instance of lack of data.

economic, legal, statistical, mathematical and IT profiles, and extensive experience in competition matters.

33. Furthermore, thanks to the analysis of the data through business intelligence tools, this Unit supports the Competition Directorate´s decision-making in all types of cases, as well as improving the effectiveness of dawn raids, for example, by obtaining information with OSINT (Open-Source Intelligence) techniques.

34. Lastly, the EIU has several anonymous channels for citizen collaboration as it is the CNMC's whistle-blower programme contact point.

35. In relation to the CNMC's resources for data treatment, the data analysis carried out by the EIU is used both for purely ex officio detection and for the reinforcement of evidence gathering, sampling and testing of allegations in any sort of cases, irrespective of whether they originate in leniency applications, complaints, information from anonymous channels, contracting bodies[6] or regional competition authorities, etc. The following section will describe the main data processing procedures and examples obtained.

36. Finally, it is also important to highlight the role that international collaboration between NCA's plays in the promotion and development of new techniques and tools.

4. Data Use and Screens. CNMC's Experience

37. Considering that until recently the CNMC only had access to winning bids and bidders, it was not possible to develop an automatized system of digital screening tool and so that, every analysis was made on a case-by-case basis, using mainly statistical methods.

38. There are many different techniques and tools that allow a simpler and accurate data treatment, which have helped to detect and investigate anticompetitive conducts, as well as to gain in efficiency and effectiveness in the daily work of this Commission.

39. Their use has been mainly focused on bid rigging detection, but it has also proven useful regarding other horizontal agreements between companies as well as for the analysis of markets reported by whistle-blowers or informants.

40. A summary of some of the used tools is outlined below, from less to more sophistication.

4.1. Statistic Analysis:

41. After the described data cleaning process, there are many tools that, (without being ex officio screenings), make possible to identify anticompetitive agreements in markets already under investigation or where there are doubts about effective competition.

42. Statistical analysis is a simple way to obtain an agile visualization of the information, facilitating patters detection compatible with collusive conduct.

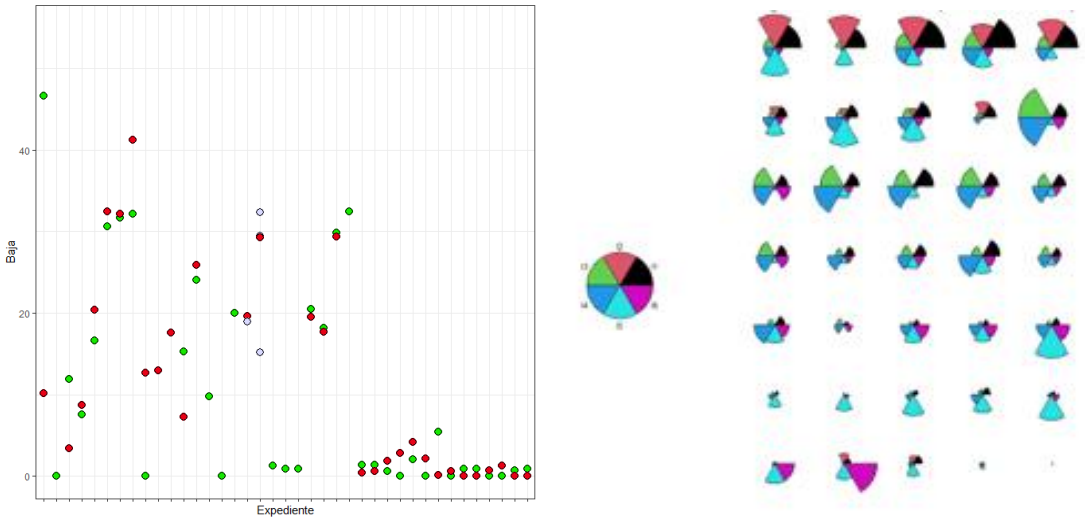
43. **Scatter plots** are a good example of this when used to confront bids with other variables such as the tender date, number of participants or technical score. Table 2 (left) compares with a scatter plot, bids of different tenders over time, identifying the companies

[6] According to articles 150 and 132 of the public procurement law, contracting bodies are obliged to send to the CNMC/ regional competition authority the possible cases of anti-competitive practices found in ongoing bids or ended procedures. This is an important starting point to some bid rigging cases.

with different colours. As shown, it clearly plots the moment when contacts between companies started.

44. Another statistical tool are **radial** diagrams. These diagrams synthesize graphically a set of variables for each observation. Once ranked by characteristics, they allow a quick visualisation and detection shared patterns between companies. Thus, figure 1 (right) shows how the companies represented in the third-row present similarities in relation to the selected variables.

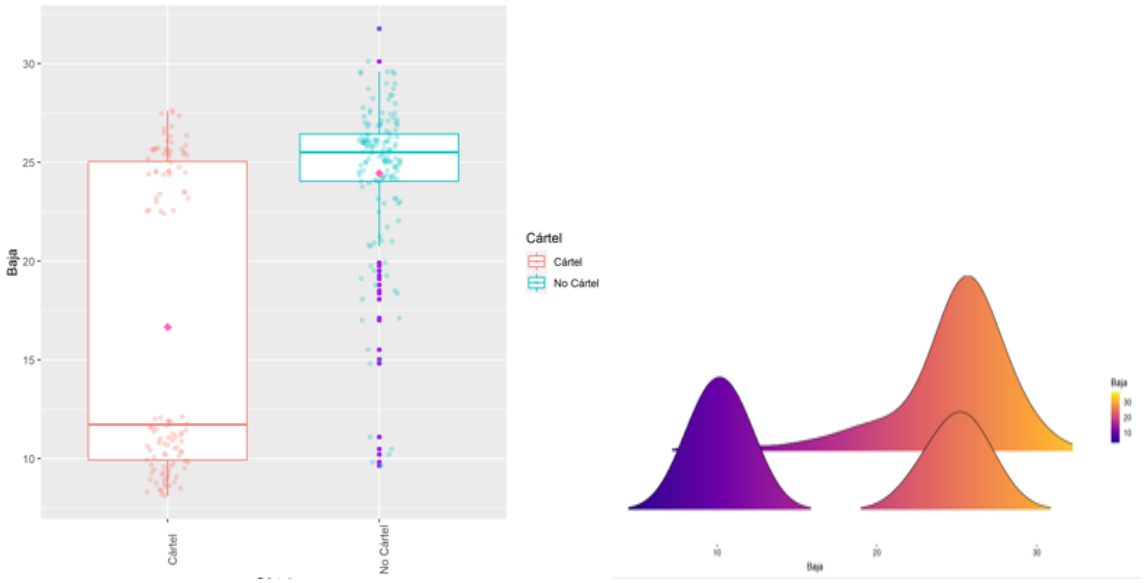
Figure 1. Scatter plots (left) and radial diagram with the companies sorted by characteristics (right)



Source: Anonymized CNMC cases

45. **Box plots graphs** are used to obtain a compilation of the distribution of different companies' bids and to detect similarities or differences between them. They are also very useful to compare cases where a group of companies supposedly members of a cartel with the rest of the firms and to detect atypical data. Figure 2, (left) shows the differences between the distribution of cartel firms bids (red) and the rest of the market (blue). The range of the non-cartel firms' low bids is wider and with many outliers (purple). However, in this specific case, cartel firms also had two clearly differentiated behaviour due to their collusive agreement scheme.

Figure 2. Box plots (left) and Density plots (right)

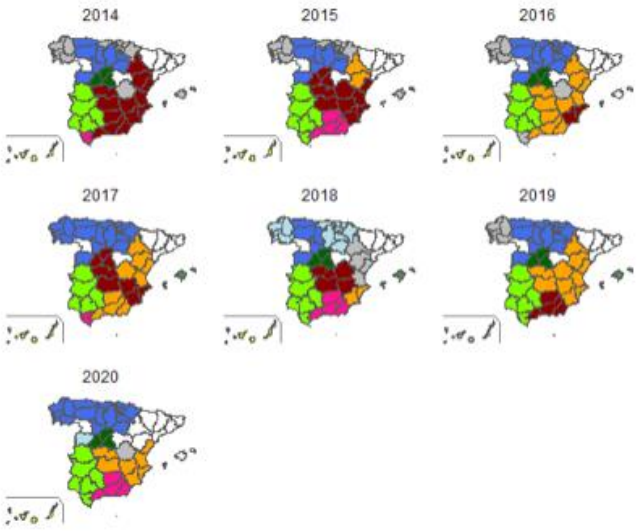


Source: Anonymized CNMC cases

46. **Density plots** complete the information of box plots studying the behaviour of bids in one or several tenders. They are especially useful to compare distributions of several datasets. The finding of differences between groups could be an indication of a cartel as shown in figure 2 (right), where is possible to observe a different distribution between cartel (first row mountains) and the remaining firms. As explained, this cartel scheme drives to a bimodal bids' distribution and with lower bids on average.

47. To analyse **allocation of lots** or tenders, there are some **association algorithms** with good results, such as the a priori algorithm that looks for relationships between enterprises without the need of having availability of losing bids. Also, in the specific case of **geographic market allocation**, maps (figure 3) give a quick view of the awarded company within different territories over the years.

Figure 3. Geographic Market allocations



Source: Anonymized CNMC case

4.2. Business Intelligence and indicators:

48. Business Intelligence (BI) has been used and considered very helpful since the origins of the EIU as general and embryonic screening method, especially in the absence of data on losing bids.

49. This type of software makes possible to plot data, compute indicators from different points of view (geographical, temporal, markets, etc.) and make filters based on different variables (contracting body, CPVs markets, subject of the contract, dates, etc.).

50. **First indicators** were developed considering the absence of losing bids and bidders. These were measures such as winning bids, number of participants or number of bidders per market. This type of screening provided an orientation for further research and the information obtained, such as graphs or tables, was used as support for research already in progress.

51. A typical result of bid rigging or collusive tendering is a concentrated market structure. That is, one or several companies winning a large part of contracts while competitors either abstain or only accompany the winning companies. It is possible to create indicators to measure market concentration such as the share of the total value of contracts awarded to the leading firm, or the addition sum of the market shares of the four top firms.

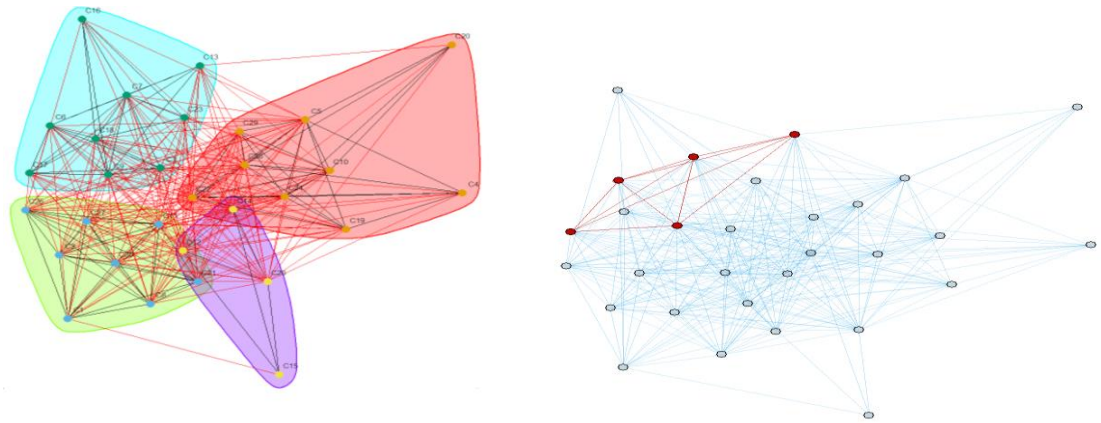
52. The current availability of **new data as the losing bids**, has led to the development of new indicators such as the **time lag** between bids submissions, **relative difference** between the winner and the second bid, **relative standard deviation or coefficient of variation**, or **relative bid range**. These indicators can be easily computed and can provide a lot of information.

4.3. Unsupervised learning algorithms:

53. New developments in artificial intelligence techniques have led to the availability of increasingly sophisticated tools for analysis. The EIU has used certain unsupervised learning algorithms, as well as other methods such as clustering algorithms. For the time being, these techniques have been used on a case-by-case basis, and with the idea of supporting the analyses carried out within a case.

54. Figure 4 (left), about application of unsupervised learning techniques that search for **clusters** of firms, shows companies that according to this technique would be more likely to be part of a cartel.

Figure 4. Cluster analysis (left) and network analysis (right) on public procurement data



Source: Anonymized CNMC case

55. On the other hand, other **network analysis techniques**, (figure 5, right), would look for groups of firms that, under some requirements, would be likely to be part of same cluster or group. For this purpose, it is important to focus on densely interconnected nodes compared to the rest of the network. The variables considered can be different, all of them internationally accepted as red flags in bid rigging detection.

4.4. gUIE.me: An instant public procurement search engine

56. As mentioned before, the **public sector procurement platform**, is the central database for procurement in Spain, containing also regional and local procurement data.

57. **Part of the information is structured** (winning bid and enterprise, initial budget, number of participants…) but **some other is not**, just accessible from the PDF, Microsoft Word, or other documents, that usually contain information as important as losing bids and bidders.

58. To make all these data (structured and not structured) available, the EIU has developed, entirely in-house, an instant **search engine** on any field in the database with a simple and friendly interface. This tool not only searches for structured data but also on any concept that appears in the attached documents whatever their format (Word, PDF, photographs...).

59. It also enables instant multiple selection criteria in different categories (type of contract, amount, contracting body, place of performance, etc.). Additionally, as it has been

developed in-house, the different weights (of relevance) given to each search field can be modified when returning the results.

60. In addition, **a name entity recognition algorithm** has been developed to read all the attached documents and find what is likely to be a non-winning bid and the bidder. Other NLP techniques are now under development, like a redefinition of the above mentioned named entity recognition, topic modelling and text classification algorithms.

61. This search engine also allows the results to be downloaded to Excel and the data can then be fed into a BI tool or statistical software.

62. Although this does not constitute per se a system of screens or red flags, it is a very important first step in that direction. It also facilitates the task enormously and allows for high increases in efficiency in the work of the EIU. Examples of its use are the delimitation of affected tenders in a conduct, the determination of their geographical or temporal scope, or the confirmation of replies to requests made to companies for other units of the Competition Directorate.

4.5. Graph based databases

63. A new field that EIU is studying is the detection of non-obvious relationships between tenders and bidders through the application of data science algorithms on graphs. For that, EIU is loading the traditional relational database into a new graph-based one.

64. Once all data will be in a graph database, there will be the possibility of running general screens similar to the ones run case-by-case described on chapter 4.3.

4.6. Balance of the use of these techniques. Next in the near future

65. Competition Authorities are well advised not to rely entirely on external factors such as incoming complaints or leniency applications. Ex officio detection is a fundamental complement to these traditional entry channels and contributes to incentivise their use triggering a virtuous circle of higher detection, higher reporting and higher compliance levels. Ex officio detection also acts as a deterrent to collusion for companies in different markets and allows NCA's to choose the sectors investigate.

66. The techniques presented in this contribution, along with many others, allow more in-depth and detailed investigations, comparing in many cases the behaviour of cartelised versus competitive firms, as well as summarising trends in public procurement in the selected markets.

67. The first and most important lesson learned is that it is vital to have a complete, clean, and reliable dataset to perform a proper screening. This dataset should as far as possible, include data from all bids (awarded or not) to develop more complex and consistent analyses. This also allows to study the anticompetitive agreements from a general point of view instead of a case-by-case basis.

68. On the other hand, to use supervised learning techniques, it is also necessary to have a repertoire of sanctioned cases for collusive agreements. If that's possible for the CA, these new techniques will lead to significant improvements in the efficiency of detection of suspicious behaviours.

69. In Spain, business intelligence indicators and the statistical techniques used when there was no information on losing bids have yielded good results. This has been a good starting point which has often made it possible to validate the behaviour observed. Nevertheless, all this work was made on case-by-case basis.

70. Furthermore, the incorporation of automated processes, such as the creation of specific algorithms, was a major step forward in terms of efficiency at the CNMC.

71. The availability of data from all participating bids and bidders opens the range of available techniques and analyses, especially derived from the development of machine learning procedures and the exploration of new tools such as graph analysis. And, it will permit conducting general analysis on pure ex officio detection. Both issues are among the objectives of the CNMC's EIU in the near future.

5. Main conclusions

72. Increasing transparency, especially in public procurement data, demands new approaches and actions by Competition Authorities.

73. Firstly, the large volume of data in public procurement databases, even if it is a good starting point, it still depends on subsequent cleaning and quality categorization processes.

74. Second, it is essential to be able to implement statistical instruments and be aware of new IT developments, which can highly help in the visualization and discovery of new bid rigging cases.

75. CAs must adapt their staff recruitment and efforts to boost these developments, including the creation of multidisciplinary teams.

76. Since 2015, the CNMC has been making its way into the world of ex officio detection of bid rigging with the creation of a procurement public database, from the download of certain information from other public procurement sources. The creation of the Economic Intelligence Unit (EIU) in 2018 has been aligned with this objective. In this regard, different techniques have been gradually used to take advantage of public databases. First, on a case-by-case basis, mainly those related to statistical analysis and indicators in business intelligence platforms. In these cases, losing bid and bidders were introduced manually. Subsequently, by developing task automation algorithms and natural language processes, which facilitate the daily activity of the EIU and increase its efficiency. More recently, based on the availability of losing bids, the door is open to new developments based on machine learning and graph analysis techniques.

77. Sharing new tools and improvements with other CAs and other public bodies is also a mutual benefit to be exploited.

78. To sum up, ex officio detection, specially related with public procurement, is called to be one of the pillars of CAs in the next years. So, the importance of combining artificial intelligence, all new tools, and specific techniques such as data use and screens with human intelligence is crucial to achieve fruitful results. In the years to come, we will have to continue using all those elements. This makes the determination to improve the use of these tools greater.