Organisation for Economic Co-operation and Development

DAF/COMP/WP3/WD(2022)37

Unclassified

English - Or. English

8 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 Romania

28 November 2022

This document reproduces a written contribution from Romania 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]

JT03506962

Romania

1. Background information and preliminary work

1. By way of background, Romania has a thriving, rapidly growing software industry representing 6% of GDP, built on the low personal taxation regime the country has adopted for software engineers to stop the brain drain in this sector. We have regional hubs for technology and many of the familiar multinationals have regional centers and even global centers in Romania. The collective industry umbrella, the Romanian Software Industry Association, is targeting 10% of GDP by the year 2025.

2. On the other hand, the public sector is playing catch-up in terms of keeping up with the speed of development of the private sector, which has been one step ahead already in the race towards going digital, since it is more agile and better equipped financially and in terms of human resources. Their advancements in technology can make collusion harder to detect so we had to ramp up continuously our technological capabilities to create equal fire power when confronting more sophisticated cartels.

3. The standard of proof required by courts can no longer be upheld just by following the paper trail of documents, the competition authority is now required to comb through vast amounts of data and to correlate them in search of proof of cartels or collusion.

4. So, using tech driven investigative tools has been a favorite topic for internal discussions ever since 2016-2017, and RCC staff has had several debates regarding the various tools which could be employed in various screening scenarios.

5. Furthermore, there were several instances where the findings of studies using econometric analysis were used to support the detection of a cartel agreement.

6. The team of our Chief Economist Unit used cluster analysis, statistical hypothesis, normality and symmetry tests to highlight improbable results in a series of public procurement procedures, which later on were subject to a full-on infringement investigation and sanctions were applied for bid rigging.

7. However, such analytical methods need resources to process large amounts of data and therefore are not easily applied if not supported by digital tools.

8. This is why the plan was to use Big Data technologies, which RCC was studying as part of a market study. The authority had an initial attempt with a smaller project on bid rigging, which was meant to be developed in partnership with the public institution managing the e-procurement system (SEAP). The interface issues, teething problems with the software platform and the difficulty of integrating the data proved to be the type of hurdles which were at that time too difficult to surmount without consistent financial resources.

9. Using the experience garnered from the earlier project, RCC started the preparation for a bigger, EU funded project with many more workstreams integrated and made to work jointly. In parallel, we started working on the list of indicators which it was thought that might be used by the platform and a small team of academics[1] was contracted to help us draft a methodology.

[1] RCC would like to acknowledge Mihaly Fazekas, Silvia Fierascu, Bence Toth and the team of Government Transparency Institute (GTI) for their contribution to the RCC BD Platform.

10. Their input and experience were invaluable and constituted the foundation on which the RCC BD platform was built.

2. Main components of the Big Data platform

11. The operational case management component, which represents one of the main sources of internal data for the specific analyzes carried out through the BD Platform, was completed during 2020. Also, during the same year, the integration of this component with the institution's Electronic Archive was made; together with the facility of electronic signing of all documents on the workflows, this step represents the main "engine" behind digitizing the day-to-day activity that will lead to a “paperless” competition authority.

12. During 2021, the development of the Interoperability Platform was completed, a platform that integrates data sources of external information obtained from partner institutions such as the National Trade Register Office (ONRC), the National Fiscal Authority (ANAF), the Public Procurement Electronic Portal (SEAP), the National Council for Solving Complaints in public procurement (CNSC) and the Ministry of Justice (MJ), which are capitalized as such or within the advanced analysis process of Big Data platform.

13. The main component of the project, the Big Data Platform, was completed at the end of 2021. This component allows the specialized staff of the RCC to access/use data for the initiation and development of cases based on 5 analysis modules: cartel screening, bid rigging detection, sector investigations, mergers and structural and commercial links between enterprises.

14. At the same time, the platform allows advanced analyzes on the entire volume of data, making correlations on a predictive basis between actions and events which serve two objectives: predictive analysis for early detection of anticompetitive practices and support for current investigations.

15. Given the specificities of the data collected by the platform, its development required an update of the initial methodology developed in-house, based on quantitative and qualitative indicators specific to the analysis modules. Some of the qualitative indicators could not be used, since they needed considerable human input and correlations of data which was unavailable and some quantitative indicators were either adjusted to match the data or aggregated to simplify the process.

2.1. Resources

16. First of all, significant financial outlays were required. The implementation of the Big Data platform, along with its supporting internal Case management, electronic archive and interoperability platform, was funded via an EU structural funds project.

17. Next, RCC employed a massive number of staff in this enterprise when implementing the project and it is expected to continue to use a significant team for sustainability; moreover, a team of external consultants, both academics and IT developers, was employed through the project. RCC’s own human resources needed an upskill to be able to work with the new systems, so the project included a training component for each of the three main components, totaling more than 650 hours of training.

18. Last but not least, one of the most valuable resources was time. It took RCC around two years to come up with the idea, identify funding and design the project and four years for the implementation.

2.2. Sources of data

19. The data which is part of our online architecture comes from the National Trade Registry, the fiscal authority, the electronic bidding platform and the interface with MJ.

20. Apart from the external sources, the platform also utilizes most of the data RCC already stores since we automated our internal workflows and digitized our 20yrs archives and gave access to the platform to those sources of information. The daily press report which RCC receives as part of an external contract is also uploaded in the platform. In addition, RCC already had some digital databases from previous projects which harbor important data, such as the Price Monitor app for fuels and food products and the State Aid Registry.

21. In the future, once RCC staff is more comfortable with the use of the platform, a potential avenue worth exploring is to upload investigation documents into private and secure workspaces, where more advanced operations could be performed, such as text analytics.

2.3. BD platform features and early results

22. The BD platform allows for two types of screening: automatic and manual.

23. The automatic screens feed 10 types of reports (for example cluster analysis for procurement procedures, financial statistics on company level, market shares at county/national level, top 10 companies etc.) and 5 dashboards (sector inquiries, bid rigging, cartels, mergers and structural and commercial links between enterprises).

2.3.1. Example of a cluster analysis for procurement procedures

24. Detailed below is an extract from an unsupervised cluster analysis for a sample of public procurement procedures.

25. The indicators used in the example analysis are:

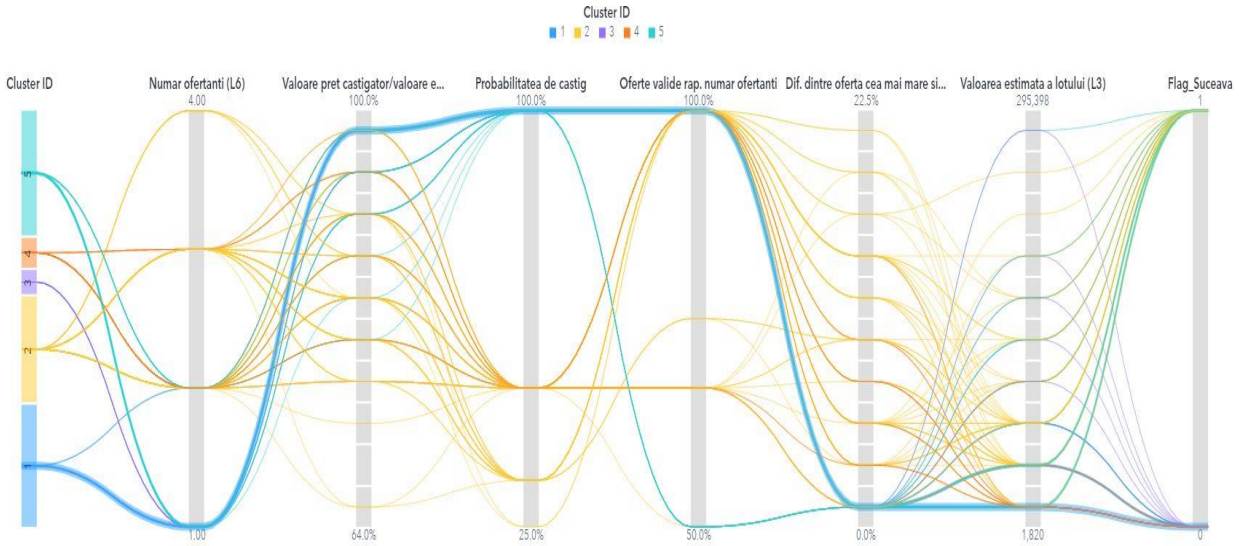
- Estimated lot value
- Probability of winning
- Winning price value/estimated value
- Number of bidders
- Valid offers related to the number of bidders
- The difference between the highest and the lowest offer, compared to the estimated value
- The difference between the best offer and the second-best offer, compared to the estimated value

26. The highly correlated variables were removed from the analysis, the outliers from the P1 and P99 percentiles were removed and the data standardized. Following these stages, the example uses a final base of analysis of 297 public procurement procedures from the CPV code 77211100-3 – Logging services.

27. In order to group procedures/objects with similar characteristics into classes, the K-Means partitioning algorithm was used as a clustering algorithm. This algorithm chooses k cluster centroids and assigns the objects to it by choosing the centroid closest to the respective object. As the objects are assigned to a cluster, its centroid can migrate, so the

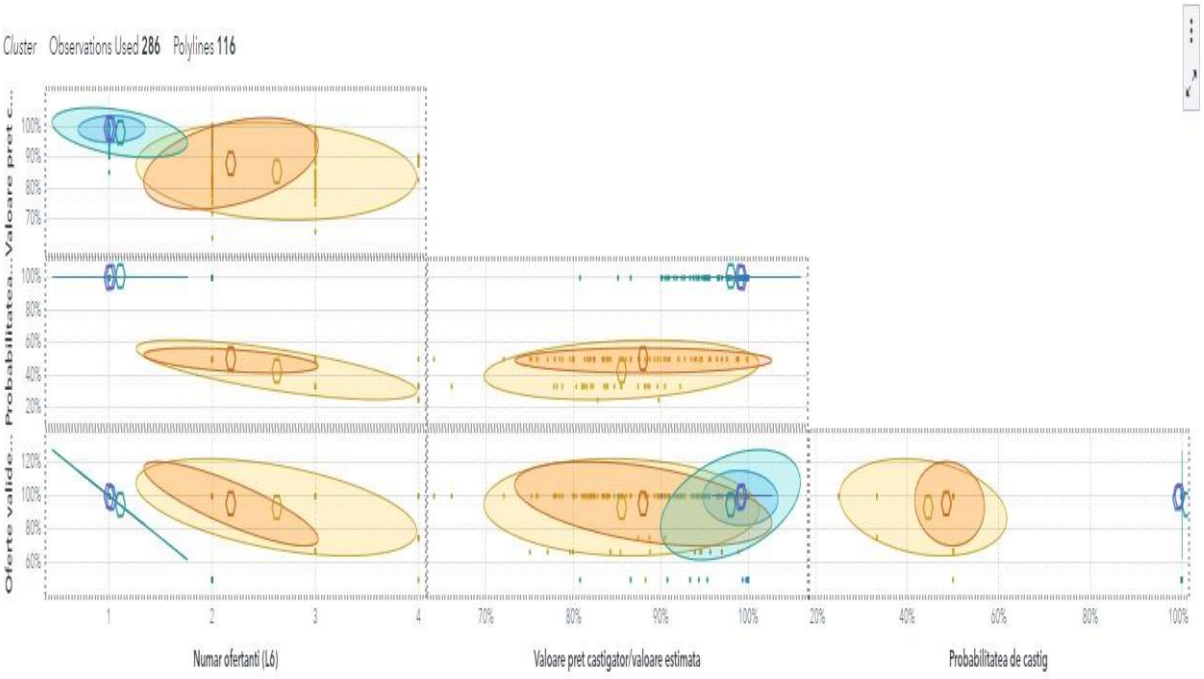
algorithm will be repeated until the centroids from two successive steps do not differ significantly. K represents the number of classes intended to obtain after applying the method. In this example 5 clusters were created. Each cluster is characterized by the indicators based on which they were built (see image below for a standardized average graph showing the color-coded clusters and the indicators).

Fig.1 Standardized average graph



28. The dashboard offers multiple graphical representations and possibilities to group the indicators, to facilitate the identification of previously undetected patterns or to further explore procedures in a specific cluster (see below).

Fig. 2 Clusters grouped using sets of two indicators



29. This clustering method allows for automatically discovering natural grouping into data. Later, experts can investigate further a selected sample of relevant procurement processes for manual review.

2.3.2. Considerations on alerts and early results

30. It is true though that in most cases the automatic results of screening will serve only as a motivation for the case handlers to further explore the red flags raised and to manually investigate. The findings of such screens will seldom constitute legal proof of the infringement without correlating additional evidence. This why the platform is constructed with a general screening area, where the dashboards are also located, and where users can view reports, alerts, indicators, companies and define markets. From the general screening area, the user can then refine its search to visualize the network for the selected companies, the automatic indicators computed for the respective companies and/or markets and add manual alerts. These types of analyses can then be migrated to a private investigation area, where searches can be manually refined further.

31. Automatic alerts are constructed on scores, which are computed based on different algorithms for each alert and market indicator. If the score is higher than a certain threshold, an alert is being generated. The work which is currently underway is testing the alerts and refining them accordingly by either adjusting the scores or the algorithms generating the scores, in order to reduce the number of false positives.

32. Even in this initial stage of utilization, the red flags raised in the merger area already delivered some encouraging results, since algorithms in this section were designed according to the step-by-step procedures for merger analysis.

33. The red flag system for mergers analyzed differences in company data (business register data from the Trade register and financial data from the Fiscal authority) recorded from September until December 2021, whereby the September dataset was considered the baseline. The analysis raised 10 alerts of apparently unnotified mergers. These alerts were further analyzed and, out of 10 transactions, 3 had already been notified at the RCC, which validated the process, and 4 transactions were found to be false positives, because a more in-depth network analysis showed them to be intra-group transactions. However, 3 transactions were mergers which were already put into practice and had not been notified to the RCC. These 3 mergers are currently under investigation for failure to notify.

34. The sections of the platform dedicated to sector inquiries and to links between undertakings are also areas where results are also more within reach for regular users. For example, this year's RCC report on the state of competition in essential economic sectors used the data aggregated by the platform to assess the profitability and capitalization indicators of companies on the Romanian market for 2017-2021.

3. Challenges

35. Some of the challenges are already obvious, as presented in the earlier chapters – money, people, data and time. The main obstacle to digitization is the lack of interoperability of information systems of public institutions. Another problem is the lack of IT specialists in public institutions, due to less attractive salaries. Also, access to the latest technologies is limited by financial resources, so most of RCC investments in IT were made by attracting EU funds. As such, acquisitions needed to be made on a larger scale (since EU projects cannot be initiated so frequently) and imply all the requirements and resources needed to manage and implement an EU-funded project of a large magnitude. If lack of IT specialists would not have been an issue, digitization could have been done with

open-source software or in-house developments of existing software. Since the RCC IT eco-system is based mostly on COTS software, costs for acquisition and ulterior maintenance are high and need constant budgeting efforts.

36. But other challenges came once the project was completed. One challenge is data management, since the datasets need to be updated on a regular basis and big administrative data has errors that need to be cleaned up. RCC experts have had to explore the specific datasets to find major types of mistakes and transform them into a standardized form.

37. For instance, a specific company may appear with different spellings: abbreviated name, long name, misspelled name, with or without legal form etc. This causes two practical problems. First, a user wants to filter the data for a specific company, so multiple filters must be applied to find all relevant observations which makes the process more complex unnecessarily. Second, merging of procurement data to other databases such as financial data is especially hard without IDs when a specific company has more names meaning more IDs.

38. Moreover, if legal considerations occur when using platform results, the experts need to document and justify the methodology of cleaning data, since the cleanup implies an intervention on the original dataset.

39. It is true though that in most cases the automatic results of screening will serve only as a motivation for the case handlers to further explore the red flags raised and to manually investigate. The findings of such screens will seldom constitute legal proof of the infringement without correlating additional evidence. Therefore, even if interventions on the original dataset may raise legal considerations, such interventions can be documented and justified.

40. The algorithms designed need testing in real life, the RCC staff needs to integrate the platform in their daily interaction so they gain the confidence to use it, and the more refined analysis needs advanced analytical skills which are scarce. So, building them in the future might be one of the areas for institutional improvement.

4. Lessons learned and going forward

41. Acquiring entire structured databases from other public authorities has been an obstacle due to GDPR concerns and this stifles interoperability. RCC was already using before an interoperability platform which allowed to interrogate databases of other institutions. The inquiries were however specific to a particular company, it did not involve large data sets. The Big Data platform, on the other hand, needs access to entire sets of data to screen and issue red flags. To alleviate GDPR concerns as to why access to privileged information of so many companies is needed in one go, RCC BD team had to carefully justify the process and explain the legal underpinnings.

42. Maintaining and improving data sets is an ongoing exercise - long-term commitment is needed and resources allocated. Both for data management and analysis, analytical needs are too specific, so it is essential to invest into data analytical skills that go beyond a mere understanding on how to use the BD platform. This can pose a problem for the Romanian competition authority since the salaries in the public administration are not attractive enough for data scientists and not even for data analysts. The alternative is to train its own resources, but that takes time, or to outsource, which is expensive and poses serious security and confidentiality issues.

43. On a more positive note, one of the biggest and immediate added value of the BD platform is the quick access to integrated administrative datasets. Since the mere utilization

of the BD platform is relatively straightforward with the trainings already provided to our staff, we feel that even without advanced analytical instruments the platform is a very powerful addition to the RCC investigative toolset.

44. Regular implementation of a detection method based on simple screens may hopefully have an additional deterrent effect. If trying to "beat" the detection method is still possible, it may increase the coordination costs among cartel participants and make cartelization more difficult. This added value of potential deterrence could be taken into account when considering publicizing the existence and use of digital screening in the authority, even if there are concerns that, if the markets are aware, companies will try to hide data. In the case of RCC, most of the datasets used are public, either structured or unstructured, so obscuring potential indicators of competition infringement is not an easy strategy.

45. Besides the GDPR issue, the need for specific cooperation between competition authorities may also be an avenue worth exploring. The ECN WG on Digital Investigation and Artificial intelligence has already taken some steps in this regard, and although traditionally there is a natural reluctance to share intellectual products which required significant investments to obtain, sharing trained algorithms could save a lot of time and resources and help overcome some of the challenges presented above.

46. On the RCC part, the Romanian authority is definitely open to sharing our work and learning from others. Sharing national data may not have relevance for the work of other competition authorities except in cross-border cooperation and there are significant limitations related to confidentiality, but sharing historical datasets and our software architecture could be useful for testing and training algorithms. If we would be able to learn from the others and incorporate new screening criteria, which have already been tested and proven valuable, in our collective screening toolbox, we might be able to save resources and keep up with what will probably be a digital shield and sword competition between NCAs and companies.