OECD Public Governance Reviews

Countering Public Grant Fraud in Spain

MACHINE LEARNING FOR ASSESSING RISKS AND TARGETING CONTROL ACTIVITIES

OECD

OECD Public Governance Reviews

Countering Public Grant Fraud in Spain

MACHINE LEARNING FOR ASSESSING RISKS AND TARGETING CONTROL ACTIVITIES

OECD
BETTER POLICIES FOR BETTER LIVES

The project was co-funded by the European Union via the Structural Reform Support Programme (REFORM/IM2020/006). This publication was produced with the financial assistance of the European Union. The views expressed herein can in no way be taken to reflect the official opinion of the European Union.

Please cite this publication as:
OECD (2021), *Countering Public Grant Fraud in Spain: Machine Learning for Assessing Risks and Targeting Control Activities*, OECD Public Governance Reviews, OECD Publishing, Paris, *https://doi.org/10.1787/0ea22484-en*.

ISBN 978-92-64-77835-1 (print)
ISBN 978-92-64-55436-8 (pdf)

OECD Public Governance Reviews
ISSN 2219-0406 (print)
ISSN 2219-0414 (online)

Photo credits: Cover © 2-Q STOCK/Shutterstock and Gearings image © OECD, designed by Christophe Brilhault.

Corrigenda to publications may be found on line at: *www.oecd.org/about/publishing/corrigenda.htm*.

Foreword

Fraud in public grant programmes diverts taxpayers' money away from essential services and reduces benefits for well-meaning recipients. When individual beneficiaries, private providers or government officials defraud grant programmes, they not only undermine the integrity of the programme itself, but they also risk eroding trust in government. In the wake of the COVID-19 pandemic, marked by a high volume of accelerated spending, fraud risks have become a pressing concern for governments worldwide.

In this environment, public control and audit bodies play a vital role to ensure money is well spent and vulnerabilities are spotted and addressed quickly. In Spain, the General Comptroller of the State Administration (*Intervención General de la Administración del Estado*, IGAE) is at the forefront of the country's efforts to prevent and detect fraud. Its oversight mandate focuses on high-risk areas where fraud commonly lurks, including the focus of this report, public grantmaking.

Taking a risk-based approach is essential for directing limited resources. To this end, modern control bodies like the IGAE are increasingly reliant on data and analytics as fundamental tools for assessing risks. By leveraging data and analytics for enhancing control and audit processes, the IGAE is better equipped to identify risks and target its resources where they will have the most impact. This report reflects the initiative and commitment of the IGAE to tap into cutting-edge approaches, including state-of-the-art methodologies in artificial intelligence and machine learning.

This document was reviewed by the OECD Working Party of Senior Public Integrity Officials (SPIO) on 01 November 2021and declassified by the Public Governance Committee on 23 November 2021. It was prepared for publication by the OECD Secretariat. The project was co-funded by the European Union via the Structural Reform Support Programme (REFORM/IM2020/006). This publication was produced with the financial assistance of the European Union. The views expressed herein can in no way be taken to reflect the official opinion of the European Union.

Acknowledgements

Under the direction of Elsa Pilichowski, OECD Director for Public Governance, and Julio Bacio Terracino, Acting Head of the Public Sector Integrity Division, this project was led by Gavin Ugale, who also drafted Chapter 1 and edited throughout the report. Dr. Mihaly Fazekas, Assistant Professor at Central European University and Scientific Director at the Government Transparency Institute, designed the machine learning model described in Chapter 2 and drafted Chapter 3, with support from Viktoriia Poltoratskaia. Varun Banthia supported the research and drafting process. Meral Gedik, Andrea Uhrhammer, Laura Völker and Elisabeth de Vega Alavedra provided editorial assistance. Charles Victor and Aman Johal provided administrative support, and Balazs Gyimesi contributed to communications and publication design.

The OECD is grateful to colleagues in the General Comptroller of the State Administration of Spain (IGAE) for their fruitful co-operation and leadership. In particular, the OECD would like to thank Isabel Silva Urien and Ismael García Cebada, as well as their teams, including Carlos Collado Molinero, Pablo Lanza Suárez and Israel Barroso Pérez. The OECD would also like to thank Ciresica Feyer of the European Commission's Directorate General for Structural Reform Support (DG REFORM) for her guidance throughout the project and input on the draft report.

Table of contents

FIGURES

TABLES

Follow OECD Publications on:

http://twitter.com/OECD_Pubs

http://www.facebook.com/OECDPublications

http://www.linkedin.com/groups/OECD-Publications-4645871

http://www.youtube.com/oecdilibrary

http://www.oecd.org/oecddirect/

Abbreviations and acronyms

AEAT	State Agency of Tax Administration *Agencia Estatal de Administración Tributaria*
BDNS	National Subsidies Database *Base de Datos Nacional de Subvenciones*
COSO	Committee of the Sponsoring Organisation of the Treadway Commission
EU	European Union
IGAE	General Comptroller of the State Administration *Intervención General de la Administración del Estado*
OIP	Office of Finance and Information Technology *Oficina de Informática Presupuestaria*
OLAF	European Anti-Fraud Office
ONA	National Audit Office *Oficina Nacional de Auditoría*
ONC	Public Accounts Office *Oficina Nacional de Contabilidad*
SAIs	Supreme Audit Institutions
SHAP	SHapley Additive exPlanations
SNPSAP	National System of Publicity for Subsidies and Public Aid *Sistema Nacional de Publicidad de Subvenciones y Ayudas Públicas*
TGSS	General Treasury of Social Security *Tesorería General de la Seguridad Social*

Executive summary

Fraud is by nature a hidden activity, so how can authorities detect and mitigate risks effectively? This report identifies ways for Spain's General Comptroller of the State Administration (*Intervención General de la Administración del Estado,* IGAE) to tackle this challenge, using state-of the-art machine learning models, and effectively target its control activities to the highest fraud risks found in public grants and subsidies.

There are few reliable figures for country-level fraud levels, given the complexities of measuring something that is intentionally concealed. Often countries rely on broader proxy measurements, such as the extent of reported irregularities in specific programmes or sectors. Nonetheless, available figures suggest considerable challenges and fraud risks for governments. For instance, in many countries that assess the extent of fraud in social benefit programmes, such as France, the United Kingdom and the United States, estimates of fraud reach into the hundreds of millions of euros. In its *32nd Annual Report on the protection of the European Union's financial interests: Fight against fraud 2020*, the European Commission reported EUR 375 million as fraudulent linked to revenue and expenditures. Fraud levels in EU Member States are likely to be much higher when taking into account national funds and public expenditures.

Control bodies, such as the IGAE, are on the frontline of the governmental efforts to prevent and detect fraud. They have a unique government-wide vantage point to spot fraud risks and strengthen the effectiveness, efficiency and economy of government spending through ex-ante and ex-post evaluations. To do this job effectively in a digital age, oversight bodies face considerable pressure to keep pace with both evolving risks and new technologies. In Spain, like other EU Member States, the Recovery, Transformation and Resilience Plan puts specific emphasis on the need to improve the mechanisms and tools to prevent, detect and correct for the risks inherent in public grants, including fraud, corruption, conflicts of interest and double funding.

In this context, the IGAE and the OECD, with the support of the European Commission, worked together to identify ways for the IGAE to strengthen its assessments of fraud risks in public grants and subsidies, with the ultimate goal of more targeted control activities. The project focused on supporting the IGAE in making use of existing data, and identifying ways that it could expand its analysis to consider new data sources, fraud risks and methodologies. Chapter 1 briefly describes the IGAE's context and mandate, as well as its approach for assessing risks and planning its control activities. It also highlights several overarching considerations for the IGAE to enhance its use of data and analytics, regardless of whether it adopts the machine learning model in Chapter 2, with a focus on assessing grant fraud risks. They include:

- Strengthen data governance and management for assessing grant fraud risks, starting with quick wins like improving its data dictionaries, clarity of unique identifiers and data controls specifically for fraud risk analysis.
- Build capacity for data-driven risk assessments, in particular, developing structured datasets and ideally a capacity that brings together expertise related to grant-making processes, fraud risks, analytics and visualisation.
- Beware of pitfalls concerning composite risk indicators as well as biases, which can include biases in machine learning models.

Chapter 2 lays out a proof-of-concept for a data-driven risk model for the IGAE to adopt in part or in its entirety. The methodology makes use of data at the IGAE's disposal, thereby implicitly accounting for the IGAE's current context. The machine learning model accounts for risks across the grant cycle to the extent the data allowed. The process of developing the proof-of-concept for the risk model led to several insights and the identification of areas for improvement, including the following:

- Establish a ready-made dataset for fraud risk identification, which this project has started as a pilot and can form the basis for future risk analysis with fewer investments in resources and time.
- Expand the IGAE's use of indicators across the entire grant cycle, including enhancing data and indicators that go beyond descriptive features and reveal behaviours (e.g. conflicts of interest).
- Invest in continuous improvement of the machine learning risk model, if adopted, to ensure a truly random sample, account for new data and risks, and address biases, among other considerations.
- Consider network analyses and making use of a broader set of methodologies, including those that take advantage of company data.

Finally, Chapter 3 offers a roadmap for complementing existing IGAE grants data in order to improve its risk assessment models. Specifically, it outlines datasets that can be matched to existing IGAE grants data, thereby enhancing the analytical sophistication and improving the precision of the IGAE's risk assessment. The guidance and recommendations in the report draw from OECD fact-finding interviews, analyses of the IGAE's context and available data, the experiences of other government entities and international leading practices.

1 Risk-based control in Spain: A foundation for improved analytics

This chapter provides an overview of the General Comptroller of the State Administration (Intervención General de la Administración del Estado, IGAE) and its oversight of public grants and subsidies in Spain. It describes the IGAE's current approach to risk-based planning, and highlights preconditions and considerations for the IGAE to advance its use of grant data for assessing fraud risks. This includes considerations and recommendations for ensuring effective data governance and data management, as well as building capacity for using machine learning models.

Introduction

The General Comptroller of the State Administration (*Intervención General de la Administración del Estado*, IGAE) exercises internal control over the economic and financial management of the Spanish government. This includes the central government, dependent autonomous bodies in the central administration, state entities under public law, and public business entities. As part of its mandate, the IGAE carries out control activities to ensure sound financial management and compliance with, inter alia, the Organic Law of Budgetary Stability and Financial Sustainability (*La Ley Orgánica de Estabilidad Presupuestaria y Sostenibilidad Financiera*), the General Law of Grants (*General de Subvenciones*) and legislation of the European Union (OECD, 2014[1]). The IGAE also investigates high-risk areas for potential fraud and irregularities, including public grants and subsidies that support the achievement of Spain's public policy goals.[1] d

The public grants and subsidies that the IGAE oversees amounts to EUR 89 860 million of the total annual budget, and involves thousands of beneficiaries and entities. Given the size of this audit universe and the high volume of transactions related to grant disbursements, the IGAE has developed a risk-based approach to help it target the highest risks and manage its resources efficiently. The risk criteria the IGAE has developed takes into account the potential for fraud and irregularities based on predetermined criteria, as described in this chapter.

The IGAE has developed a risk-based approach for its control activities, but opportunities remain for it to make better use of existing data and new methodologies to further target its resources to high-risk areas. This chapters explores key considerations for the IGAE to advance its use of data and analytics. As described in Chapter 2, the project focused on a specific methodology, inspired by machine learning; however, the considerations in this chapter are more generally applicable regardless of the technique or methodology. Moreover, while the OECD project focused on enhancing detection of grant fraud risks, the insights from this chapter and the next are applicable to other types of risk analyses when reliable data were available.

Overview of the grant cycle and the IGAE's oversight responsibilities

The IGAE follows a decentralised operating model, with three central service functions delivering its core areas of responsibility at the central government level, including the National Audit Office (*Oficina Nacional de Auditoría*, ONA), the Public Accounts Office (*Oficina Nacional de Contabilidad*, ONC), and the Office of Finance and Information Technology (*Oficina de Informática Presupuestaria*, OIP). The IGAE has both ex-ante and ex-post responsibilities:

- *Ex-ante*, by controlling, before they are approved, activities in the performance of expenditures, revenues, payments and investments, or the general application of public funds, to ensure that management complies with all applicable laws. Ex-ante control is therefore preventive, taking place prior to the adoption of various economic activities, such as contracts, grants, agreements, charges and payroll, among others. It can be exercised in a limited fashion, by examining certain key aspects of economic and financial activities, or it can be exercised in full, by examining all documentation linked to a financial act.
- *Ex-post*, by verifying on an ongoing basis the status and operation of public sector entities to verify compliance with applicable regulations and that management conforms to the principles of sound financial management, in particular the achievement of the objective of budgetary and financial stability. The IGAE performs public audits, which can take various forms, including annual accounting regularity audits (reviewing accounting information to verify its relevance to accounting standards), compliance audits (verifying the legality of budget management, procurement, personnel, revenue and grant management) and performance audits (examining operations and

procedures to assess financial and economic rationality and relevance to the principles of good governance as a means to detect deficiencies and make recommendations to correct them).

The main results of the IGAE's audits are summarised in an annual report. When infractions are detected that could result in corruption or fraud, a special report is sent to the Ministry of Finance and Civil Service (*Ministerio de Hacienda y Función Pública*) in addition to the controlled entity. This reporting promotes improvements over time in the techniques and procedures of economic and financial management as recommendations are acted upon. There are collaboration mechanisms between the IGAE, comptrollers of the autonomous communities and local comptrollers (OECD, 2014[1]).

The mandate to control grants rests predominantly with the ONA and the IGAE's Grants Monitoring and Reporting Division (*División de Control e Información de Subvenciones*). However, there are "Delegated Interventions" (Intervenciones Delegadas) at regional or provincial levels, as well as those integrated into ministries and public sector organisations. These entities act as financial controllers and are responsible for ongoing monitoring of financial controls and public internal audits (IGAE, 2020[2]). In addition, Delegated Interventions are tasked with exercising control on public expenditures to third-party organisations, including public grants, loans, and guarantees.

Article 140.2 of Law 47/2003 General Budget (*General Presupuestaria*) gives the IGAE the power to execute internal control of the public sector with full autonomy vis-à-vis the authorities and other entities whose management it controls (Government of Spain, 2003[3]). This power includes the authority to implement control activities related to beneficiaries of grants, in accordance with articles 141 of the 140.2 of Law 47/2003 (Government of Spain, 2003[3])and 44 of Law 38/2003 (Government of Spain, 2003[4]) General Grants (*General de Subvenciones*) (IGAE, 2020[5]).

The administration or granting body oversees the general processes of each phase of the grant cycle, and the granting body is responsible for oversight of the beneficiary to ensure compliance with the terms of the grant. For instance, early in the grant cycle, concerns addressed can include whether the grant-awarding agency has correctly generated the grant, and if the grant was awarded, applied and checked accurately. In addition to the oversight by the awarding agency, external bodies, such as the legislature, court of auditors, or other audit bodies, provide additional oversight and controls. The IGAE's model for overseeing the public grants spans the grant cycle, which generally consists of the following phases:

1. *Competition*—The conditions a grant beneficiary must meet in order to receive a grant are defined by the awarding agency. The grant-awarding agency approves these conditions, and a request for applications is opened.
2. *Selection*—Candidates are reviewed and selected based on the quality of their applications against the original set of criteria.
3. *Grant execution*—If the applicant already meets all of the requirements for the grant, or the grant has a provision for an advance, a payment is then made and the beneficiary must begin the activities required by the grant immediately.
4. *Monitoring*—After the requirements of the grant are delivered, the beneficiary must present a justification of how funds were spent. The granting agency will review this justification and adjudicate whether any final payments need to be made or if funds need to be clawed back. The latter can take place if an activity was not completed as stipulated by the initial grant.

The IGAE's investigations and control activities across the grant cycle serve different purposes. For instance, they aim to verify whether the beneficiary obtained and is managing the subsidy correctly. The IGAE may also assess whether the subsidy was justified, and that the operations covered by the subsidy are legitimate and real. The IGAE will also investigate whether the beneficiary had failed to report material facts to the administration that could affect the financing of the subsidy. (IGAE, 2020[5])

Transparency is emphasised in laws and in practice. For instance, Royal Decree 130/2019 (Government of Spain, 2019[6]) integrates many of the aforementioned laws and reiterates the provisions concerning

transparency, access to public information and good governance. This law, along with EU 651/2014 (European Union, 2014[7]) and 702/2014 (European Union, 2014[8]) dictate that data on these grants and their disbursement must be published publicly on the National System of Publicity for Subsidies and Public Aid (*Sistema Nacional de Publicidad de Subvenciones y Ayudas Públicas*, SNPSAP) each year (Ministerio de Hacienda y Función Pública, IGAE, 2021[9]).

The IGAE's approach to risk-based planning

The IGAE prepares a plan each year on which ex-ante and ex-post controls will be tested. This plan is based on which controls address higher risks and which contribute most effectively to the advancement of the body's four overarching goals. These goals include: 1) combat fraud; 2) increase awareness of control activities among grantees and granting bodies; 3) seek additional value in the control beyond simple verification or repetition; and 4) account for the principles of decentralisation by using all resources, media and tools available for control activities. Previously selected testing which was incomplete from the previous year will also carry-over into the yearly plan. The IGAE's annual plans are subject to change throughout the year if new unforeseen risks emerge (IGAE, 2020[5]). For instance, in 2021, the IGAE selected two controls to evaluate: one of which is that no disqualified entities have been awarded a subsidy, and the other being that no grants have been awarded that exceed the European Commission's regulatory thresholds (IGAE, 2020[5]).

The IGAE typically plans its control activities based on the following analysis of the National Subsidies Database (*Base de Datos Nacional de Subvenciones*, BDNS), which is a database that has information on all national grants and their recipients and is under the management of the IGAE. The IGAE also relies on *CincoNet* and *Presya*, which are Spain's accounting system and loan accounting systems, respectively, as well as complaints. Examples of sources of complaints are the State Agency of Tax Administration (*Agencia Estatal de Administración Tributaria*, AEAT), individual whistle-blowers, granting organisations, and money laundering investigators. Information on beneficial ownership is available from a variety of sources (e.g. a database of the General Council of Notaries, *el Consejo General del Notariado*), and in the future, the Ministry of Justice (*el Ministerio de Justicia*) is developing a Registry of Beneficial Ownership in the future that consolidates different sources. Experience from previous years and contextual knowledge also help the IGAE to determine which areas are high risk and have control weaknesses.

The IGAE's goals, priorities and resource limitations are also considered when planning activities for the year ahead (IGAE, 2020[5]). To promote the efficient use of its resources, the IGAE adopted a risk-based approach that it describes in its 2021 Financial Audit and Control Plan of Subsidies and Public Aid. Highlighting international frameworks, including those of the Committee of the Sponsoring Organisation of the Treadway Commission (COSO) and the European Commission's Anti-Fraud Strategy, the plan outlines the IGAE's three main considerations:

1. *Grants with the highest perceived risk*—as risk indicators, the IGAE considers the amount granted, the level of fraud noted in previous years, characteristics of the grant calls and of the granting, justification and verification procedures.
2. *The visibility of the control*—the IGAE considers the visibility and impact of the control activity, recognising that high-visibility activities can act as a deterrent (i.e. beneficiaries and other stakeholders are more aware of the IGAE's surveillance) and they can lead to better management.
3. *The "profitability of the available means"*—this broadly refers to the IGAE's consideration of the efficiency of its control activities and the decentralised structure referred to as "Peripheral Services" (*los Servicios Periféricos*), which includes collaboration with line ministries and departments in regional territories (IGAE, 2020[5]).

In planning and executing its work, the IGAE must follow certain parameters that shape its control activities. Its mandate is limited to the control of subsidies and public aid, including loans, contemplated in Title III of the General Subsidies Law (*Ley General de Subvenciones*, LGS). In addition, the IGAE's control activities for 2021 generally focus on 2018 or after, recognising that some grants have multi-year execution periods. The IGAE's control activities focus primarily on grants and aid financed with national funds, although it is possible that a subsidised action may have also received funding from the European Union (EU) (IGAE, 2020[5]).

IGAE officials highlighted three key areas of fraud risk that are of particular concern in Spain's public grant-making programmes: 1) over-billing of hours by grantees; 2) double-financing; and 3) excess billing by contractors or third-parties.

- There is a risk that grantees bill additional hours over the actual service provided. Organisations receiving grant funding must report back to the granting agency on how many hours of work were completed by staff on the relevant project. This figure has implications for how much funding the grantee receives. However, as many organisations' operations are only in part funded by grants, there exists a risk that these employee hours incurred as a direct result of grant-related work could be overstated. For example, an organisation could falsely claim that wage costs which would still have been incurred in absentia of any grant, were instead, a direct result of the public funding (see Box 1.1 below for the experience of the U.S. Centre for Medicare and Medicaid Services). The IGAE attempts to control for this risk by mandating work reports on the hours utilised and applying maximum thresholds; however, these approaches can only partially mitigate the risks. IGAE officials highlighted the need for improved data to help detect this type of fraud, including data for wage hours, total company revenue, and typical personnel expenses prior to receiving the grant. This information could be added to the BDNS to support further analysis, officials said. This could include comparison of grantees to their peers to find those inefficiently using labour hours, or before-and-after tests to assess discrepancies between what a firm claims in grant documents and the wages it actually bills.
- A second area of concern is that grantees could receive funding from two or more sources, both public and private, at a level that exceeds incurred costs and results in undue profit. The BDNS helps the IGAE to combat this practice, as it includes a list of all national grants given to a single organisation. However, the BDNS does not include grants from the EU. IGAE officials noted that having this additional information on all grants given to an organisation and the total income of each from all sources would be particularly useful in identifying areas of high risk. While this declaration is a requirement for large grantees already, it is not for smaller ones. An expansion of this mandatory disclosure to all grant recipients could be achieved through a self-declaration by the grantee, a web search, or an analysis of the organisation's financial statements.
- Excess billing or outsourcing occurs when a supplier to the grantee overcharges for a particular service or supply, either by charging greater than market value or providing the less than stated amount. This fraud risk is particularly hard to detect, since it is often accompanied by a legitimate paper trail. IGAE officials highlighted the need for leveraging new technologies and data as a means of identifying cases in which this may occur, and to better analyse the environment in which firms and their suppliers operate, the geographical context and the relationships between them. Indeed, analysing relationships can be useful for broader risk analysis, going beyond the analysis of excess billing alone. This can includes relationships between suppliers, beneficiaries, beneficiaries' subsidiaries or related companies, and granting organisations. These types or relationships can lead to misappropriation, or organisations being granted an excess of funding. IGAE officials noted that the creation of a database which tracks these kinds of relationships would be useful in identifying areas of high fraud risk. See Chapter 2 for an example and further discussion on networking analyses techniques.

Box 1.1. Targeting of overbilling by the U.S. Centre for Medicare and Medicaid Services

When governments fund third parties, one frequent area of risk is the grantee overbilling work hours, whether in error or with malicious intent. In the United States, the Centre for Medicare and Medicaid Services (CMS) uses a predictive analytics system to try and capture overstatements of this nature. The Fraud Prevention System (FPS), uses a variety of data to evaluate a number of metrics, including:

- Rules-based differentiators, such as identifying credit cards or accounts that have been associated with fraudulent behaviour in the past.
- Anomaly identification, such as flagging beneficiaries that, when compared, bill larger amounts than similar entities.
- Predictive analytics, which identifies beneficiaries that have similar characteristics to known bad actors.
- Network analysis through which phone numbers and addresses of beneficiaries are compared to those of known bad actors.

By examining these traits and behaviours, the CMS has identified a number of entities with high-risk billing practices. After being flagged by the analytics programme and upon further, more traditional forms, of investigation, a number of entities have been blocked from further billing or from practicing all together. The FPS has allowed the CMS to allocate its resources efficiently and effectively. Ultimately, the programme is estimated to have saved taxpayers over USD 200 million, which is a USD 5 return for every USD 1 of investment in the system.

Source: (Centers for Medicare & Medicaid Services, 2014[10])

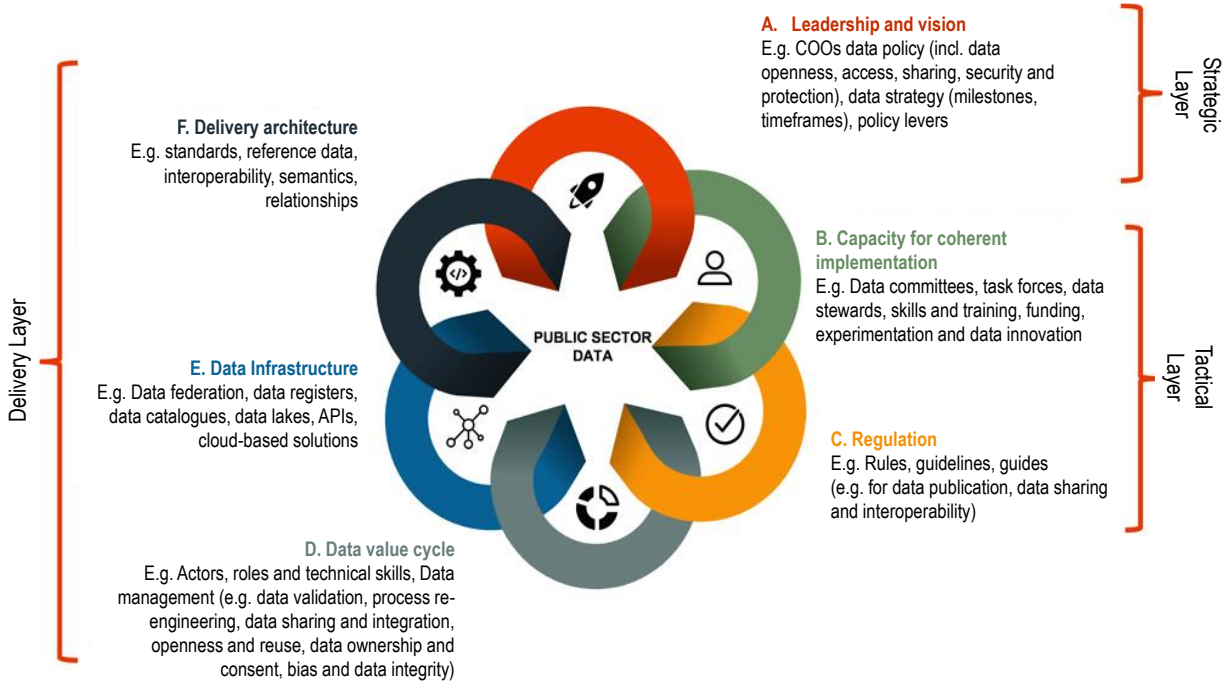
Common considerations for using data and analytics to assess risks

The IGAE is primarily a data consumer in that it relies on data inputs from other government entities to conduct its oversight work and assess risks. As discussed, much of this data are captured in the BDNS, but the IGAE also makes use of other sources, such as accounting systems, loan databases and data on complaints. The IGAE also maintains its own records on the result of control activities and sanctioned cases. IGAE officials highlighted quality checks and controls in place that are meant to ensure the reliability of the data it uses. However, while supporting the IGAE to develop the risk methodology described in Chapter 2, the OECD identified areas where the IGAE could take additional steps to enhance its use of data and analytics regardless of the specific technique or methodology. Broadly, as elaborated in this section, this includes: 1) improvements to the IGAE's data governance and management; 2) further building its capacity for analytics using data and analytics; and 3) taking into account pitfalls concerning advanced forms of risk assessments, such as limitations of using composite risk indicators and biases.

Strengthen data governance and management

Data governance, and more specifically data management, is the cornerstone for effective analytics, including the approach described in Chapter 2. Regardless of the specific methodology, any "data-driven" approach relies on these elements. The model described in Figure 1.1 highlights the values of all organisational, policy and technical aspects for successful data governance.

Figure 1.1. Data governance in the public sector



Source: (OECD, 2019[11])

The data governance model above is relevant from both a whole-of-government and institutional perspective. For audit institutions, data governance and data management are at the forefront of their everyday work. International standards and guidance, particularly those advanced by supreme audit institutions (SAIs), highlight the need for effective data governance to help audit bodies to keep pace with the digitalisation of government and society.[2] Government entities beyond SAIs are also tackling the same issues and developing their own data governance framework. For instance, in New Zealand, the lead agency for government-held data (Stats NZ) developed a data governance framework for government that promotes better data management and encourages government to adopt a "whole-of-data life cycle approach." The framework encourages public officials to think more strategically about the governance, management, quality and accountability of the data they use over the entire data life cycle (i.e. from the design and source of the data to its storing, publication and disposal) (OECD, 2019[11]). In terms of data quality, several guiding principles include:

- *Relevancy*: the extent to which the data meets the needs of the organisation and its stakeholders.
- *Accuracy and reliability*: the degree to which the data correctly and consistently describes the phenomenon being examined.
- *Timelines and punctuality*: the speed at which data can be obtained, and the reliability of this measurement.
- *Accessibility and clarity*: the ease to access, the clarity and the affordability of the data available.
- *Coherence and comparability*: the consistency of the data and the ease with which it can be combined and compared with other data.
- *Availability of metadata*: the ease with which the underlying information about the data, its structure, and attributes can be found or understood (INTOSAI, 2019[12]).

Using data from multiple sources that are prepared independently from one another can lead to an array of challenges for control bodies when applied to fraud risk detection. The IGAE administers the BDNS and

uses it for its own risk analysis, but it is not solely responsible for inputting the data into the BDNS. Public bodies, the Local Administration (*la Administración Local*), administration of autonomous communities, public sector foundations, among others, are all required to provide information to the BDNS. The IGAE does not conduct data reliability assessments on all data. As a data consumer, some of the data quality issues that were apparent in the data the IGAE uses, such as errors or missing values, are the responsibility of the agency that inputs the data. Nonetheless, audit and control bodies have an obligation to test the reliability and validity of data according to international standards, such as those of the International Auditing and Assurance Standards Board (IAASB) or the Committee of the Sponsoring Organisation of the Treadway Commission (COSO). Moreover, Spain's own standards for collecting audit evidence, such as International Auditing Norm 500 (*Norma Internacional de Auditoría 500*),[3] emphasises the need for auditors to assess reliability, accuracy and completeness of data. Therefore, even though the IGAE may be dependent to some extent on the data governance, management and quality checks of data producers (i.e. government entities or other institutions), it also must take steps to independently assess the data it obtains.

As illustrated in Chapter 2, interpreting and cleansing the data for enhancing the IGAE's fraud risk model was time consuming and resource intensive. During the course of this process, "quick win" improvements to the IGAE's data management, such as having a data dictionary that clearly describes data fields or ensuring that unique identifiers are uniformly applied across datasets, became evident. In general, data of poor quality can reflect issues like missing observations, incorrect information or misnamed variables. Any of these concerns could hinder audit or control bodies from conducting meaningful and accurate analyses of risks and controls. For instance, in the IGAE context, missing values in the data, while common, was a major issue identified while working with various databases to develop the risk model. Missing information or data points can reflect errors or simple oversight by the entity that inputted the data, but they can also be due to purposeful omission. Implementing checks and controls to prevent this from occurring could also serve an additional means of detecting and preventing fraud. From a methodological perspective, reliance on data of poor quality could lead to ineffective sampling, for example, meaning that a number of instances of grant fraud could go unnoticed every year. Inaccurate or incomplete data could also negatively bias more advanced techniques, such as the machine learning approach elaborated in Chapter 2, resulting in models with low predictive power and ultimately the inefficient allocation of the IGAE's resources.

The IGAE could take additional steps to ensure that the data in the systems and sources it uses are reliable. Put in context, confirming that data are reliable means the IGAE would deem it sufficient and appropriate specifically for fraud risk analysis and the methodology it selects. In other words, are the data complete, accurate and truly describe the key concepts under scrutiny? As a data consumer, the IGAE could work with the institutions and organisations that source the data it relies on to address some of the issues described above and in Chapter 2, and ensure the existence of sound internal controls over the data. This includes the policies and procedures that govern data collection, management, storage and use.

Generally, such controls can be categorised in three ways: 1) general controls, 2) application controls, and 3) user controls (United States Government Accountability Office, 2019[13]) General controls apply to the institution's information systems as a whole, while application controls are those built into the application to make certain that all actions within it are valid, accurate and complete. User controls are those administered by individuals to improve the reliability of the information system. By understanding the controls that are already in place, the IGAE can have better assurance regarding the reliability of the data specifically for assessing fraud risks. Moreover, drawing from the OECD's experience working with the data the IGAE uses for fraud detection, the IGAE could pay special attention to the following issues when adjudicating the reliability of the data it uses:

- Verify the total number of records provided against summary statistics.
- Check for missing observations, accounting for all necessary columns or rows.
- Confirm that none of the records are duplicated.

- Search for dates outside of the desired range.
- Search for values that are extreme outliers.

The IGAE can also look at documentation or manuals explaining how the information systems are designed, but in this case it would also need to verify that the way the system is functioning in actuality does indeed adhere to this benchmark. As another check, data could also be traced back to its source material to ensure the two are consistent (United States Government Accountability Office, 2019[13]).

Build capacity for data-driven risk assessments and analytics, particularly competencies for working with large-scale datasets and data visualisation

Data architecture, data infrastructure and capacity for implementation were highlighted by IGAE officials as some of their top priorities for enhancing data use and analytics in general. These areas were the focus of several OECD recommendations for the IGAE and the ONA to strengthen its continuous supervision system, in part, by automating processes for importing data, as well as enhancing efforts to validate and corroborate self-reported data (OECD, 2021[14]). In the context of assessing fraud risks, given the storage and scale of most grants and related datasets the IGAE uses or could access in the future, such as company registry data, the ability of government servers to manage the volume of data in a timely, reliable manner is critical for data extraction. For large datasets of several million records, even basic data cleaning and analytical work can require the use of high-capacity servers. IGAE officials highlighted the need to enhance the IGAE's data infrastructure. However, for purposes of this project and assessing fraud risks in public grant data, the existing infrastructure is sufficient for more advanced forms of risk analysis, as evidenced by the machine learning methodology described in Chapter 2.

As a more immediate need to implement the said methodology and similar analytics, the IGAE could build its internal digital competencies to manipulate large-scale datasets (i.e. hundreds of thousands or millions of observations) and to implement advanced statistical methods, such as Random Forests, as described in Chapter 2. The pre-processing phase—data creation, extraction, merging and organisation of dataset that comes before the actual analysis—is time-consuming, costly and requires data literacy to process and clean the data. Costs often depend on the quality and openness of government data systems. With some exceptions, the IGAE has the authority to access many databases that can be used for fraud detection, but taking the time to process poor quality data can drive up costs.

In addition to data quality, cost drivers can include the existence of a digitised, centralised and structured grant datasets, as well as the format of storing them and the corresponding ease of extracting the relevant fields. For this project, the OECD supported the IGAE to create a database that can be used for fraud risk analysis, regardless of the methodology used, thereby reducing such costs in the future. However, data, like risks themselves, are not static and they require the right mix of technical skills and risk expertise to be routinely updated. For instance, to further improve its capacity for carrying out data-driven fraud risk assessments, the IGAE could continue building a multi-disciplinary team with expertise in grant operations, fraud risk management, analytics and data visualisation.

The methodology in Chapter 2 made use of open source software (i.e. Python and R). While many audit institutions rely on paid software (e.g. IDEA, ACL, SAS or Stata), there is no one-size-fits all solution and many entities in search of a more robust tool than Excel have developed effective analytics based on open source tools. In general, the objectives of the analysis, as well as the skills and expertise of auditors, will determine which tool is most appropriate. For instance, the Austrian Court of Auditors (ACA) developed a tool to monitor the financial health of Austrian municipalities. The tool operates mainly through the statistics software R and enables criteria-based comparison of municipalities and identification of those that pose the highest financial risk. The ACA found that R software was better equipped for analysing big data than Excel, was less prone to error and the R codes could be readily re-used in future evaluations, with minor adaptations. The learning curve for ACA analysts was significant, according to ACA officials, given the level of detailed technical expertise required. Nonetheless, having in-house expertise in these applications

and coding languages has become a standard skillset for many audit institutions that have advanced their analytics capacities in recent years.

The capacity to leverage data and analytics goes hand-in-hand with data visualisation skills. Visualising data in a way that helps users to understand and act on results requires knowledge of data visualisation principles as well as familiarity with, if not expertise in, specialised software that can produce dashboards and facilitates auditors' understanding of risks (e.g. R Shiny package, or Tableau). IGAE officials highlighted the need for such tools and dashboards to support analyses of the BDNS as one of their top priorities and needs. Currently, the IGAE makes little use of data visualisations to assess grand fraud risks. Network analyses to identify conflicts of interest is one area that lends itself well to visualising risks (see Chapter 2).

Users that have in-depth knowledge about grant processes, available databases and risks are critical for building an effective analytics capacity and data-driven approach to risk assessments. The IGAE has a team with a strong foundation in all these areas, but could invest further in expertise in analytics and data visualisation in order to advance its digital capacities further. Creating, validity testing, and analysing fraud risk models requires both an in-depth understanding of the grant giving and implementing process as well as advanced analytic skills. Specific knowledge about grants and subsidies helps to understand the scope of data and variable definitions, as well as the regulatory framework that governs the grant cycle. These various capacity issues, many of which reflect the needs of the IGAE, highlight the importance of having clear objectives and priorities when developing an analytics capacity.

While new data-driven approaches can be a catalyst for broader change, making effective use of data and analytics requires more than simply introducing new tools, techniques or data sources. Moreover, questions about building an analytics capacity would likely need to account for other aspects of the IGAE's work beyond the scope of this project. For instance, how the IGAE builds its capacity to enhance its analytics for assessing fraud risks could likely tie into its broader digitalisation strategy, goals and resources for enhancing data architecture and infrastructure, or institutional objectives for more targeted, effective control activities. Box 1.2 describes the experience of the European Union's Internal Audit Service strategy for enhancing its analytics function by taking an institution-wide approach.

Box 1.2. Developing a strategy for analytics at the European Union's Internal Audit Service

The European Union's Internal Audit Services (IAS) has made strides to advance its use of analytics and technology in its investigations and audits over the past few years. This was achieved by, early on, devising and adhering to a strong and cohesive analytics strategy. To begin, the existing Information Technology team carried out an extensive analysis of areas for improvement, including innovations and new technologies that the service could incorporate in its work. IAS also established an internal group to continue this work, including discovering ways in which data and technology could be used on novel engagements, to stay abreast of current best practices, and to make the department more efficient through analytics. To drive this effort, IAS created a long-term strategy around analytics focused on three key areas: 1) developing a robust inventory of knowledge and skills; 2) starting pilot projects; and 3) knowledge sharing. Creating a singular organisation-wide strategy helped the IAS to more effectively plan audits, among other benefits.

Source: (Barrigon, 2020[15])

Beware of pitfalls concerning composite risk indicators as well as biases

While risk-based control is part of the IGAE's annual plan, selecting audits and investigations based on perceived risks is ultimately aimed at maximising the value-for-money of taxpayer money. It is therefore critical for the IGAE to be mindful of some of the pitfalls inherent in typical approaches to risk assessments, and to reduce the risk of both false positives and false negatives. One of the most frequent ways of creating (composite) risk indicators is based on manually selecting observed features of well-known salient cases and generalising them by applying the same indicators to the full dataset of cases.

This approach suffers from two major drawbacks. First, it causes the so-called selection bias, meaning that particular cases were taken into considerations, with assumption that their characteristics are generalisable to other observations, without any proof that these are typical or representative of all types of fraudulent schemes. Second, such approaches typically fail to take into account the prevalence of selected risk indicators (or red flags) among clean and unknown cases. In other words, they often produce high false positive rates, meaning they often signal fraud risks when there is no fraud. Third, typically such approaches apply a simple averaging of individual red flags to produce a composite score as they lack the understanding of how different indicators coincide with each other or which ones are more important.

While not the only approach, the methodology described in Chapter 2 was selected because it addresses these shortcoming, and as discussed below, it allows the IGAE to work around some of the peculiarities of the data it uses. The machine learning method in Chapter 2 generalises from all past proven cases (i.e. sanctioned cases) to identify which factors influenced the probability of being sanctioned. This approach leads to a single risk score composed of all relevant features in the data, with weights of each feature defined to maximise predictive power. The approach also explicitly addresses the problem of false positives and false negatives, learning from both proven positive (sanctioned) and likely negative cases (non-sanctioned). Nonetheless, no methodology is completely free from the risk of bias or inaccuracies; however, being mindful of these and the inherent tendencies of specific methodologies concerning these issues can help the IGAE in taking an informed approach to strengthening its current risk assessment methodology. Box 1.3 explores further how the IGAE can control for biases in its models, drawing from international leading practices.

Box 1.3. Addressing biases in machine learning models

Machine learning models are trained based on the data that is available, so they themselves can be inherently biased. The author of the algorithm can also amplify these biases further, purposefully or subconsciously. This is of particular concern for auditors or fraud practitioners, for whom objectivity is of the utmost importance. A number of institutions, including audit bodies and think tanks (e.g. the Brookings Institution), have issued guidance about how to audit artificial intelligence and how to check for biases in algorithms to enhance machine learning model. These include, but are not limited to, the following:

- Algorithms can be periodically and independently audited. The audit could include evaluating the data collection process, monitoring how the programme works, and checking whether it is fairly evaluating sensitive subgroups.
- The programme could be compared to risk-assessments prepared by humans to see if it is actually more effective.
- Algorithms can be checked for compliance to non-discrimination laws.
- Algorithm operators can make attempts to increase human interaction with the program, striving to ensure the code and metrics being used are understood, and that their relation to key social inequities are being considered.

- The operating agency could consider drafting a formal bias impact statement to document its conscious consideration and strategy when managing this challenge.

According to the Brookings Institution, some questions that can be pondered and included in such a statement in order to assess and control for biases include:

- What will the automated decision do?
- Who is the audience for the algorithm and who will be most affected by it?
- Does the organisation have training data to make the correct predictions about the decision?
- Is the training data sufficiently diverse and reliable? What is the data lifecycle of the algorithm?
- Which groups may be treated unfairly or may be impacted disproportionately by the training processes of the model and ensuing analysis?
- How will potential biases be detected?
- How and when will the algorithm be tested? Who will be the targets for testing?
- What will be the threshold for measuring and correcting for bias in the algorithm?
- What are the operator incentives?
- What will is to be gained from the development of the algorithm?
- What are the potential bad outcomes and how will the organisation become aware of these?
- How open (e.g., in code or intent) will the design process of the algorithm be to internal and external stakeholders
- What intervention will be taken if it is predicted that there might be bad outcomes associated with the development or deployment of the algorithm?
- How are other stakeholders being engaged?
- What's the feedback loop for the algorithm for developers, users and stakeholders?
- Is there a role for civil society organisations in the design of the algorithm?
- Has diversity been considered in the design and execution?
- Will the algorithm have implications for cultural groups and play out differently in cultural contexts?
- Is the design team representative enough to capture these nuances and predict the application of the algorithm within different cultural contexts? If not, what steps are being taken to make these scenarios more salient and understandable to designers?
- Given the algorithm's purpose, is the training data sufficiently diverse?
- Are there statutory guardrails that organisations should be reviewing to ensure that the algorithm is both legal and ethical?

Source: (Canadian Audit and Accountability Foundation, 2019[16]); (Lee, Resnick and Barton, 2019[17])

Conclusion

The IGAE has developed a solid foundation to advance its use of data and analytics for assessing fraud risks in public grant data. The skills and knowledge it has in-house, particularly with respect to the grant-making processes, existing risks and the intricacies of relevant databases, are key elements of the capacity and expertise needed for effectively assessing grant fraud risks. There is no analytical tool or method that can replace this knowledge or expert judgement. Moreover, by some accounts, on-the-spot checks and internal fraud reporting mechanisms are perceived to be the most effective fraud detection measure, ranking higher than data analytics or data mining (Dozhdeva and Mendez, 2020[18]). Nonetheless, with an increasingly digital government and society, oversight bodies like the IGAE will have to evolve out of necessity as opposed to by choice.

Building on its strong foundation of expertise and knowledge, the IGAE could consider adding capacities for taking advantage of the full potential of existing databases at its disposal, in particular, strengthening its capacity for working with multiple large datasets and visualisation of data. At the same time, the IGAE could continue to improve its data management, and it checks on the quality of data to facilitate the merging of datasets and assessing fraud risk in public grants. These are actions that would help the IGAE to mature from an analytics perspective regardless of whether it decides to adopt the specific methodology in Chapter 2. Advancing its use of data and analytics would help the IGAE not only to collect more predictive insights about the risks in public grant programmes, but also to be more efficient and effective in its use of taxpayer money.

References

Barrigon, F. (2020), "Innovation and digital auditing – the journey of the European Commission's IAS towards state-of-the-art technologies", *ECA Journal*, Vol. 1/2020, pp. 97-101, https://www.eca.europa.eu/Lists/ECADocuments/JOURNAL20_01/JOURNAL20_01.pdf. [15]

Canadian Audit and Accountability Foundation (2019), *Artificial Intelligence and Auditing: Overview of Potential Impact on Public Sector Auditors*, https://caaf-fcar.ca/en/performance-audit/research-and-methodology/research-highlights/3455-research-highlights-3. [16]

Centers for Medicare & Medicaid Services (2014), *Report to Congress, Fraud Prevention System, Second Implementation Year*, https://www.cms.gov/About-CMS/Components/CPI/Widgets/Fraud_Prevention_System_2ndYear.pdf (accessed on 13 August 2021). [10]

Dozhdeva, V. and C. Mendez (2020), *Is fraud risk management in cohesion policy effective and proporitionate?*, https://www.eprc-strath.eu/public/dam/jcr:dbcbcfde-e024-44a0-a11b-b12456ffe0c5/EPRP%20121%20-%20IQ_Net_Thematic%20paper%2047(2).pdf. [18]

European Commission Anti-Fraud Office (OLAF) (2017), *Handbook on Reporting on Irregularities in Shared Management*, https://www.eu-skladi.si/sl/dokumenti/navodila/handbook-irregularity-reporting-final.pdf (accessed on 13 August 2021). [19]

European Union (2014), *Commission Regulation (EU) No 651/2014*, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02014R0651-20210405 (accessed on 13 August 2021). [7]

European Union (2014), *Commission Regulation (EU) No 702/2014*, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02014R0702-20201210 (accessed on 13 August 2021). [8]

Government of Spain (2019), *Royal Decree 130/2019 (Real Decreto 130/2019)*, https://www.boe.es/eli/es/rd/2019/03/08/130 (accessed on 13 August 2021). [6]

Government of Spain (2003), *Law 38/2003, General Subsidies (Ley 38/2003, de 17 de noviembre, General de Subvenciones)*, https://www.boe.es/buscar/pdf/2003/BOE-A-2003-20977-consolidado.pdf (accessed on 13 August 2021). [4]

Government of Spain (2003), *Law 47/2003, of November 26, General Budgetary*, https://www.boe.es/buscar/act.php?id=BOE-A-2003-21614&p=20201231&tn=6. [3]

IGAE (2020), *Activity report 2019 (Memoria de actividades 2019)*, https://www.igae.pap.hacienda.gob.es/sitios/igae/es-ES/QuienesSomos/Documents/Memoria_2019.pdf. [2]

IGAE (2020), *Approval Of The Audit And Financial Control Plan Of Subsidies 2021 (Aprueban El Plan De Auditorías Y Control Financiero De Subvenciones 2021)*, https://www.igae.pap.hacienda.gob.es/sitios/igae/es-ES/Control/CFPyAP/Documents/Resoluci%C3%B3n%20Plan%20Auditor%C3%ADa%20Pbca%20y%20CFP%202021.pdf (accessed on 13 August 2021). [5]

INTOSAI (2019), *Training Tool on Environmental Data: Resources and Options for Supreme Audit Institutions*, https://www.environmental-auditing.org/media/113693/23g-wgea_environmental-data_2019-fin.pdf (accessed on 13 August 2021). [12]

Lee, N., P. Resnick and G. Barton (2019), "Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms", *Brookings Institute*, https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/ (accessed on 13 August 2021). [17]

Ministerio de Hacienda y Función Pública, IGAE (2021), *National System of Publicity for Subsidies and Public Aid (Sistema Nacional de Publicidad de Subvenciones y Ayudas Públicas)*. [9]

OECD (2021), *Enhancing Public Accountability in Spain Through Continuous Supervision*, OECD Public Governance Reviews, OECD Publishing, Paris, https://doi.org/10.1787/825740cc-en. [14]

OECD (2019), *The Path to Becoming a Data-Driven Public Sector*, OECD Digital Government Studies, OECD Publishing, Paris, https://dx.doi.org/10.1787/059814a7-en. [11]

OECD (2014), *Spain: From Administrative Reform to Continuous Improvement*, OECD Public Governance Reviews, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264210592-en. [1]

United States Government Accountability Office (2019), *Assessing Data Reliability*, https://www.gao.gov/assets/gao-20-283g.pdf (accessed on 13 August 2021). [13]

Notes

[1] The European Union defines irregularities as “any infringement of a provision of Community law resulting from an act or omission by an economic operator, which has, or would have, the effect of prejudicing the general budget of the Communities or budgets managed by them, either by reducing or losing revenue accruing from own resources collected directly on behalf of the Communities, or by an unjustified item of expenditure.” Alternatively, fraud is considered to be “in respect to expenditure, any intentional act or omission relating to the use or presentation of false, incorrect or incomplete statements or documents, which has as its effect the misappropriation or wrongful retention of funds from the general budget of the EU or budgets managed by, or on behalf of, the EU, or non-disclosure of information in violation of a specific obligation, with the same effect, or the misapplication of such funds for purposes other than those for which they were originally granted.” (European Commission Anti-Fraud Office (OLAF), 2017[19]).

[2] For instance, see the African Organisation of Supreme Audit Institutions research report on integrating big data into public sector auditing (https://afrosai-e.org.za/wp-content/uploads/2020/12/Research-Paper-Integrating-Big-Data-in-Public-Sector-Auditing.pdf); the training tool on environmental data published by the INTOSAI Working Group on Environmental Auditing (https://www.environmental-auditing.org/media/113693/23g-wgea_environmental-data_2019-fin.pdf); or the experiences of the Netherlands Court of Audit in developing an audit framework for algorithms (http://intosaijournal.org/developing-an-audit-framework-for-algorithms/).

[3] International Auditing Norm 500 was adapted from the International Standards on Auditing issued by the International Federation of Accounts through the IAASB.

2 Fraud in public grants: Piloting a data-driven risk model in Spain

This chapter presents a proof-of-concept for a risk model that the General Comptroller of the State Administration (Intervención General de la Administración del Estado, IGAE) of Spain can employ to assess fraud risks and detect likely fraud cases. The chapter presents an overview of the machine learning methodology that underlies the risk model, as well as a detailed account of how the model was built, based on data that are readily available to the IGAE. The chapter concludes with a discussion about the results of the model and recommendations for the IGAE to build on the proof-of-concept.

Introduction

Data-driven fraud risk assessment frameworks can have multiple uses, among which identifying investigative priorities is central. When investigative resources are scarce and a random selection of cases for investigation is likely to yield a low success rate (e.g. because fraud is rare on the target population), a risk-based case selection can deliver material benefits. To this end, a risk score assigned to all potentially investigated cases can feed into the prioritisation of cases for investigation. This typically does not imply a full automation of case selection; rather analytics offers a crucial input into the organisational decision-making process.

For a large-scale risk assessment to deliver benefits, it has to be accurate enough to be used for risk scoring transactions or organisations on an ongoing basis, including new cases. In general, risk scores can be defined to be valid for such purposes either by explicitly defining the risk factors from known relationships and risk descriptions (e.g. grant recipient organisation's ultimate owner being based in a tax haven) or by defining the combination of risk features through statistical means, including machine learning from past investigations. Either way, what is crucial is that the risk model not only considers known fraudulent cases and their features but it also takes into account the features of a much larger group of cases which were not investigated, hence whose fraud status is unknown. In short, indicator validation and continuous improvement are of crucial importance, as shown in this chapter.

When developing new analytical approaches, often insights and learning comes from doing. This is why many audit institutions, for instance, have established "Innovation Labs" and internal communities of practice to test and experiment with new audit techniques, analytics and technologies and uses of data. This incremental approach allows audit and control bodies to take measured risks and contain costs, before either scaling up or winding down pilot initiatives. In this spirit and in response to the interest of the General Comptroller of the State Administration (*Intervención General de la Administración del Estado*, IGAE) in strengthening its use of data to detect grant fraud risks, this chapter presents a proof-of-concept for a data-driven risk model for the IGAE to adopt in part or in its entirety.

The methodology in this chapter aims to make use of the data that was already at the IGAE's disposal, including the National Subsidies Database (*Base de Datos Nacional de Subvenciones*, BDNS), and in doing so, implicitly takes into account the IGAE's context. As noted in Chapter 1, like any investment to improve data governance, data management, or analytics, this approach may require investments in skills and digital capacity. For this reason, the chapter provides a detailed account of all stages of the methodology and its development to support the IGAE's own assessment of what it is able to do with its existing resources and skills in-house. In addition, the process of developing the proof-of-concept for the risk model led to several insights and the identification of areas for improvement, which are addressed in the results section.

Overview of the machine learning model

A brief primer on machine learning for risk assessments

The IGAE's current approach to assessing fraud risks, described in Chapter 1, is outlined in its 2021 Financial Audit and Control Plan of Subsidies and Public Aid. The IGAE considers the grant amount, previous levels of fraud and other qualitative factors, such as the justification and verification procedures. The machine learning model described in this chapter advances a more data-driven approach, which can complement the IGAE's existing processes. In reality, given resource constraints, the IGAE can only carry out a finite number of control activities in any year. The machine learning methodology described in this chapter should not replace auditors' judgement. For instance, the model can highlight cases of likely fraud, but the auditor will also need to assess which of these cases would be the most profitable in terms of

further investigatory or control activities. Taking this nuance into account, the model can be a useful input into auditors' decision-making, and help the IGAE to target its resources more effectively.

The risk model developed to support the IGAE is based on a random forest methodology. Random forests is a supervised machine learning method which predicts the output by constructing multiple decision trees with given features (Breiman, 2001[1]). It is particularly well suited for datasets with a large number of explanatory variables or potential risk indicators. By using random forests, it is possible to include a wide list of explanatory factors of different types (numeric and categorical).

Selection of the methodology

In order to analyse the data using machine learning methods such as random forests, the dataset was cleaned by removing missing values and variables lacking variability (i.e. where the variables take almost always the same value in the entire dataset). Random forests allow for working with a large number of observations as well as variables, performing algorithm training on a reduced and balanced sample, and testing models on a set-aside sample. Random Forest algorithms are sensitive to missing values. For this reason, variables with high missing rates were dropped. The method is also sensitive to imbalance in the dependent variable (i.e. sanctioned versus non-sanctioned), as described below. In general, the approach can be broken down into the following steps:

1. Identify which grantees were sanctioned, and then mark all awards of the sanctioned organisations from the last 2 to 3 years prior to sanctions. In this period, proven fraudulent activity is very likely to have taken place. This gives a full set of proven positive cases (sanctioned awards); however it leaves a very large sample of unlabelled cases (non-sanctioned). Some of these cases likely should have been sanctioned, but were not investigated, and others are true negative cases where sanctioning would not have occurred even if they were investigated. In other words, the dataset is strongly imbalanced. In most of the cases, it is unknown if the award was not sanctioned, either because it was not investigated, or because it was yet no violations were discovered. Therefore, the majority of observations are neither positives nor negatives, but rather they are unlabelled.
2. Choose the method that fits the particular problem in the data, that is, an imbalanced sample and the presence of a large, unlabelled subsample. For these purposes, a positive unlabelled (PU) bagging model is applied. This machine learning method enables training a model on random samples of observations, both positive and unlabelled, in order to assign likely negative status (not sanctioned) and likely positive (sanctioned) status to unlabelled cases. Box 2.1 provides additional background on PU bagging and random forest models.
3. After the labels are assigned, use the relabelled dataset to train the model and identify factors that influence the probability of being sanctioned. The influence can be both positive and negative. The model then calculates the probability for each award to be sanctioned on any number of observations.
4. Once the model is trained and achieves a sufficient accuracy, apply it to the full dataset of awards in order to predict a fraud risk score for all observations.[1]

Box 2.1. Overview of positive unlabelled learning and bagging

Positive unlabelled (PU) learning is a semi-supervised machine learning technique, which allows working with highly unbalanced data (Elkan and Noto, 2008[2]). PU learning could be used in cases when the majority of all available observations belongs to unlabelled cases. For example, this includes situations when a binary variable (i.e. values of 1 and 0) has positive observations (1) that appear only in case of treatment, and when it is unknown whether the remaining negative cases (0) were treated but remained negative, or they were not treated at all. PU learning observes all the positive and negative cases, identifies the most typical characteristics referring to each, and relabels observations accordingly.

A PU bagging approach consists of several steps (Li and Hua, 2014[3]). First, it involves building a classifier by analysing the variety and combination of factors influencing positive and negative outcomes. To build a classifier, a subset of data are created, consisting of all positive cases and a random sample of unlabelled ones. This classifier is further applied to the rest of unlabelled cases to assign the probability scores for the rest of the observations. Each step is repeated several times, and then the average score received by each observation is calculated.

After relabelling, all observations are divided into training and testing samples. The ratio of the split is flexible, but it is usually between 60-70% for the training sample and 30-40% for the testing sample. Next, the random forests method is applied to the training set of data. The parameters of the model can be specified manually, including the number of trees, maximum number of features in each individual tree and the size of terminal nodes. The choice of the parameters depends on the overall size of the data, namely, the number of observations and indicators included in the model. After applying the random forests method to the training sample, the output probabilities can be predicted for the rest of the data.

Additionally, to identify the impact of each feature, SHAP values (Shapley Additive Explanations) can be calculated once the model is constructed. SHAP values show how much and in which direction (positive or negative) a given indicator changed the predicted output. To estimate the model fit, such parameters as accuracy, recall and precision should be calculated. All of them calculate the number of correctly predicted scores in either absolute or relative numbers.

Source: (Mordelet and Vert, 2014[4])

Consideration of strengths, weaknesses and assumptions

The validity of the analysis depends on two factors: the quality of the learning dataset and the availability of the relevant award, grant and grantee features. First, the main indicator differentiating fraudulent cases from non-fraudulent ones is the presence of sanctions. For positive-unlabelled learning to produce valid results, it was assumed that positive cases have been selected at random, hence are a representative sample of all positive cases. This also implies that if the observed sanctions sample missed out on some typical fraud schemes (i.e. not even one example is to be found among observed sanctions cases), the machine learning model will not capture such fraud types, hence will be biased. Similarly, if cases were selected following a particular variable, say the size of the grantee, the model will overestimate the importance of such a variable in the risk prediction. In other words, supervised machine learning uses the information given by proven cases, therefore if there is a bias in the sample of sanctioned awards - it will be replicated in the prediction process.

Second, the machine learning model can only learn features of fraud which are captured by the data. The presence of certain indicators in the dataset influences the predictive power of the model: if some crucial characteristics are missing from the data, they will not be taken into account by the model. Missing features or indicators also imply that the final list of influential indicators may be biased, overstating the importance of those features which are correlated with influential but unobserved features (e.g. if a particular region is found to be of higher risk, it may actually mean that some entities in that region have risky features, say links to corrupt politicians, rather than the region itself, its culture, administrative structures etc. being more prone to fraud.). Nevertheless, the chosen machine learning method based on Random Forests is particularly well suited for large datasets with a big number of explanatory variables or potential risk indicators (James et al., 2015[5]) It is possible to include a wide list of explanatory factors of different types (numeric and categorical).

Developing a proof-of-concept for a data-driven risk model

Identifying relevant data sources and variables for assessing grant fraud risks

The data provided by the IGAE consists of 17 datasets covering different pieces of information on awards, third parties, projects, grants and grantees. They could be grouped into three main categories.

- The first category consists of seven datasets that cover information about the grant, such as location, type of economic activity, objectives and instruments.[2]
- The second category covers awards information, including information on refunds, projects, returns and details of the awards to beneficiaries.[3]
- The third category includes datasets that cover information on the beneficiaries themselves, which can include a range of actors responsible for implementing a project (e.g. a government entity, contractor or sub-contractor), such as whether a beneficiary was sanctioned or disqualified, as well as the type of economic activity, location, and so forth.[4]

In total, these datasets consist of around 100 variables covering details of the awards (amount, date of receiving, type of economic activity, etc.), grant calls (publicity, type of economic support, regulatory base, etc.), and details of the third parties (location, legal nature, economic activities, etc.). The time period covered is 2018-2020.

All three groups of datasets present different levels of data: the first category covers grant-level information, and each grant could embrace a few awards. The second category includes award-level, and could be linked to the main dataset BDNS_CONCESIONES by unique award IDs. Finally, the last category is third party-level, and the same third party can receive multiple awards. Therefore, for the sake of merging all datasets between each other, the award level was used as a main unit of analysis, providing unique IDs.

The list of variables relevant for fraud risk assessment could be divided into background and risk indicators. Background indicators are needed to describe specific characteristics of grants, grantors, grantees and third parties which are potentially associated with sanctions. Risk indicators refer to certain phases of grant advertisement, selection, execution and monitoring. Table 2.1 shows the full list of background indicators, Table 2.2 shows risk indicators that could be extracted from IGAE's datasets.

Table 2.1. Background indicators

Indicator group	Indicator name	variables code	variables header
Grantor	Call organiser	CON 705; CON 710; CSU 100	Organising body; Granting body; Granting body
Grantee	Beneficiary	CON 580; CSU 120	Types of beneficiary; ID of beneficiary
	Concession ID	PRO 110; PAG 100; DEV 100; REI 100	Concession Identification
	Project ID	PRO 130; EJE 110	Project Identification
	project description	PRO 210	Project description
	project location	PRO 260	Geographic region (project)
	Execution ID	EJE 120	Executor Identification
	year	EJE 130	Execution year
	Disqualified ID	INH 100	Disqualified ID
	Disqualification date	INH 210	Disqualification date
	Disqualifying body	INH 220	Identify the administrative or judicial origin of the disabling body
	Disqualification period	INH 230; INH 240	Disqualification start date; Disqualification end date
Grant	grant value	CSU 220; EJE 210	Grant amount; Grant amount to the executing agency per year
	grant regulatory base	CON 250; CON 260	Description BBBR; URL BBBR
	grant identification	CON 290	Call Title
	grant publication	CON 300; CON 310	Sent for Publication; Official Source
		CON 335	Title in Spanish, Text in Spanish
		CON 335; CON 340	Date of Publication; Link to the Publication
	signature data and place	CON 351; CON 352	Data of Signature; Location of Signature
	application	CON 440; CON 460	Start date; End date
	state aid	CON 490; CON 495; CON 515	Condition of State Aid; aid authorisation; EU aid identifier
	call sectors	CON 550	Sectors of Economy
	grant location	CON 570	Geographic regions
	deadline	CON 600	Deadline for concession justification
	nominative grant	CON 610	Concession of nominative grant nature
	EU funds	CON 690	EU fund financing amount
	regulations	CSU 110	Regulation ID
	payment date	PAG 210	Date of payment
	payed amount	PAG 220	Amount of payment
	retention	PAG 230	Tax withholding
	return date	DEV 210	Date of return
	return amount	DEV 220	Amount of return
	interest rate	DEV 230	Amount of interest rate
	reimbursement date	REI 210	Date of reimbursement
	reimbursement cause	REI 220	The cause for reimbursement
	reimbursed amount	REI 230	The amount of reimbursement
Third party	country	TER 100; TER 250	Third party country; Country of domicile
	id	TER 110	Third party ID
	name	TER 240	Third party name; Third party business name
	surname	TER 210	Third party first surname; Third party second surname
	address	TER 252; TER 254; TER 256; TER 258; TER 310	Address of domicile; Postal code; Municipality; Province; Region
	type	TER 280; TER 290	Legal status; Third party type
	activity	TER 320	Sector of Economy

Source: Author

Table 2.2. Risk indicators

Phase	Indicator name	Indicator definition	Variables (code)	Variables (header)
Competition	lack of advertisement	No proper electronic advertisement of the grant scheme	CON 310; CON 420; CON 620	Official Source; Open Application; Concession publicity conditions
Selection	Procedure of selection	Inadequate standard and procedure for the selection	CON 560; CON 540; SAN 110; SAN 100; SAN 210	Help instrument; Public purpose; Penalty discriminator; Identification of the sanctioned; Date of sanction resolution
	Improper selection	Improper selection of subsidy recipients	CSU 120; CSU 130; PAG 110; DEV 110; INH 110; CON 630	Beneficiary; Grant award discriminator; Payment discriminator; Return discriminator; Disqualification discriminator; Gender impact
Execution	Unmonitored transactions	Transactions that bypass normal review procedures, or are otherwise unmonitored	CON 502; CON 503; CON 504	EU Regulation exemption by aid category; Objectives of exemption; Regulation of exemption by amount
	Inconsistent payments	Payment is unreasonably expensive or does not relate to the grant programme objectives	CSU 250; CSU 240; CSU 220; PRO 220; PRO 240: PRO 250; EJE 220; EJE 240; EJE 250	Equivalent grant award aid; Financeable cost of the activity (grant); Grant award amount; Amount of grant for project; Project costs; Equivalent aid (project); Amount of grant for executing agency per year; Const of the project assigned to executing agency per year; Equivalent aid (executor)
	Rounded payments	A recipient of a reimbursement grant that draws grant funds using numbers rounded to the nearest hundred, thousand, or greater may indicate funds are not being drawn on a reimbursement basis.	CSU 250; CSU 240; CSU 220	Equivalent grant award aid; Financeable cost of the activity (grant); Grant award amount
Monitoring	Sanctions	High number of systematic violations by recipient	SAN 250, SAN 280; SAN 440; SAN 450	Fine for minor infractions; Fine for serious infractions; Publicity of sanctioning; Deadline for advertising the sanction

Source: Author

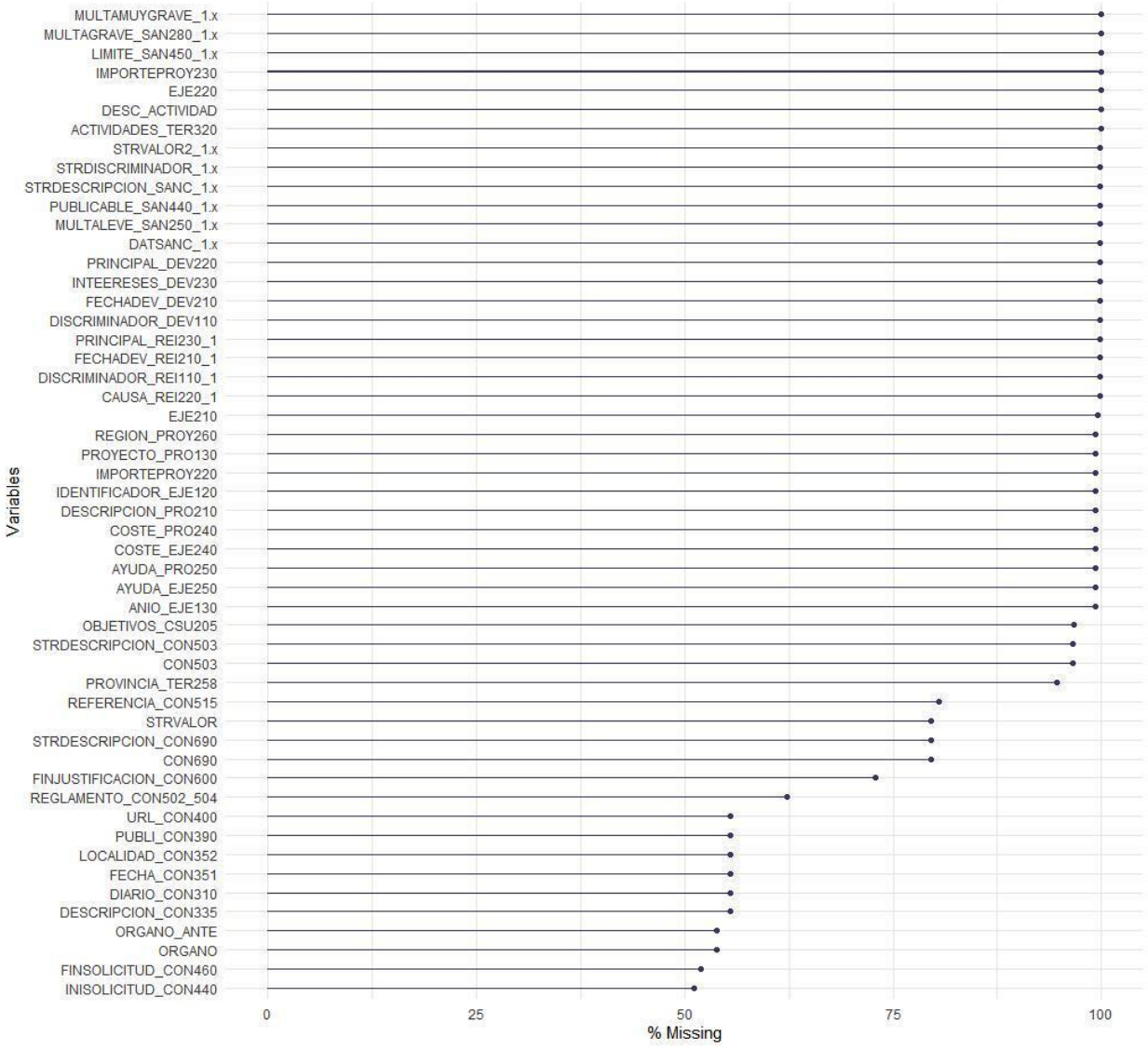
Merging, cleaning and understanding the limitations of the data

The first step of data processing was to merge all the datasets provided by Spanish government to the main dataset covering all grants and awards, BDNS_CONCESIONES. In order to do that each dataset was aligned to the same unit of analysis - award ID. When multiple observations were related to the same award ID (e.g. one award relating to multiple sectors), the data was aggregated, for example by placing each duplicate observation in a separate column. When observations related to a series of award IDs (e.g. when the dataset contained information on calls for applications), the relevant characteristics were copied over to all awards relating to that higher level observation. The merged, but uncleaned dataset contains 1 792 546 awards and 152 variables. The full list of variables included could be found in Annex B.

The next step of data processing was data cleaning. This involved removing variables with high missing rate or low variance. Such data problems impact a high number of variables as shown in Figure 2.1. All variables with a missing rate higher than 50% were removed as they would have introduced a high degree

of noise into the analysis. Most of these variables with high missing rate correspond to sanctions and project description. Additionally, some of the variables showed very low variance less than 0.3, which means that they carry very little relevant information for the subsequent analysis (i.e. in technical terms: their discriminant value is low as they do not vary sufficiently between sanctioned and non-sanctioned observations). Finally, text variables which are not directly relevant for risk scoring were removed such as text descriptors of categorical variables (e.g. sector descriptions in Spanish) and free text variables carrying little relevant information (e.g. title of the call for grant applications).

Figure 2.1. Missing values rates



Source: Author

As the analytical methods used can be sensitive to missing information, only those observations, i.e. awards, which had no missing values on all the variables considered in the analysis were retained. After conducting all these steps of data processing, the final dataset used in the analysis consists of 1 050 470 observations, awards, and 60 variables for the years from 2018 to 2020 (inclusive).

Using existing data to create new indicators

While most indicators used in the analysis directly derive from the data received, a few indicators were also calculated by combining other variables. The first group of such calculated indicators refer to the size and number of awards received by the same beneficiary. The second group consists of location-related variables: the territorial level of grantor and grantee: national, regional or local. Additionally, a binary variable was created for identifying if the execution of the contract was located in the same place as the third party. Third, an indicator was calculated capturing the month of grant award which can indicate seasonality in spending and the corresponding risks. Finally, the sector of the award was cleaned to only capture the highest level of the NACE classification. See Annex B for a table that describes all the variables from these additional variable calculations and the previously described data processing steps. This is the final list of variables used for risk modelling.

Defining the dependent variable based on sanctioning status

The main dependent variable used for the analysis is a binary variable indicating if the third party receiving the award was sanctioned or not; with the sanctioning interpreted as a reliable indication of fraud in an award. The variable turns value “1” if the third party was sanctioned for corresponding award, as well as for all previous awards received by the same party, as fraudulent practices have taken place earlier than the date of sanction. In case if the third party was not sanctioned, the corresponding award gets value “0” in the dummy variable. The classes in the sanction variable are very imbalanced - it shows 1 031 cases of sanctions against 1 049 439 cases of no sanctions.

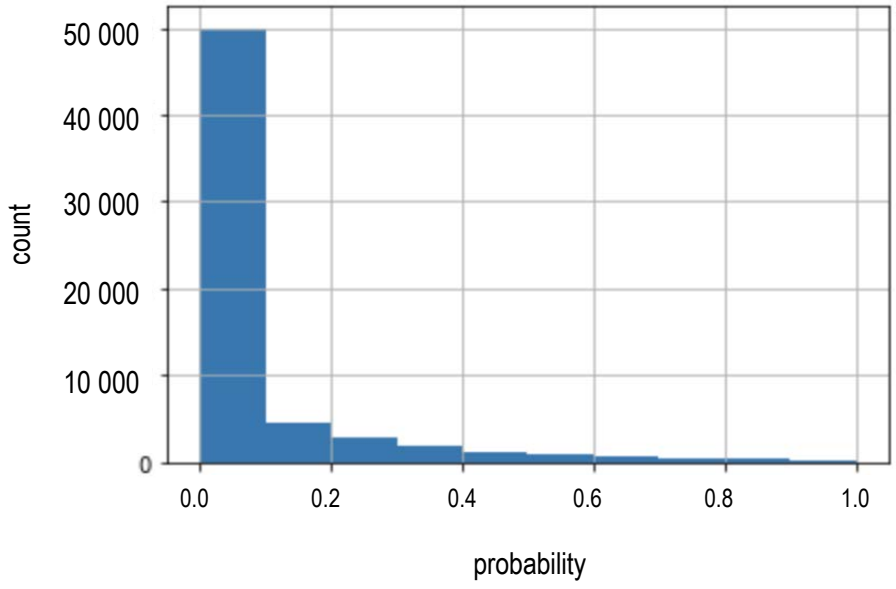
In order to make the random forest algorithm run efficiently, a random sample was drawn of 90 000 awards from the unlabelled portion of the dataset. Hence, the training dataset used in the below analysis makes use of the initial sample of 91 031 awards, consisting of 1 031 positive (known sanctions) and 90 000 unlabelled (unclear fraud status) awards.

Assigning sanctions status to unlabelled awards

In order to assign positive and negative labels to unlabelled observations, the positive unlabelled learning methodology was used. This method starts off by creating a training subset of the data, consisting of all positive cases and a random sample of unlabelled cases. On this sample, PU bagging builds a classifier which assigns the probability of sanctioning to each award, based on which it is possible to assign the positive and negative label (sanctioning probability >50% → positive label). These steps are repeated 1 000 times in order to build a reliable classifier which identifies the likely negative and likely positive cases in the unlabelled sample (please note that the average predicted sanctioning probability across all models will become the eventual predicted score).

As a result of running these algorithms, all unlabelled cases received a sanctioning probability and hence a likely sanctioning label (positive vs negative). For the training dataset, Figure 2.2 presents the distribution of sanctioning (i.e. fraud) probabilities. This highlights that most awards are considered low to very low risk with only a handful of awards receiving a high risk score. In other words, most awards can be classified as non-sanctioned while very few awards receive the sanctioned label: compared to the initial positive-unlabelled sample, the number of likely positive (sanctioned) cases increased to 4 430 with the 86 601 being identified as likely negative (not sanctioned).

Figure 2.2. PU bagging classifier: sanctioning probability prediction on the initial sample



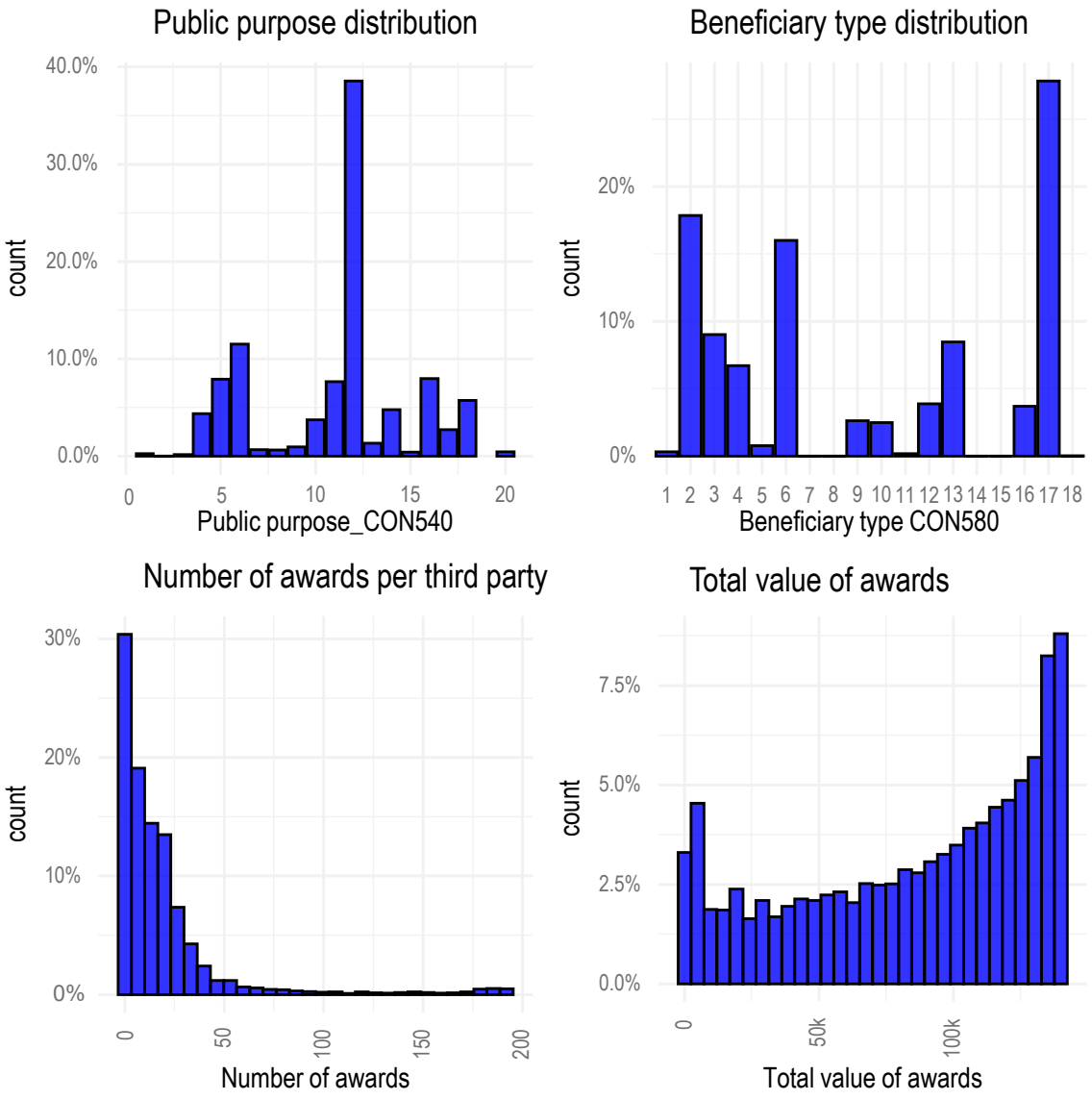
Source: Author

Identifying the most impactful variables

Once the dataset of positive and unlabelled awards is relabelled and only positive and negative cases remain in the dataset following the methods above, a new Random Forest model is run and tested for its accuracy. This means that the relabelled dataset of 91 031 awards was split into a training sample (70%) and a test sample (30%). The Random Forest algorithm will be trained on the former and test its accuracy on the latter sample which it has not 'seen'. The best model consists of 1 000 trees and uses 106 variables at each run.

This best Random Forest model identified the most important variables for predicting the probability of sanctions. For the purposes of modelling, each categorical variable was transformed into a set of binary variables, so that they correspond to a single category of the categorical variable. Numeric variables were used as is, without transformation. The most impactful variables in the best Random Forest model are Public_purpose_CON540,Nawards_TER_110, Amount_awards_TER110, and Third_party_legal_Spain_TER280. Their distributions are presented in Figure 2.3.

Figure 2.3. Distributions of the most impactful variables



Source: Author

These distributions show that many of the most important variables have rather uneven distributions. For example, the number of awards falls overwhelmingly below 50 with very few beneficiaries having more than 50 grants awarded. Similarly, the public purpose variable has a small number of prevalent categories such as 12 (agriculture). Moreover, the number of awards received by the same beneficiary is not correlated with the total value of awards, meaning that the average amount of distributed awards is relatively low with some awards being of very high value. The next section goes one step further and discusses the impacts of these impactful variables on sanctioning (fraud) probability.

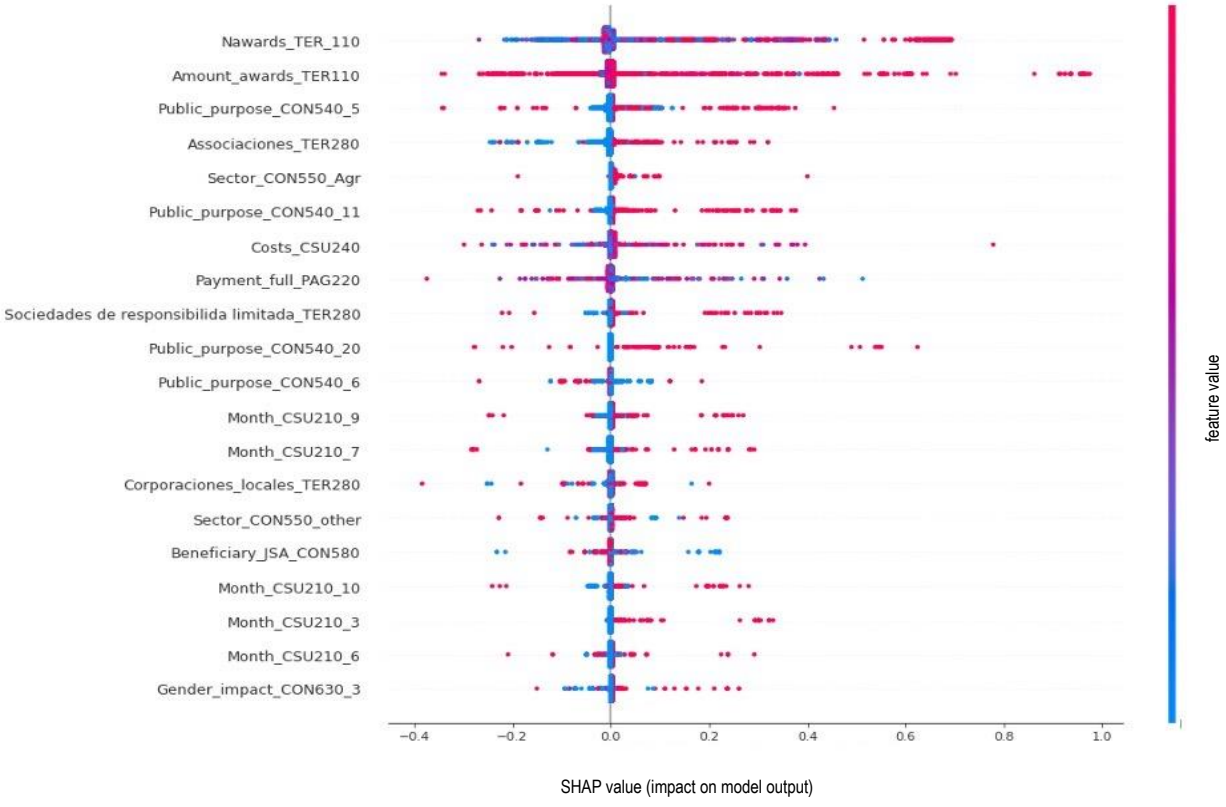
Testing the model on an unseen dataset

The best model trained on the training dataset was tested on unseen data, the test set (30% of the sample). On this test dataset, the best Random Forest model achieved:[5]

- accuracy = 95% (accuracy is the number of correctly predicted labels out of all predictions made), and
- recall = 93% (recall is the number of labels that the classifier identified correctly divided by the total number of observations with the same label).

Such results lead us to the conclusion that the model is of high quality. After establishing overall model quality, attention is turned to the impact of individual predictors on sanctioning (fraud) probability. Please note that Random Forest models capture a range of non-linear and interacted effects so interpreting relationships between predictors and the outcome is a multi-faceted and complicated matter. To show the impact of each impactful predictor on model output, the latest machine learning literature was followed and calculated Shapley Additive Explanations values (SHAP) (Lundberg and Lee, 2017[6]) and plotted them. SHAP values help to identify the individual contribution of each feature to the model and their importance for the prediction. The Shapley plot in Figure 2.4 displays the probability of sanctions (i.e. likely fraud) as a function of different values of each impactful predictor.

Figure 2.4. SHAP values: Variable importance and effect direction

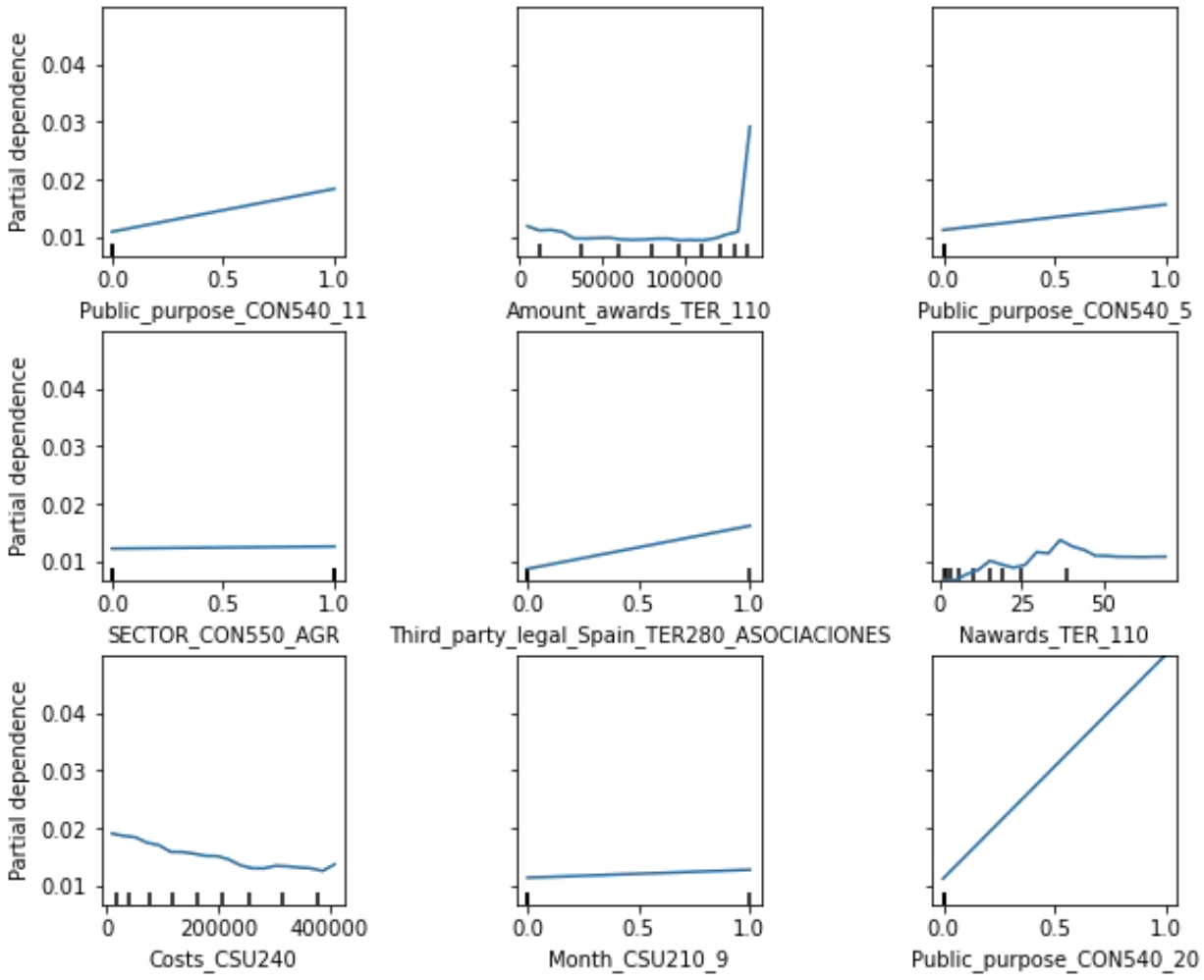


Source: Author

Figure 2.4 highlights that the most significant positive impact on the probability of sanctions is provided by the number of awards, as well as the overall award value received by the same third party. Regarding the rest of the predictors, the probability of sanctions positively correlated with association and limited liability

companies among third parties, as well as with the agrarian sector of the economy. Costs of the grant are negatively associated with the probability of being sanctioned, which means that higher prices of the projects are not correlated with higher risks. On the contrary, public purposes of the award, such as culture (11), social services (5), international co-operation for development and culture (20) and employment promotion (6), are related to higher probability of sanctions. Additionally, grants awarded in September and July are associated with higher chances of sanction, with a similar tendency in October, March and June. More detailed visualisations of important variables' influence on probability of sanctions (fraud) is shown in Figure 2.5.

Figure 2.5. Using partial dependence plots to depict the impact of selected variables on the probability of fraud



Source: Author

Finalising the list of indicators for the risk model

To complete the description of the risk assessment model, the final list of 29 valid indicators used by the model are described according to six groups (Table 2.3), referring to phases during which the potential fraud might occur, or the features of the participating organisations: competition, selection, execution and monitoring phase; grant giving body and recipient organisation (third party).

Table 2.3. Final list of indicators

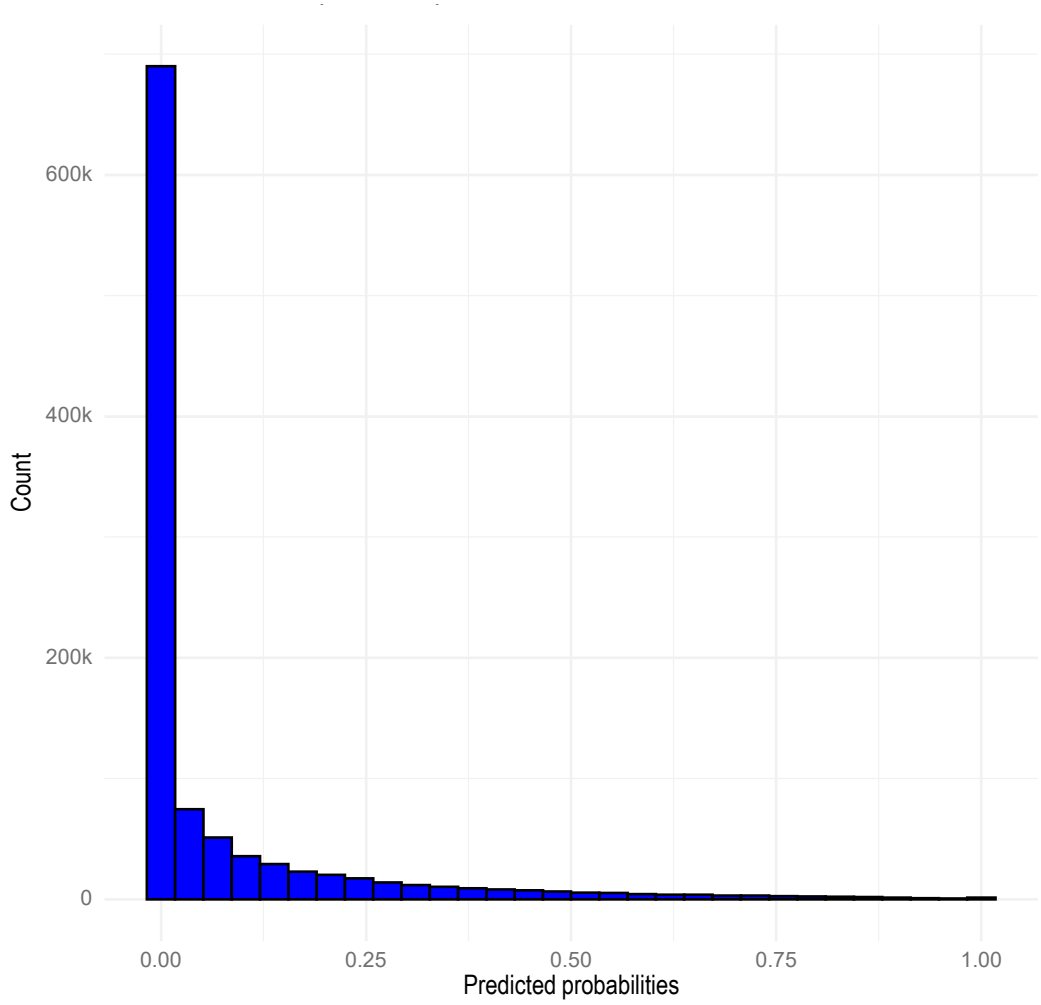
Phase	Variable	Variable description	Frauds
Competition phase	CON420, CON490, CON620	Open admission, Condition of State Aid, Public call	Absence of open admission or public call lead to a less transparent process for monitoring, and therefore are more predisposed to fraudulent activities.
Selection phase	CON540, CON580, CON610, CON630, SECTOR_CON550_AGR...EXTRATER, Month_CSU210	Public purpose, Type of beneficiary, Nominative grant, Gender impact, Sector of economy, Month of award	Type and date of call, sector of economy, as well as type of beneficiary could be correlated with certain fraudulent practices
Grant subsidy/execution	CSU240, CSU220, CSU250, PAG220, PAG230, CON560, LOCAL_IMPL	Nominative grant, Costs, Grant award, Grant aid, Full amount paid, Tax retention, Help instrument, Local Implementation	Grants of high costs could potentially be more predisposed to fraudulent activities. If the implementation takes place in the same location as the grantor, it could be a sign of a corruption scheme.
Grant giving body	NATIONAL_CSU260,REGIONAL_CSU260, MUNICIAPAL_CSU260	Grant award level	Administrative capacities in certain regions could be insufficient for effective monitoring over the call
Recipient organisation	TER100, TER250, TER280, TER290, NATIONAL_TER310, REGIONAL_TER310, MUNICIAPAL_TER310, Amount_awards_TER110, Naward_TER110	Third party country, Third party location, Third party legal status, Third party type, Third party level, Number of awards, Amount of awards	Structure and type of third party organisation, as well as locality could be correlated with fraudulent activities. Parties receiving more awards of bigger size could be potentially more fraudulent than others.
Monitoring	SAN_dum	Sanctioned awards	Captures the fraudulent activity of the third party

Source: Author

Demonstrating results and considerations for further development

The power of the proposed risk assessment methodology is best shown by using the final best Random Forest model to assign a fraud risk score to all awards in 2018 to 2020 with sufficient data quality. Hence, the final distribution of predicted probability of sanctions is presented in Figure 2.6 for the observed 1 050 470 awards. In this broad sample, the model predicts no fraud (sanctions=0) for 1 008 318 awards, while it predicts fraud (sanctions=1) for 42 152 awards using the threshold of 50% sanctions probability for distinguishing between sanctions and non-sanctions.

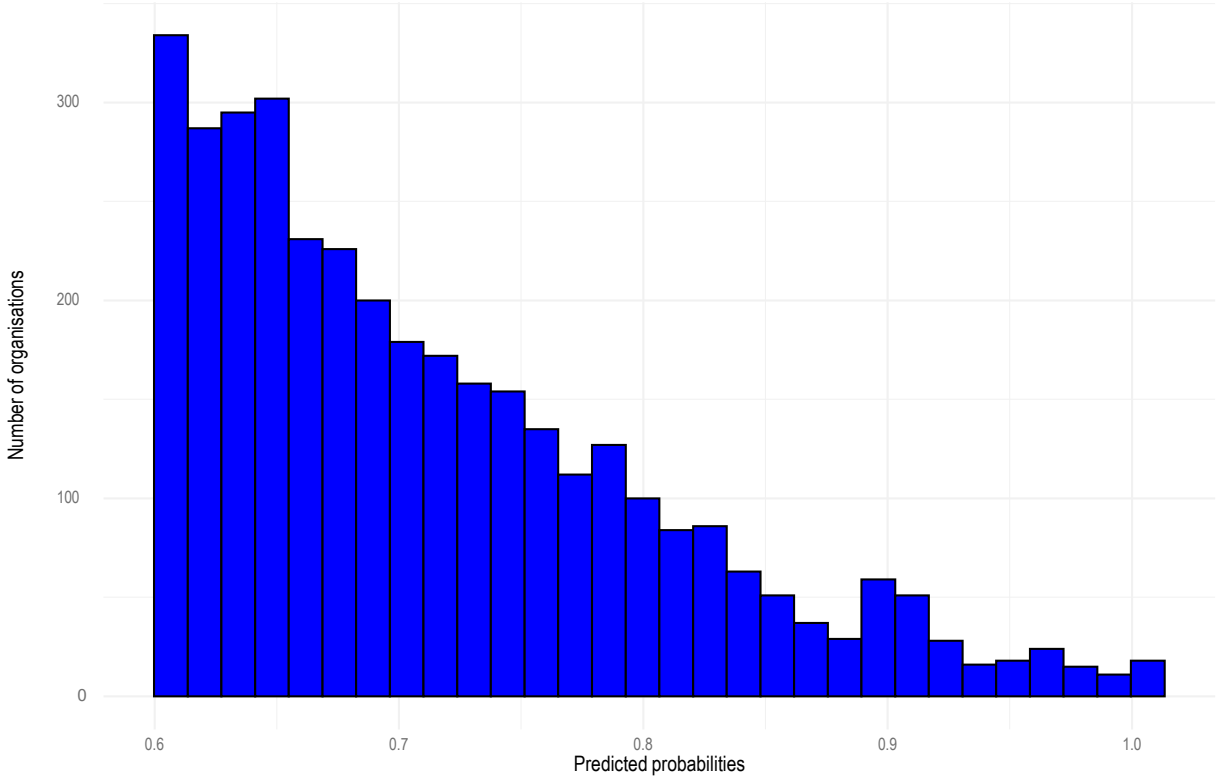
Figure 2.6. Distribution of predicted probabilities for all awards, award level, 2018-2020



Source: Author

Given that risks tend to cluster on the level of organisations and that investigations often look at all the grants received by an organisation, exploring predicted fraud probabilities on the level of grantees adds further value to the model. In order to offer an overview of this level of aggregation, we show the distribution of predicted fraud risks by grantee with high-risk probabilities in Figure 2.7. It shows that among high-risk grantees, risk probabilities are unevenly distributed. The bulk of high-risk grant recipients have features that indicate a 60% to 70% probability of being fraudulent, and a very small group of organisations at the right tail of the distribution have nearly 100 % likelihood of being fraudulent based on the model. These organisations, the top 10 of which are shown in Table 2.4 below, pose the highest risk and are the most suitable candidates for further review and potential investigation based on the predictive model. In addition to these, the organisations the IGAE targets for further investigation would depend on where it sets its risk threshold, and potentially other factors, such as financial implications (see sub-section “Combine predicted risk scores with financial information”).

Figure 2.7. Distribution of average predicted probabilities for high-risk organizations, third-party level, 2018-2020



Source: Author

Table 2.4. Top 10 organisations by average value of awards

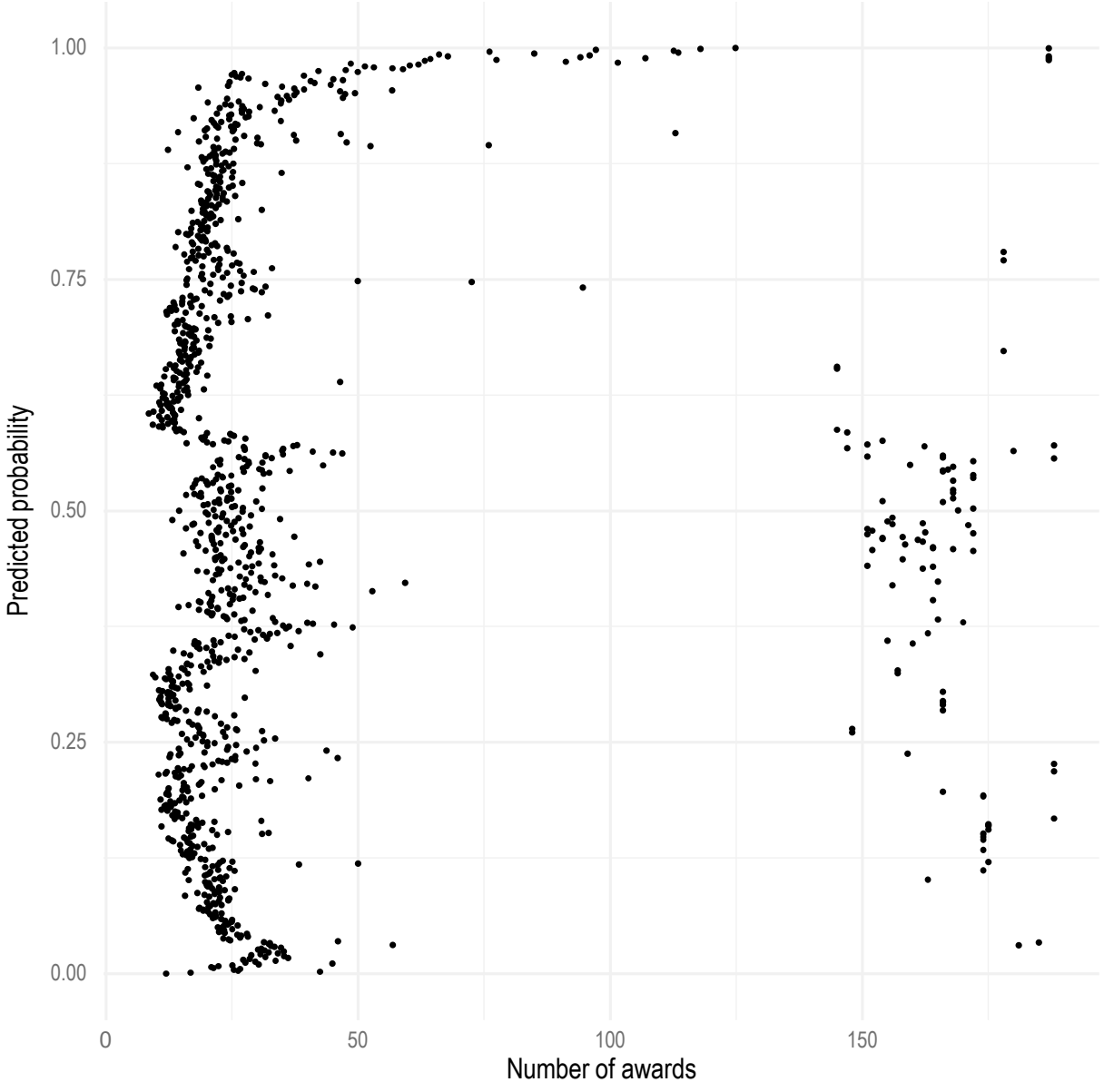
ID Generated	Predicted probability (average per third party)
22568	1
46462	1
60626	1
101336	1
102140	1
129947	1
144235	0.996
152526	0.988
159661	1
167691	1

Source: Author

As expected, the overwhelming majority of awards are estimated by the model as non-risky, yet a few thousands of awards are flagged as risky in addition to the 1 031 observed sanctioned awards. Taking into consideration the most important variables in the model, a closer look is taken at the distribution of predicted fraud probabilities. First, Figure 2.8 shows the distribution of the number of awards received by the same third party in relation to their probabilities of sanction. Interestingly, the model predicts high

sanctioning probability for both large and small entities. The majority of awards are located in the left side of the graph, with 0 to 50 awards per third party and relatively even probabilities of being sanctioned for this group of observations. Starting from 50 awards, the probability increases to almost 100%, with a decrease to around 50% when the number gets over 150 awards. This might be explained by the reliability of third parties—if these organisations are shown to be reliable for a long period of time, they receive more awards as trustworthy ones. Whereas for the first 50 awards the process of evaluation is taking place. It is also conceivable that after a certain threshold of 50 awards per third party, the investigations take place more frequently and therefore potentially sanctioned awards are more likely to appear.

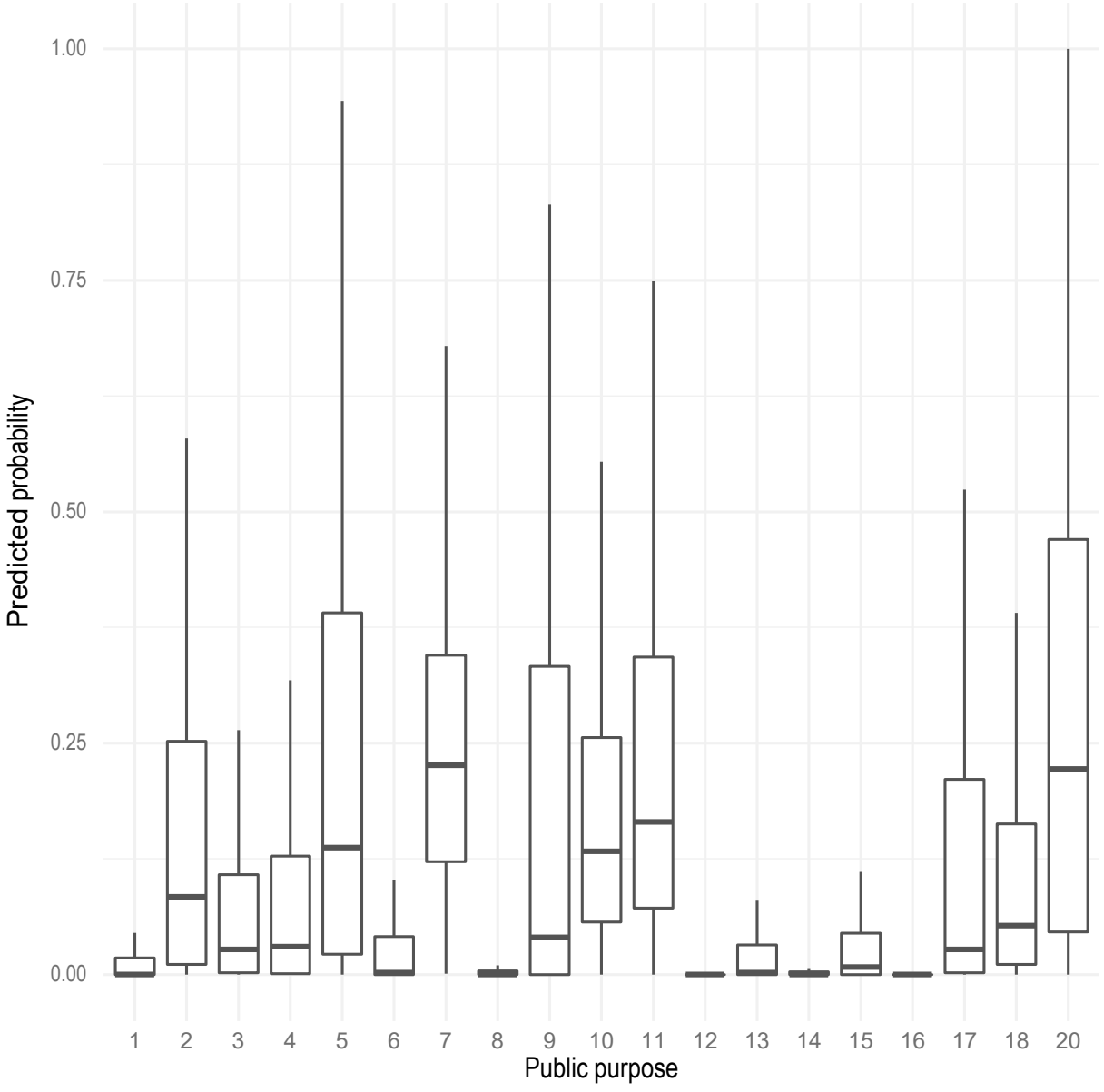
Figure 2.8. Distribution of number of awards by probability of sanctions



Source: Author

Second, coming to another important variable—public purpose of the call—the distribution of predicted probabilities is presented in Figure 2.9. Two categories show the most extended risk of sanctions: social services (5) and international co-operation for development and culture (20). Importantly, these are not the most frequent categories among awards - the most frequent one is agriculture (12), which shows the lowest predicted risk.

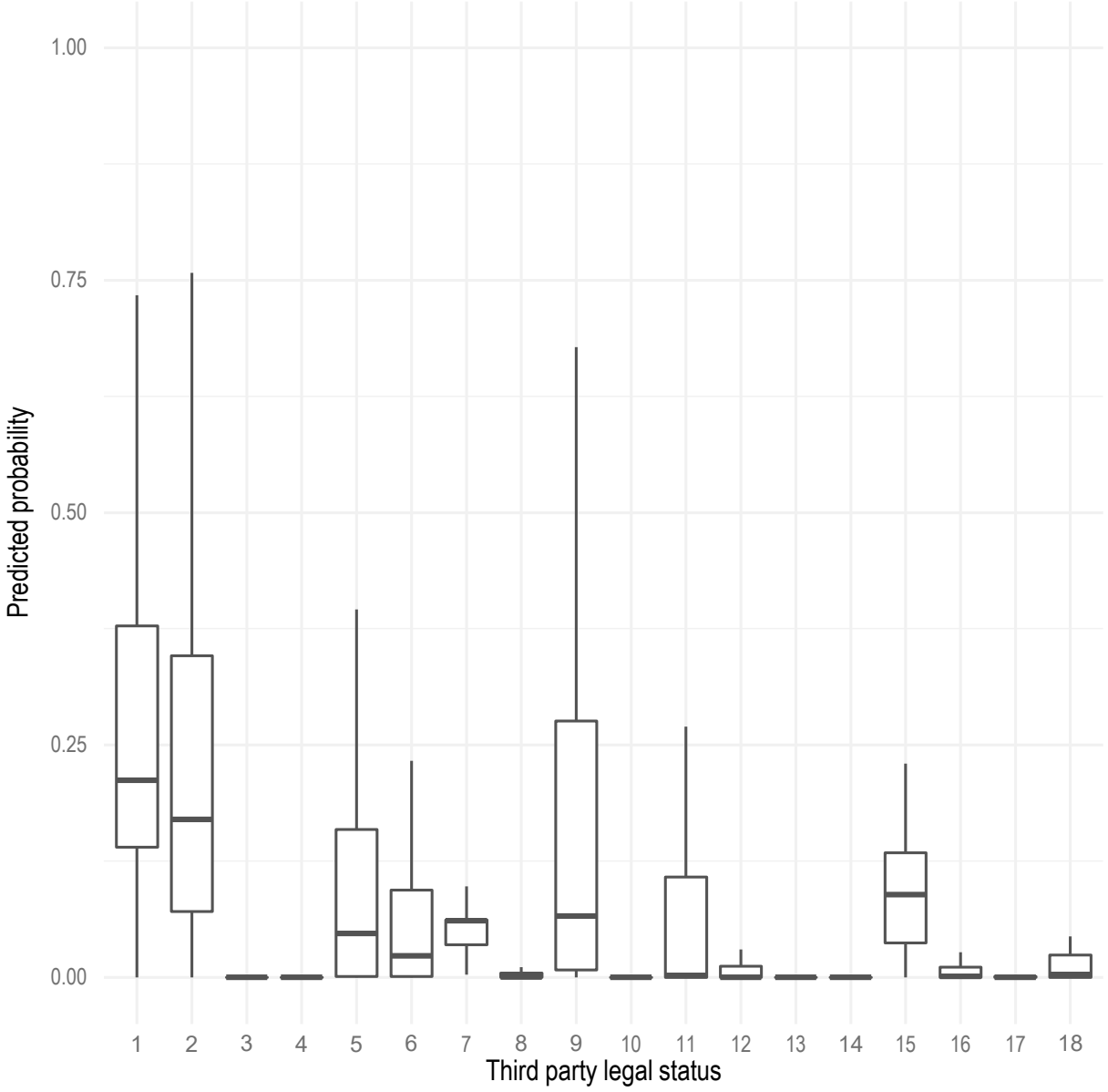
Figure 2.9. Distribution of public purpose of the call over probability of sanctions



Note: 1 - Justice, 2- Defence, 3 -Citizen Security and Penitentiary Institutions, 4- Other economic benefits, 5-Social Services and Social Promotion, 6-Employment Promotion, 7-Unemployment, 8-Access to housing, 9 -Health, 10 -Education, 11- Culture, 12 - Agriculture, Fishing and Food, 13 - Industry and Energy, 14-Commerce, Tourism and SMEs, 15-Transportation subsidies, 16 - Infrastructure, 17- Research, development and innovation, 18 - Other economic actions, 20 - International co-operation for development and culture
Source: Author

Third, the legal status of the third party is another important variable identified by the model (Figure 2.10). The second category - associations - showed significant positive impact on the probability of sanctions in the presented model. Two other types of third parties are prone to higher risks too: bodies of the state administration and autonomous communities (1) and public organisations (9). While association is also the most frequent category for this variable, category 1 and 9 are the least frequent ones yet showing the high probability of being sanctioned.

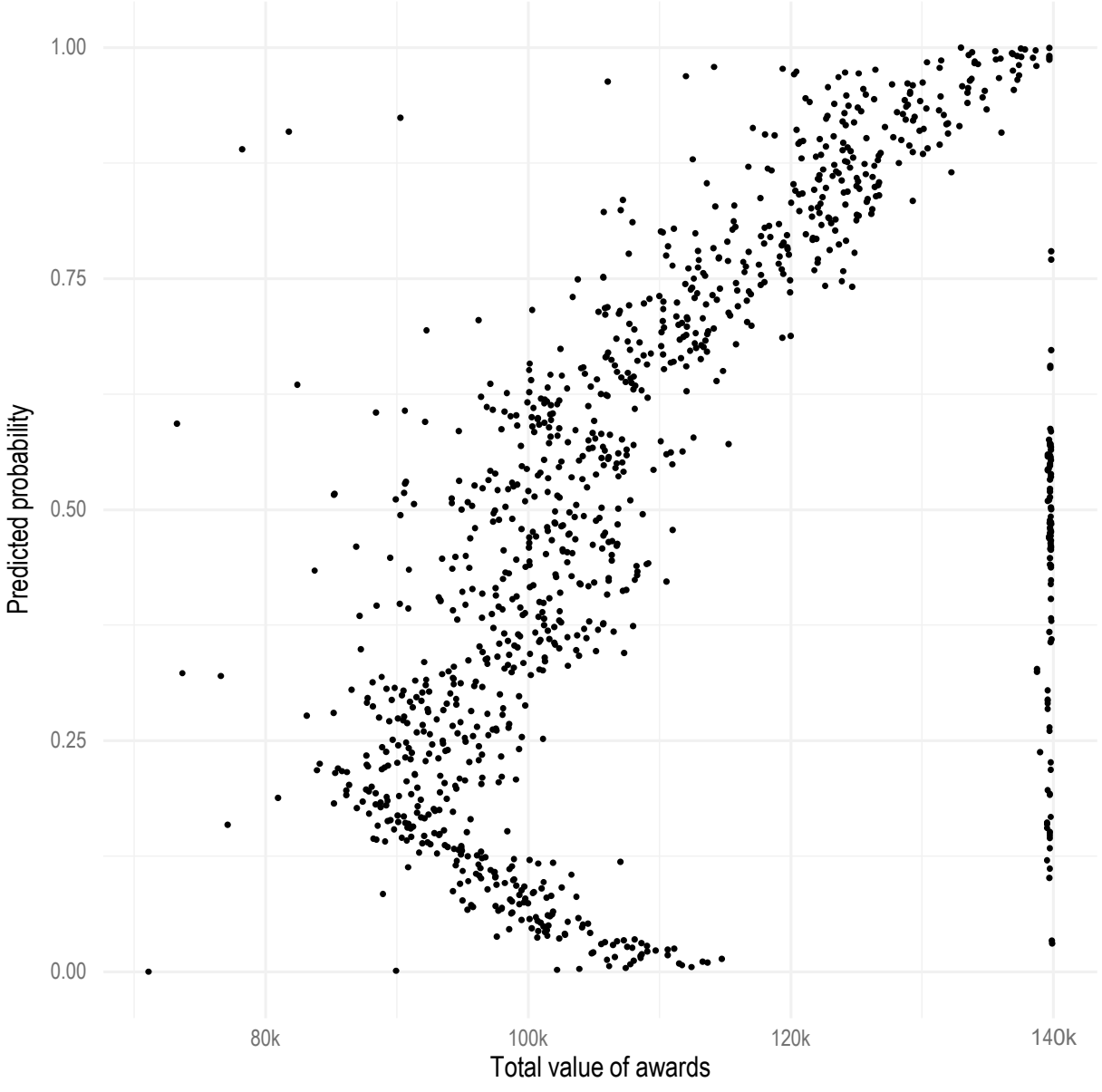
Figure 2.10. Distribution of third parties' legal status over probability of sanctions



Note: 1 - Bodies of the state administration and autonomous communities, 2 - Associations, 3 - Communities of property, inheritances and other entities without legal personality, 4 - Communities of owners under horizontal property regime, 5 - Religious institutions, 6 - Local corporation, 7 - Foreign entity, 8 - Permanent establishment of non-resident entity in Spanish territory, 9 - Public Organisations, 10 - Other types, 11 - Legal person with identification not generated by Spanish authorities (AEAT or Police), 12 - Anonymous companies, 13 - Civil organisations, 14 - Collective organisations, 15 - Commanded companies, 16 – Co-operative companies, 17 - Limited liability companies, 18 - Temporary unions of companies
Source: Author

Finally, the total value of the received awards by the third party has also been found to have a significant impact on the probability of sanctions (Figure 2.11). There is a sustained growth in probabilities of sanction starting from EUR 90 000. Additionally there is a divergence in predicted probabilities in between EUR 85 000 and EUR 110 000, which shows that until EUR 110 000 not all of the awards are risky. Finally, for the maximum total value of observed awards (EUR 140 000), the predicted probability of sanctions is distributed evenly between 0.05 and 0.76. This is very similar to what was observed in the distribution of the number of awards - the highest number was associated with even distribution of risks.

Figure 2.11. Distribution of the overall size of awards received by the same third party over probability of sanctions

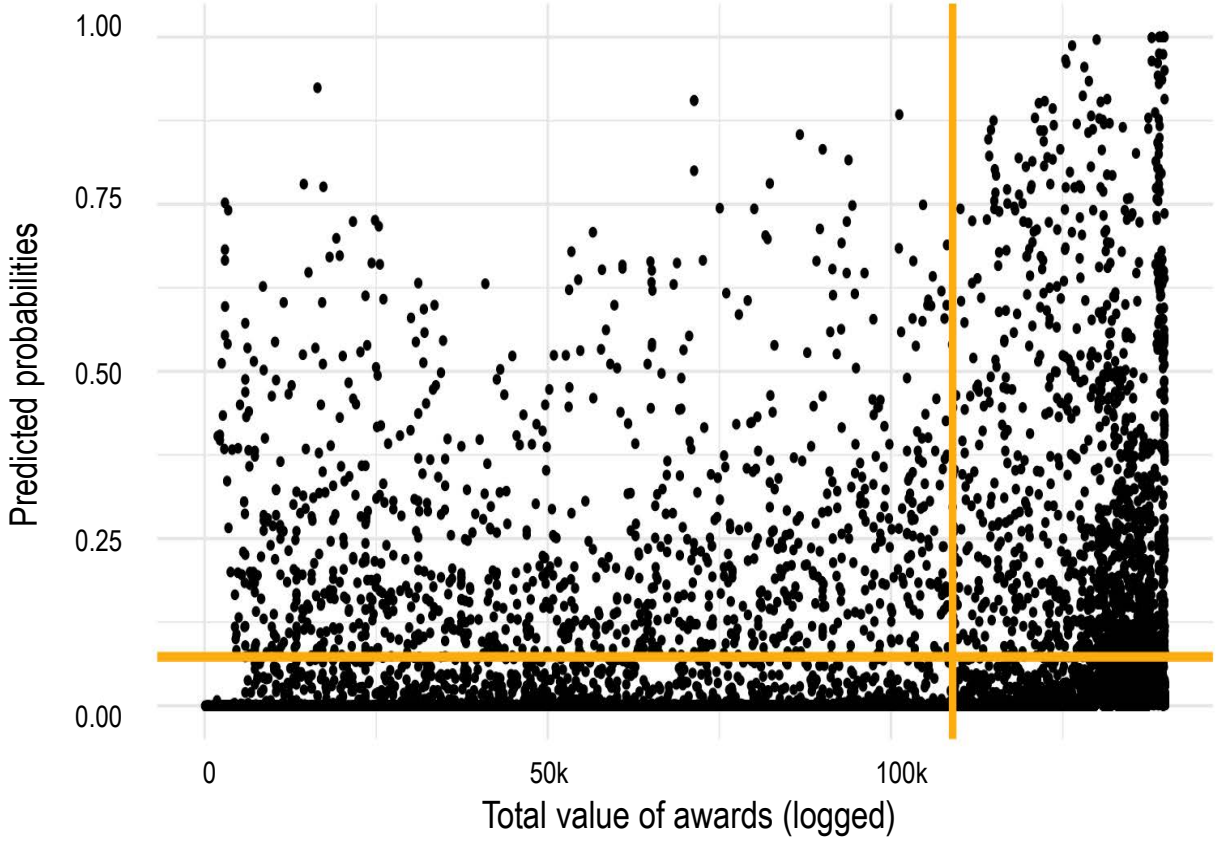


Source: Author

Combine predicted risk scores with financial information

Fraud risks represent the key variable of interest for the IGAE and hence serve as the main dependent variable for the model described so far. Nevertheless, they only represent one of the key dimensions according to which investigative targets can be selected. A second key dimension that could be considered is the total value of the grant as an indication of the potential financial impact of fraud for the Spanish government. Combining the estimated fraud risk scores with the total value of the award allows decision makers and investigators to simultaneously consider the prevalence of risks and their likely financial implications (Fazekas, M., Ugale, G, & Zhao, A., 2019[7]). The simplest approach to look at these 2 dimensions simultaneously is to draw a scatterplot with these 2 variables also highlighting their average values (Figure 2.12). The top right quadrant includes those awards which not only have high risk but also have high award values. This is the group of greatest interest for the IGAE's future investigations as they are most likely to include fraudulent grants with large financial implications.

Figure 2.12. The distribution of awards by predicted risk score and total award value



Source: Author

Establish a ready-made dataset for future fraud risk detection

In order to further improve the data-driven fraud risk assessment framework of IGAE, a number of short and medium term reforms could be implemented which would enhance the quality and scope of data underlying the risk models. The OECD's development of a risk model has already helped the IGAE to advance in addressing some of these issues, and the resulting dataset from the OECD's work can be a

starting point for the IGAE. Nonetheless, datasets are not static and new data sources may become available. Therefore, these points are relevant outside of the context of this project.

First, existing data could be better and faster combined into a single dataset ready for fraud risk modelling. Currently, almost all the datasets cover different units of analysis such as awards, calls or organisations. In order for the IGAE to merge them, each dataset should be aligned to the same level with unique IDs to avoid redundant multiplication of observations in the merged dataset. During data processing for this report, the data was transformed from long to wide format whenever it was needed. Yet this approach suffers from a major drawback, that is a high missing rate for IDs without multiple observations per single ID. To solve this problem aggregation is needed, especially for factor variables which cannot be calculated as means or medians.

Second, lowering missing rates on all variables collected by IGAE is paramount. As discussed in Chapter 1, defining data quality standards and enforcing them, in collaboration with the relevant data owners, would assure that there are no variables with high missing rates such as 40-50%. Third, some datasets (for instance, on projects) consist of a very small number of observations, which prevents their analysis in conjunction with the main dataset (i.e. when matched they result in a very high missing rate). Similarly, data on third parties is very limited and needs further enhancement.

Expand the IGAE's use of indicators across the grant cycle

As noted, the final list of indicators includes 29 variables. These variables are mostly categorical, although there are several numeric variables, such as costs and payments. Most of the variables analysed are descriptive due to the data available. There is limited data that can provide insights into the behaviours of organisations and people, such as conflicts of interests among actors who receive or benefit from the grants. This is one of the largest gaps in the current data available to the IGAE and is one of the most limiting factors on its risk analysis, regardless of the methodology taken. Table 2.5 shows additional behavioural indicators that could be used for fraud risk assessment that span the grant cycle and could help to refine the IGAE's risk model.

Table 2.5. Behavioural indicators for assessing fraud risks for each phase of the grant cycle

Indicator group	Indicator name	Indicator definition
Competition phase	lack of competition	Only one applicant for a grant competition
Selection phase	spending concentration	Excessive number and value of payments to a single vendor
	political influence	Only beneficiaries linked to the government get their applications granted
Grant/subsidy execution phase	excessive payments	Salary or other compensation for personal services that exceed agency-approved amounts, or are higher than compensation for other comparable services that are not grant-funded
	large early payments	A recipient that draws all or most of the grant funds shortly after the grant is awarded may be characteristic of a higher risk, unless this practice is allowed by the grant programme
	Deadline amendment	Request by the contractor to amend deadlines and contract terms
	Large transaction	One single transaction accounts for more than half of the total project costs
	Late expenditures	Expenditures outside of the allowable project period
	Unusual transactions	Questionable or unusual transactions immediately preceding the end of a grant award period may indicate that fraudsters waited until the end of the project to draw down grant funds to cover unallowable costs
Recipient organisation	new company	Final beneficiary set up immediately prior to the application for the subsidy
	double funding	Evidence that recipients are funding grant projects with more than one grant

Indicator group	Indicator name	Indicator definition
	Financial viability	A recipient that has questionable financial viability, such as a high percentage of assets funded by debt or insufficient cash flow
Contractors and Consultants	non-competitive procurement	Recipients that expend funds on unapproved purchases, or unapproved sole-source or no-bid procurement sub-contracts
	consultant subcontracts	The use of generic, non-specific, or nebulous consultants
	insufficient documentation	Insufficient justification and documentation for payments made to contractors/consultants, such as hours worked and activities
Monitoring and audits	audit queries	Multiple queries from law enforcement or audit offices which cannot be answered
	non-cooperation with auditors	Recipient staff who are unco-operative with monitoring activities or aggressive towards grant auditors or managers

Source: Author

There are a variety of sources and examples that can support the IGAE in refining its risk indicators. In the European Union, the European Anti-Fraud Office (OLAF) created a Compendium of Anonymised Cases in 2011 that still has relevance today. The Compendium lists the results of OLAF's investigations and includes information about financial frauds. Two high-risk phases of potential fraudulent behaviour could be identified—the selection phase and execution phase. During the selection phase, OLAF encouraged a close inspection of supporting declarations and official documentations, as well as to make sure the final beneficiary was not set up or created immediately prior to publication of the subsidy. During the execution phase, OLAF suggested consideratoin of the financial difficulties of the contractor, single big transactions covering almost half of all project costs, as well as the use of subsidies for other purposes (European Anti-Fraud Office (OLAF), 2011[8]). The Compendium illustrates the reality that a lot of fraud is simply recycled versions of similar schemes. Indeed, in its *32nd Annual Report on the protection of the European Union's financial interests — Fight against fraud-2020,* the European Commission noted that among the fraudulent irregularities related to healthcare infrastructure and the COVID-19 pandemic, the most commonly detected issues concerned supporting documentation (European Commission, 2021[9]). Box 2.2 provides additional insights from the experience of the Grant Fraud Committee of the Financial Fraud Enforcement Task Force, which was set up to tackle fraud in the wake of the 2008 financial crisis.

Box 2.2. The Grant Fraud Committee of the U.S. Financial Fraud Enforcement Task Force

In the United States, the Grant Fraud Committee of the Financial Fraud Enforcement Task Force identified several key areas to monitor and identify fraudulent activities:

- structure of recipient organisation and grant program
- payment requests or drawdown of grunt funds
- monitoring reports and activities
- transaction-level activities
- contracts and consultants.

Among the first category, the Grant Fraud Committee suggested monitoring the design of the project, as well as financial viability of the recipient, internal control, organisations' personnel and potential conflicts of interest. In regards to payment requests, the attention should be paid towards timing of grant drawing, as well as justifying documentation, exceeding expenditures and rounding the numbers for grants drawing. While performing monitoring activities, the responsiveness and co-operation of the recipient is a key indicator, as well as presence of internal controls and audit history of the company. When it comes to transaction-level activities, excessive, unusual and unmonitored transactions could be marked as potential risks, as well as double funding (more than one grant covering the same project).

Finally, in regards to contracts and consultants, the Grant Fraud Committee suggests looking at related party transactions, spending on non-specific consultants and grants recipients with deficiencies in their procurement systems. In case of data monitoring, Grant Fraud Committee (2012) identifies the following frauds risks:

- excessive number and value of payments to a single vendor
- payments to unapproved vendors
- transactions that bypass normal review procedures, or are otherwise unmonitored or reviewed by another person
- purchases that appear illogical considering the nature of the grant programme
- expenditures outside of the allowable project period
- checks and transactions that occur several times per month
- checks issued to multiple vendors at the same address.

Note: In 2018, the Task Force on Market Integrity and Consumer Fraud replaced the Financial Fraud Enforcement Task Force.
Source: (Financial Fraud Enforcement Task Force, 2012[10])

Invest in continuous improvement of the risk model

As the validity of data-driven models depends on the completed sanctioning activities, if sanctions missed out on relevant fraud schemes or resulted in a biased sample of investigations, any risk assessment model would also be biased. Hence, obtaining a truly random sample of investigations and sanctions is of central importance. To this end, the IGAE could select a percentage of the investigated cases each year using the IGAE's traditional sampling techniques, or a data-driven selection method such as the one presented above. The remainder of the investigated cases could be picked by complete random selection. This approach would strike a balance between maximising the utility of investigative resources through better targeting, while also investing in future improvements of the risk assessment model by providing a better training sample. It would also give the IGAE a better sense of how the model is performing. As this effort was a proof-of-concept, additional technical steps could be considered:

- Improve the quality of variables in the full dataset in order to be able to include more indicators in the model and hence improve model quality.
- Take into consideration the unbalanced nature of classes in the dependent variable (positive/negative): use PU bagging techniques to avoid inaccuracy in modelling.
- Repeat the modelling exercise regularly as new data, including sanctions as well as awards, become available in order to keep the risk assessment up-to-date.
- Analytical models do not paint a complete picture and can have biases as they learn from past enforcement action (see Chapter 1). The IGAE could supplement the models with qualitative methods and expert judgement. This allows fraud specialists at IGAE to contribute with their expert understanding of fraud schemes, latest events, and the broader context.

While models may be vulnerable to biases themselves, they can also help to control for biases. Specifically, data-driven sample selection, including those that use machine learning, not only would help the IGAE to maximise the efficacy of control resources, but it would also help to correct for some biases in the learning dataset. For example, if fraud types which are not covered by investigations are known, their features can be manually entered into the database to provide sufficient input for the algorithm to learn from. Moreover, if the selection of investigations in the sanctions dataset emphasises certain variables, say the size of the grant, under-sampling large grants and oversampling small grants can counteract biased selection of investigated cases.

Consider network analyses and making use of a broader set of methodologies

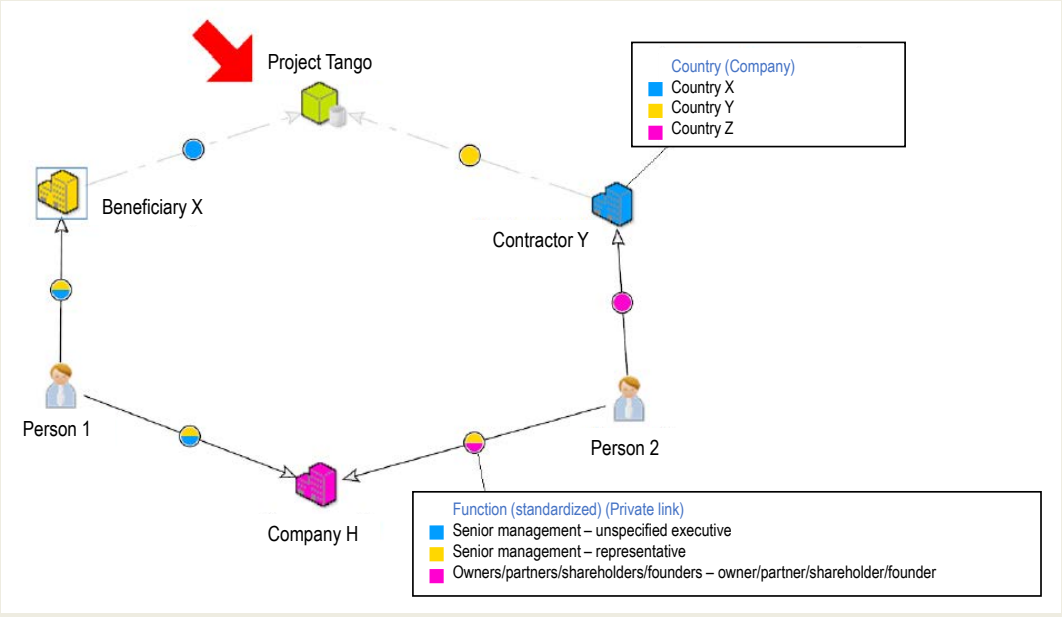
Network and data science techniques have been increasingly used to study economic crime such as corruption, fraud, collusion, organised criminality, or tax evasion to mention a few major areas (Wachs, Fazekas and Kertész, 2020[11]). Exploring networks without advanced analytics already promises great advantages for fraud detection such as tracing potential conflicts of interest (see Box 2.3).

Box 2.3. Using data to investigate conflict of interest

When the individuals behind public and private organisations parties to the grant making and implementation process are known, a range of potential conflicts of interest relationships can be uncovered. While in-depth investigations can reveal such relationships, risk screening is greatly facilitated by matching large-scale datasets containing: 1) all public officials playing a significant role in preparing, assessing, awarding, and monitoring grants and subsidies; and 2) all private officials playing a significant role in the companies which submit grants applications, receive and implement grants.

Gathering, cleaning and linking such datasets and maintaining the underlying data pipelines involve can come with considerable costs. However, once such a dataset and a simple graphical interface is available, which is the case for the EU's ARACHNE tool, it can greatly speed up the screening and investigation of risky relationships among grant makers and grant recipients. For example, it is possible to quickly and efficiently look at projects, the implementing contractors and the persons participating in the preparation of the call and assessment of applications.

Figure 2.13. Visualising conflicts of interest



Source: (European Union, 2016[12])

Analysing large-scale networks of contractual or personal relations can reveal hidden patterns which serve as risk indicators on their own or complement other risk indicators (Fazekas and Tóth, 2016[13]); (Fazekas and Wachs, 2020[14]). For example, tendering risk indicators such as the incidence of single bidding in public procurement can be superposed over clusters of linked buyers and suppliers in public procurement

to identify high risk cliques. Figure 2.14 below shows a visualisation of the public procurement contracting market of buyers and suppliers in Hungary. Such diagrams provide a visual snapshot of the data that signal potential high-risk relationships for further investigation. For instance, the red lines highlight a higher than average single bidding rate in that relationship. In addition, there is a cluster of high corruption risk actors in the top (i.e. dense organisational contracting relationships coinciding with high single bidding rates in those relationships).

Figure 2.14. Hungarian public procurement contracting market of buyers and suppliers in 2014

Source : (Wachs, Fazekas and Kertész, 2020[11])

The IGAE could compile relevant datasets such as company ownership data and link them to its core grants and subsidies data in order to make use of such network analytic techniques. As individuals move across the private and public sectors and there are a range of other ways grantees can establish connections to grant making bodies, tracing overt or hidden networks offers a key tool for improving fraud risk assessment in Spain. As discussed, this is one area where the IGAE currently has gaps in its data, and so use of network analyses will also depend on the ability of the IGAE to address these gaps. Recent developments in Spain suggest that improvements are already underway. For instance, in May 2020, the IGAE and the General Treasury of Social Security (*Tesorería General de la Seguridad Social*, TGSS) signed an agreement on the transfer of information, establishing more collaborative conditions for financial control of subsidies and public aid. The agreement stipulates direct access to the TGSS databases to facilitate the IGAE's work to detect fraud and irregularities (Ministry of the Presidency of Spain, 2021[15]). Advancing with similar agreements with other public and private entities, particularly to obtain company data and data that reflects behavioural indicators as discussed above, would be critical inputs for strengthening future risk models.

Conclusion

This chapter presents a proof-of concept for the IGAE to enhance its approach for assessing fraud risks in public grant data, drawing from leading practices in analytics. The process of developing the risk model led to a number of insights about the IGAE's current capacity for analytics, as well as data management and ensuring the quality of data for purposes of assessing fraud risks, as noted in Chapter 1. The development of the risk model also demonstrated gaps in fraud risk indicators and databases that, if addressed, could help the IGAE to improve its fraud risk assessments regardless of the specific methodology it chooses. In particular, the IGAE could incorporate additional behavioural indicators for assessing fraud risks across each phase of the grant cycle, drawing from international experiences and academic literature. In addition, the IGAE could compile company data and adopt the methodologies described for conducting network analyses as a means for identifying conflicts of interest, drawing from the procurement examples in this chapter.

The chapter also covers a number of technical considerations for the IGAE if it decides to adopt a machine learning model. Much of the heavy lifting has been done as part of this pilot in terms of data processing and data cleaning. The IGAE now has a working dataset to use for fraud risk analysis that is already an improvement on what it had available prior to the projects. The features and limitations of the IGAE's data drove much of the rationale for selecting the approach described. While it has limitations due to the quality of the learning datasets and features of the grant data, the methodology was designed to minimise false positives and false negatives, and overall, it has high predictive power for identifying potential fraud in Spain's public grant data. While implementing this approach requires additional capacities, detailed in Chapter 1, the proof-of-concept successfully demonstrates what is possible with a modest investment and provides a basis for the IGAE to adopt a truly data-driven fraud risk assessment. Chapter 3 explores further how the IGAE can improve the accuracy of the model by integrating additional data that can be used for detecting possible fraud.

References

Breiman, L. (2001), "Random Forests", *Machine Learning*, Vol. 45/1, pp. 5-32, https://link.springer.com/article/10.1023/a:1010933404324. [1]

Elkan, C. and K. Noto (2008), "Learning classifiers from only positive and unlabeled data", *KDD '08: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 213-220, https://dl.acm.org/doi/10.1145/1401890.1401920. [2]

European Anti-Fraud Office (OLAF) (2011), *Compendium of Anonymised Cases*, https://ec.europa.eu/sfc/sites/default/files/sfc-files/OLAF-Intern-2011.pdf (accessed on 13 August 2021). [8]

European Commission (2021), *32nd Annual Report on the protection of the European Union's financial interests: Fight against fraud 2020*, https://ec.europa.eu/anti-fraud/sites/default/files/pif_report_2020_en.pdf (accessed on 13 August 2021). [9]

European Union (2016), *Arachne, Be Distinctive*, http://www.ec.europa.eu/social/BlobServlet?docId=15317&langId=en (accessed on 13 August 2021). [12]

Fazekas, M., Ugale, G, & Zhao, A. (2019), *Analytics or Integrity: Data-Driven Decisions for Enhancing Corruption and Fraud Risk Assessments*, OECD Publishing, Paris, https://www.oecd.org/gov/ethics/analytics-for-integrity.pdf. [7]

Fazekas, M. and I. Tóth (2016), "From Corruption to State Capture: A New Analytical Framework with Empirical Applications from Hungary", *Political Research Quarterly*, Vol. 69/2, pp. 320-334, https://scholar.google.com/citations?view_op=view_citation&hl=en&user=H1FpS2AAAAAJ&citation_for_view=H1FpS2AAAAAJ:SP6oXDckpogC. [13]

Fazekas, M. and J. Wachs (2020), "Corruption and the network structure of public contracting markets across government change", *Politics and Governance*, Vol. 8/2, pp. 153-166, https://scholar.google.fr/citations?view_op=view_citation&hl=fr&user=PY3YH2kAAAAJ&citation_for_view=PY3YH2kAAAAJ:ZeXyd9-uunAC. [14]

Financial Fraud Enforcement Task Force (2012), *Reducing Grant Fraud Risk: A Framework For Grant Training*, https://www.oversight.gov/sites/default/files/oig-reports/Grant-Fraud-Training-Framework.pdf. [10]

James, G. et al. (2015), *Chapter 8*, https://link.springer.com/book/10.1007/978-1-4614-7138-7. [5]

Li, C. and X. Hua (2014), "Towards positive unlabeled learning", *International Conference on Advanced Data Mining and*, pp. 573–587. [3]

Lundberg, S. and S. Lee (2017), *A unified approach to interpreting model predictions*, https://scholar.google.com/citations?view_op=view_citation&hl=en&user=ESRugcEAAAAJ&citation_for_view=ESRugcEAAAAJ:dfsIfKJdRG4C. [6]

Ministry of the Presidency of Spain (2021), *Resolution of May 26, 2020, of the Undersecretariat, which publishes the Agreement between the General Treasury of Social Security and the General Intervention of the State Administration, on the transfer of information*, https://www.boe.es/diario_boe/txt.php?id=BOE-A-2020-5748 (accessed on 4 July 2021). [15]

Mordelet, F. and J. Vert (2014), "A bagging SVM to learn from positive and unlabeled examples.", *Pattern Recognition Letters*, Vol. 37, pp. 201-209, https://www.sciencedirect.com/science/article/pii/S0167865513002432. [4]

Wachs, J., M. Fazekas and J. Kertész (2020), "Corruption risk in contracting markets: a network science perspective", *International Journal of Data Science and Analytics*, pp. 1-16, https://scholar.google.fr/citations?view_op=view_citation&hl=fr&user=PY3YH2kAAAAJ&citation_for_view=PY3YH2kAAAAJ:QIV2ME_5wuYC. [11]

Notes

[1] For cleaning and merging the data, R 3.6.3 was used with the following package: readxl, tidyverse (dplyr), flipTime, tibble, data.table. For modeling both R 3.6.3 and Python3 softwares were used for different stages of analysis. To build random forests in R, randomForest and xgboost packages were used. For positive unlabeled learning in Python3 libraries pandas, numpy, baggingPU (module BaggingClassifierPU), sklearn.tree (modules DecisionTreeClassifier, DecisionTreeRegressor, precision_score, recall_score, accuracy_score, train_test_split, RandomForestClassifier) were employed.

[2] Names of these datasets include BDNS_CONV_ACTIVIDADES, BDNS_CONV_ANUNCIOS, DNS_CONV_FONDOS_CON690, BDNS_CONV_OBJETIVOS_CON503, BDNS_CONV_TIPOBEN_CON590, BDNS_CONV_REGIONES_CON570, BDNS_CONV_INTRUMENTOS_CON560.

[3] Names of these datasets include BDNS_PROYECTOS, BDNS_PAGOS, BDNS_REINTEGRO, BDNS_DEVOLUCIONES.

[4] Names of these datasets include BDNS_INHABILITACIONES, BDNS_SANCIONES and BDNS_TERCERO_ACTIVIDADES_TER320.

[5] The very high accuracy rate of 95% is largely due to the fact that the sample is imbalanced, that is most cases are negative (non-sanctioned) and the model hence relatively easily can classify the bulk of the sample as non-sanctioned. However, it is harder for the model to predict sanctioned cases correctly given that they are so much more rare. For this case, recall score is more helpful to evaluate the model performance, as it calculates the number of members of a class that the classifier identified correctly divided by the total number of members in that class.

3 Looking ahead: A roadmap of datasets to enhance the fraud risk model of Spain's Comptroller General

This chapter explores additional datasets that the General Comptroller of the State Administration (Intervención General de la Administración del Estado, IGAE) of Spain can use to enhance the risk model described in Chapter 2. The chapter provides a road map and indicates which databases are most promising for improving the assessment of grant fraud risks using the model, based on the accessibility, relevance and quality of the datasets. The datasets are grouped into three categories: 1) organisational data on the parties of the granting process; 2) data on personal connections and conflicts of interest; and 3) data on organisational reliability and violation of rules.

Introduction

This chapter offers a roadmap for complementing existing grants data of the General Comptroller of the State Administration (*Intervención General de la Administración del Estado*, IGAE) in order to improve risk assessment models. By implication, it outlines priority datasets which can be linked to existing IGAE grants data, enhancing analytical sophistication and improving the precision of risk assessment. As discussed in Chapter 2, machine learning models are limited by the scope and type of data included in the training sample. The model cannot precisely estimate risk probabilities based on incomplete information, because key drivers and mechanisms determining risks remain unaccounted for. Hence, the more comprehensive the initial dataset is, the more precise and accurate risk calculations become.

As the universe of potentially relevant datasets is vast, it is imperative to narrow down the list of datasets to the most relevant ones before investing considerable resources into data mapping, processing, linking and eventually incorporating into the predictive models. Three factors should be considered when selecting suitable datasets: *accessibility, relevance, and quality. Accessibility* in this context encompasses the ease with which the dataset can be gathered from its original source, which can include questions such as whether the dataset is publicly downloadable or it has to be requested. The format in which the data are available is also crucial, such as a single downloadable dataset or a series of HTML pages. *Relevance* refers to the potential of the data fields to improve analytical sophistication and precision. This has to be assessed before actually collecting the data. The ultimate test of this initial assessment is whether the data would improve the predictive accuracy of the model. When too many redundant variables are included, the final model may suffer from overfitting. Data *quality* in this context captures the rate of non-missing values and the reliability of information. Low quality data with many missing values or inaccurately collected data are likely to bias the results. This chapter will only cover the datasets that are considered to be readily available to the IGAE, relevant for the said risk model and of sufficiently high quality.

Roadmap for complementing IGAE grants data

The two previous chapters outlined the process by which machine learning can be deployed to enhance the IGAE's approach to identifying risks in grants and subsidies provision. The process of drawing on external datasets in addition to the existing internal data follows the same logic. First, background and risk indicators should be defined for each dataset to identify factors that potentially influence fraud risks. The next step is to link datasets to the existing internal dataset. In order to do so, a few things should be taken into consideration: the unit of analysis in each dataset, variable relevance, the missing rate and the variance. As discussed in Chapter 2, the missing rate should be lower than 50%, with variance of at least 35%. Moreover, to merge the new data it should be aligned to the same unit of analysis with unique IDs to avoid duplicative rows after matching. Variables that do not contain useful information (i.e. cannot be used as indicators) should be dropped.

For example, to add external datasets to the existing National Subsidies Database (*Base de Datos Nacional de Subvenciones*, BDNS), they should have identifiers matching with the ones in the BDNS data. Such IDs include identifiers of grants, Tax Identification Number (NIF) of beneficiaries and grantor names, such as municipality names. This implies some limitations, for example, it is currently impossible to match third parties by their names, and instead they can be matched only by NIFs. Additionally, matching by municipality will lead to a significant data loss, because aligning data to the same unit of analysis with unique IDs means that risk scores should be aggregated by municipality. Similar logic applies to matching by grantors' names and beneficiaries' NIF, as there are many identical values throughout the BDNS data (i.e. the same beneficiary might receive multiple grants or subsidies).

There are a few sources—some more reliable than others—that can be potentially used for adding data to the existing BDNS dataset. First, there are official sources such as the National Register of Associations

(*el Registro Nacional de Asociaciones*) of the Ministry of the Interior, which lists accredited non-governmental organisations (NGOs), the tax database of the State Tax Administration Agency (*Agencia Estatal de Administración Tributaria,* AEAT) and the Spanish Association of Foundations (*La Asociación Española de Fundaciones*), which lists accredited foundations. Some of the data are publicly accessible, whereas others are restricted only to authorised agencies.

Beneficial ownership (BO) registries and public procurement data can also be considered as trusted official sources. The advantage of working with official data directly obtained from data holders is that there is no need to verify the information provided, beyond the standard data quality checks used as part of the outlined data pipeline. Official aid data from the European Union is another example of trustworthy data.

The next group of sources are independent NGOs and associations. This information is less reliable, since the process of data collection and verification is unclear. While official sources most likely include primary data and information, secondary sources are either parsed from different sources or collected manually, often without transparency concerning how the dataset is constructed. Therefore, these datasets should be used with greater care and their validity checked more thoroughly. In Spain, examples of such sources include independent NGO evaluators as well as *FICESA*, a database of Spanish senior positions and secretariats.

Overview of the most relevant dataset groups

There are four major groups of data that are relevant for matching with the main BDNS database in order to enhance the IGAE's fraud risk assessments. Each group can provide insights on distinctly different dimensions and determinants of fraud risks. Some data creates opportunities for alternative methods of analysis, such as network analysis, revealing connections between private companies and politically exposed persons, as well as beneficial owners and associated companies. Bringing all of these datasets together offers the possibility of the most comprehensive risk assessment; however, matching only some, or even just one additional dataset, can be very useful for enhancing the IGAE's risk model, including the following groups of data:

i. **Organisational data on the parties of the granting process**. This group covers data on grantors and grantees, as well as third parties (i.e. project implementers). Potential sources of information for this group are:

- Company registry and financial information: provides information on the organisational structure and history of the company (e.g. when it was founded) and also uncovers the financial situation such as profitability of the organisation.
- Organisational data on accredited NGOs, foundations, associations: provides information on the registry features, reliability of the organisation, and financial records.

ii. **Data on personal connections and conflicts of interest**. This group can be helpful in identifying connections between officials in private organisations applying for grants and political officeholders overseeing grant giving. Connecting public and private office holders can be useful for further investigating possible conflicts of interest. Potential sources of information for this group are:

- The BO registry: can help with identifying beneficial owners, associated companies and their records.
- Politically exposed persons: helps in revealing people who are entrusted with power and are more susceptible to being involved in bribery or other corrupt practices.
- Data on senior positions and secretariats: provides names of people potentially connected to private companies through legal or beneficial ownership.

iii. **Data on organisational reliability and violation of rules.** This group can aid in predicting fraud risks by offering insights on relevant, but only indirectly related violations, such as tax payment irregularities. This group can also provide information on softer measures of reliability, such as civil society accreditation. Potential sources of information are:

- Data on bankruptcy or tax payments: shows the reliability of an organisation based on past financial records:
- Accreditations of NGOs: identifies accredited NGOs or other associations as more reliable ones.

iv. **Data on other funds and contracts**. Information on other funding sources and public contracts can reveal additional factors that influence the likelihood of fraud, such as double funding for the same activity. Moreover, corruption risks in public procurement or other funding processes can point to systematic, organisation-level weaknesses and the propensity to commit fraud. The relevant datasets in this group include:

- EU Funds: list of beneficiaries of EU aid can show if the organisation received double funding from different sources for the same project.
- Public procurement: corruption risks in public contracts received by organisations or provided by the same grantor can influence the possibility of wrongdoing in grants and subsidies.

Table 3.1 presents the most promising datasets in Spain which are either publicly accessible or their content and specifications are in the public domain. For each dataset belonging to one of the 4 dataset groups, the table contains information on the unit of measurement (i.e. what does a single row refer to), number of observations where available, ID for matching to the BDNS,[1] and the priority for the IGAE's follow-up work. The table highlights the top priority datasets on the top, considering the three main dimensions of data assessment discussed above: accessibility, relevance, and quality. Only datasets that scored high on all 3 dimensions—readily *accessible* bulk data download, highly *relevant* data scope and content, and adequate *quality*—were considered as high priorities for the IGAE.

Conversely, some datasets that scored high on only one or two dimensions were rated as medium or low priority. For instance, when data accessibility was limited, the priority was set to medium even for data that were otherwise seen to be highly relevant or of adequate quality. Ranking datasets in terms of overall priority sets the detailed roadmap for extending and enriching the current IGAE dataset and the risk model described in Chapter 2. The next sections discuss each of these datasets in detail, along with some fraud risk indicators, which can be calculated when data are matched.

Table 3.1. Short description of additional datasets

Dataset name	Dataset group	Unit of measurement	Number of observations	ID to match on to IGAE main dataset	Priority for the IGAE's follow-up work
NationalCompany Register (Registradores de Espana)	i, ii	Organisation	>5 000 000	NIF of beneficiaries, names of organisations	high
Beneficial ownership registry (*LibreBOR*)	i, ii	Organisation	>5 000 000	NIF of beneficiaries	high
Database of Spanish senior positions and secretariats(*FICESA*)	ii	Institutions and State Bodies	~100 000	Name of organisations	high
CINCO.net	iii	Organisations	should be accessed by official body	NIF of organisations	high
Public procurement data	iv	Tender	1 391 558	NIF of organisations	high
Public Bankruptcy Registry (*El Registro Público Concursal*)	iii	Organisations	website does not allow to search	NIF of organisations	medium

Dataset name	Dataset group	Unit of measurement	Number of observations	ID to match on to IGAE main dataset	Priority for the IGAE's follow-up work
Spanish Association of Foundations (La Asociación Española de Fundaciones, AEF)	iv	Foundation	15 840	Location and type of beneficiary	medium
State Tax Administration Agency (*Agencia Estatal de Administración Tributaria*, AEAT)	iii	Organisations	not in public access	NIF of organisations	medium
European Union Aid	iv	Grant or contract	40 567	Name of beneficiary, vat number	medium
National Register of Associations of the Ministry of Interior (*el Registro Nacional de Asociaciones)*	i, iii	Accredited NGO	44	CIF of organisation	low
Loyalty Foundation (Fundación Lealtad)	i, ii, iii	Accredited NGO	191	Name of organisation	low

Source: Author

Matching organisational data: More precise organisational profiles and anomaly detection

Organisational data for the parties involved in grant making include the grantors, grantees and third parties (i.e. project implementers). Matching data on organisations allows for gaining a more complete and detailed picture of organisational controls of fraud risks. It helps to identify additional organisational characteristics that might influence the probability of sanctions. For example, accounting information, size of the company and associated companies can all be useful characteristics for identifying fraud risks and improving the IGAE's risk model in the future. This group includes the following databases: the National Company Registry (*Registradores de Espana*), data from the Spanish Association of Foundations (*la Asociación Española de Fundaciones*, AEF), and the National Register of Associations (*el Registro Nacional de Asociaciones*) of the Ministry of Interior.

Company registry and financial data

One of the most relevant datasets for the IGAE's purpose and for enhancing the risk model is the National Company Register. It contains data on companies' details, capital, representatives (e.g. directors and attorneys), registered acts and filing of annual accounts (i.e. financial performance). The list of variables are presented in Table 3.2.[2]

Table 3.2. List of variables (National Company Register)

Variables	Description	Type of the variable
Name	The name of the company	Text
NIF	The NIF number of the company	Text
Date of incorporation	The date the company was incorporated	Date
Company address	The address where the company is registered	Text
Sector of economic activity	In which economic sector the company operates (NACE)	Categorical
Legal form	Official legal form of the company (national forms)	Categorical
Company status	If the company is active and operational	Categorical
Company's assets	Total value of items benefiting the company economically	Numeric
Company's liabilities	Total value of the company's obligations	Numeric
Company's income	Total amount of income generated annually	Numeric
Company's expenditures	Total amount of expenditures per year	Numeric
Changes in equity	If there were any changes in equity for the past year	Binary + text
Cash flows	Increase or decrease in the amount of money	List
Members	Includes the name of all members of the current company representation	Text
Beneficial owners	List of names of final owners of the company	Text

Source: https://sede.registradores.org/site/home

The National Company Register can be matched to the main BDNS dataset by the company's NIF number, or if that is erroneous, by the name of the organisation. Almost all data fields contained in the company dataset are relevant for the IGAE in terms of enhancing its risk model. These fields range from essential registry information, such as date of incorporation or location of headquarters, to balance sheets and income statements. Similarly, recent changes in equity and the full list of members of the company can provide additional insights on potential conflicts of interest when matched with other datasets.

With regards to essential registry information, some red flags have proven to be useful for predicting corruption and fraud risks. For example, companies which have been set up, or whose registration data has been modified shortly before applying for a grant, are higher risks. Similarly, companies registered in so-called "company graveyard" addresses can be high risk, where a very large number of companies are registered with high degrees of fluctuation (e.g. thousands of companies created and closed on the same address each month). Similarly, as discussed in Chapter 2, the type of organisation (i.e. the company legal status), as well as its overall income and size, can influence the level of fraud risks. For example, due to legislation, certain types of organisations can be less transparent or more loosely regulated (e.g. trusts or company ownership presented by bearer shares).

Regarding company financial data, the IGAE could consider a number of relevant indicators for risk prediction. First, the ratio between a company's expenditures and income can provide information as to whether the company is profitable. Companies that are not profitable are riskier beneficiaries of grants and subsidies, since they may use funds to repay their debts as opposed to financing their projects. Similarly, a negative ratio between a company's liabilities and assets suggests greater risk in terms of the appropriate use of grants. Frequent changes in equity might be a signal of internal conflicts and instability within the company, increasing the level of risks associated with grants and subsidies for such organisations. Systematic decrease in cash flows reflects stagnation or reduction in the company's activities, which also brings its reliability into question. Combining the grants data with company financial data also can reveal the relative size of the grant compared to the company, with small companies receiving large grants potentially being risky.

Register of Associations

Another organisational dataset that the IGAE could consider for its risk model, although a low priority, is the National Register of Associations (*el Registro Nacional de Asociaciones*), held by the Ministry of Interior. This is a list of organisations that have passed a review made by the Spanish Agency for International Development Cooperation (*Agencia Española de Cooperación Internacional para el Desarrollo*, AECID) in which more than 70 qualitative and quantitative criteria were used, mostly related to experience, financial solvency, transparency and human resources. The main limitation of this dataset is the small number of accredited NGOs it provides, as it only has 44 observations. They are stored in HTML format and can be easily transformed to excel or any other data formats. The list of the variables are described in Table 3.3.

Table 3.3. List of variables (National Register of Associations of the Ministry of Interior)

Variables	Description	Type of the variable
Name	What is the name of the NGO	Text
Sectors	Which sectors it's qualified for	Categorical
CIF	What is the CIF number of the NGO	Text

Source: https://www.aecid.es/EN/aecid/our-partners/ngdo/accreditation

The dataset provides two potential IDs for matching—the name of the organisation and its Customer Identification Number (CIF). Both can be used to link the data to the IGAE's grant data. The data consists of three variables, two of which are IDs and one specifies the exact sectors in which the NGO is qualified to operate. Based on this information, two binary variables can be created: 1) whether the NGO has been reviewed, and 2) whether the NGO is acting in the same area as it was qualified for (e.g. the NGO was qualified for the health sector, but receives grants for the education sector). Due to a low number of observations, significant changes in predicted risk scores are unlikely. However, if the main BDNS dataset is filtered for NGOs only, this information might influence the outcomes for this sector.

NGO evaluations

The third dataset worth considering is that of the Loyalty Foundation (*Fundación Lealtad*). This is an independent NGO evaluator, which analyses the management, governance, use of funds, economic situation, volunteering and transparency of NGOs. On the foundation's website, there is a downloadable PDF file with the list of all positively evaluated NGOs. However, this list has limited information beyond name of organisations. Therefore, a more effective approach would be to access the HTML pages of each organisation and parse data manually. There is a possibility to parse information from standardised PDFs called "full reports" for each NGO. The list of variables are described in Table 3.4.

Table 3.4. List of variables (Loyalty Foundation)

Variables	Description	Type of variable
Name	The name of the NGO	Text
Sectors	Sectors of its operation	Categorical
CIF	The CIF number of the NGO	Text
Income	The annual income of the organisation + sources	Numeric + categorical
Expenses	The annual expenses of the organisation + types of expenses	Numeric + categorical
Year	Year of origin of organisation	Date
Beneficiaries	The overall number and type of beneficiaries of this NGO	Numeric
Partners	Number of partners the NGO has	Numeric
Employees	Number of employees the NGO has	Numeric
Volunteers	Number of volunteers the NGO has	Numeric
NIF	The NIF number of the organisation	Text
Management positions	Individual(s) who represent the management of this NGO	Text
Contacts	Email, telephone, address of organisation	Text
Geographic area	Where the NGO operates	Text

Source: https://www.fundacionlealtad.org/ong/a-toda-vela/

The main IDs by which organisations can be linked to the IGAE's datasets are name of organisation and NIF. While name is available in both HTML and PDF files, NIF is stored in the full report PDF. Data on income, expenses, sector of activities, year of origin, as well as number of beneficiaries, partners and employees can add to the background information for the analysis. As before, a binary variable can be created reflecting whether the given organisation is verified by the *Fundación Lealtad*. Besides the general background information, some additional indicators can be extracted from this dataset. For instance, the ratio of expenses should be taken into consideration to assess how much is spent on administration of the NGO in comparison to its mission. High spending on administration might be a signal for higher risk scores, although on its own would not be an indicator of fraud or wrongdoing. Administrative bodies when linked to other datasets (e.g. politically exposed persons) can provide information on potential conflicts of interest.

Matching personal data for tracking connections and conflict of interest

The second group of datasets that could enhance the IGAE's risk model, described in Chapter 2, is data on personal connections and conflicts of interest. Matching data on personal connections between the public and private sectors opens up the possibility for tracking conflicts of interest. Such data can be analysed with the use of network analysis to identify if there are connections between politically exposed persons and owners of the companies receiving grants and subsidies. Some potential sources were already discussed in the previous group. The next sections will focus on the Beneficial Ownership Registry and *FICESA*, the database of Spanish senior positions and secretariats.

Beneficial Ownership (BO) Registry

The BO registry provides information for over 5 000 000 organisations registered since 2009. The short list of variables is provided in Table 3.2. There is no complete dataset in the public domain, but the source—an online platform for consulting and analysing the Official Gazette of the Mercantile Registry (*Boletín Oficial del Registro Mercantil*) called *LibreBOR*—provides API and Python script to parse the data.[3] It is possible to select those organisations that appear in the IGAE datasets, without parsing the whole dataset, which will make for a more efficient processing time.

Table 3.5. List of variables in the BO registry

Variables	Description	Type of the variable
Current and previous denomination	The name of the company, what are the previous names	Text
Registered office	The official office is registered	Text
Legal form	The legal form of the company	Categorical
Province	Province where the company operates	Text
Management positions	Names of the individual(s) in management positions	Text
Date of dissolution and reason	If the company dismissed or disintegrated - when and why it happened	Date + text
Registry data	Additional information on company registry	Text
Links to the official sources	Official source from which the data comes	Text
Beneficial owners[1]	List of names of final owners of the company	Text

Source: https://docs.librebor.me/

There are two ways for the IGAE to match the BDNS datasets to the BO registry: 1) by name of the organisation, or 2) by NIF of the beneficiary. Alternatively, it is possible to aggregate data per province and match aggregate numbers (e.g. average company size) by particular location. The BO dataset contains a lot of background information for organisations, but the most relevant one is management positions, associated organisations, and the final beneficial owners. The ownership data is best used when matched against other datasets, in particular, lists of political office holders (see next section).

In addition, the IGAE can use some of the background information as risk predictors on their own. When the names of beneficial owners of grant recipients is matched against public office holders, it is possible to identify either direct conflicts of interest (i.e. when the official works for the granting body itself) or indirect forms of potential conflict (i.e. when the related political office holder works in a higher level or supervisory body to the granting organisation). When looking at the ownership data on its own, the information on companies associated with the grantee can reveal risks if further matched to other datasets (e.g. complex forms of conflicts of interest and related risk factors).[4]

Senior bureaucrats' database

The next source is a database of Spanish senior positions and secretariats called *FICESA*. This source contains data related to senior public officials in a wide range of public organisations: state secretariats, undersecretaries, general directorates and sub-directorates, budget offices, official offices, as well as different judicial bodies for state, regional and local levels. There is no data in the public domain, and data must be requested from the data holder by filling out a form. Therefore, the format of the data and the variables the dataset contains is unclear. There was no response to attempts to contact the source. It is assumed that the IGAE would be able to gain access to the full database as a bulk download.

The only ID by which this dataset can be linked is names and, if available, additional personal features, such as date of birth. If the BDNS dataset contains data on beneficial owners, as described above, the data on official positions can be linked by persons' names. Linking the IGAE's datasets to the information on senior office holders creates the possibility to conduct network analysis and see if there are conflicts of interests between private organisations receiving grants and public bodies giving grants. It is particularly useful to use the BO registry in order to find all the associated organisations, and analyse if they are connected to politically exposed persons. For instance, the organisation receiving the grant is not connected to anyone from official bodies, but one of its related organisations could be.

Matching data on organisational reliability and violations to collate risks across different domains

Datasets with information about organisational reliability and violations of rules or laws is the third group of data that could support the IGAE to strengthen its risk model for assessing grant fraud risks. This group was covered partially above in the section about data on accredited NGOs. In addition, in this group, there are datasets on bankruptcy and taxation. Matching data on organisational reliability and violation of rules illuminates new dimensions of fraud risks relating to other domains. These datasets can help predict fraud risks in grants by exploiting correlations between accredited organisations' trustworthiness, rule following behaviours (tax debts, bankruptcy, etc.) and fraud in grants. Building on previous discussions, the next section focus on the Public Bankruptcy Registry, AEAT's tax data and accounting data from *CINCOnet*.

Bankruptcy Registry

The first dataset in this group, identified previously as a medium priority for the IGAE, is the Public Bankruptcy Registry (*El Registro Público Concursal*). The source includes information on procedural resolutions, bankruptcy and out-of-court settlements. The data can be parsed from HTML after filtering by province or court. Unfortunately, for unknown reasons, filtering does not work on the site properly, leading to page errors. Yet, the approximate list of variables is presented in Table 3.6.

Table 3.6. List of variables (Public Bankruptcy Registry)

Variables	Description	Type of variable
Name	The name of the company	Text
Identifying document	The ID of the bankruptcy document	Text
Debtor	If the company is a debt or not	Binary
Disabled	If the company is disabled or not	Binary
Administrator	If the company is an administrator of the bankruptcy or not	Binary

Source: https://www.publicidadconcursal.es/concursal-web/afectado/buscar

This dataset can be matched to the IGAE's grants data by either name of the organisation, or NIF/CIF number. The source does not provide an opportunity to look through all the cases, requiring filtering beforehand, so the easiest way to set a filter is to use province. The most relevant information for fraud risk assessments are the details on bankruptcy. The source provides location, name of organisation, court, judge and NIF/CIF or other identifiers of organisations. Unfortunately, there is no information on the date of bankruptcy proceedings, which would be especially important to analyse past grants and subsidies. After matching, the most relevant risk indicator for the IGAE would be the binary variable ('flag') reflecting if the grantee was or is currently in the state of bankruptcy. Such bankruptcy information on an organisation might signal that the awarded grant or subsidy will be misused by the beneficiary, or at the very least, inadequately administered due to other organisational pressures.

Tax data

The second dataset on rule violations is data from the State Tax Administration Agency (*Agencia Estatal de Administración Tributaria,* AEAT). This is a dataset with restricted access and only aggregated statistics are available in the public domain. Once again, for the discussion below, an assumption was made that the IGAE can obtain full access to the database in order to incorporate such data into its risk model. According to the notes the AEAT published, it has data in a disaggregated format which can be provided upon request. Aggregated data covers filing of tax returns, payment of taxes, debts and fees, tax certificates, consult tax return, etc.

Due to restricted access to the datasets it is uncertain whether the IDs are the same as in the BDNS dataset, but most likely organisations can be matched either by name or by NIF of the beneficiary. Information on timely payment of taxes, debts and fees are the most relevant for enriching predictive models on fraud risks. Late payment of taxes, as well as presence of debt in a given organisation (or associated ones) could be a signal of higher risks.

Accounting information

The third dataset belonging to this group is accounting and budgeting data from *CINCO.net*, deemed a high priority for the IGAE and improvements to risk model. The data includes expense operations and total expenditure amount in the current year, revenue amount in the current year, cash flows, non-budgetary operations, third-party expenses, general data of third parties, etc. Like the AEAT's data, this data is not in the public domain; however, the Ministry of Finance and Civil Service (*Ministerio de Hacienda y Función Pública*) manages *CINCO.net* and the IGAE has direct access to it.

The organisations in this database can be matched by names or NIF of the beneficiary to the BDNS. Yet, due to restricted access of the data, it is difficult to assess the quality and content of matching variables. Besides general background information on revenues and expenditures, *CINCO.net* provides data on reimbursement of other grants provided by different organisations in Spain. This can be particularly useful in assessments of potential risks in future subsidies and grants provision, such as double-funding of operations or the large value of grants received compared the revenue.

Matching data on public contracts and other grants enables tracing double funding and related risks

The final group of datasets encompasses a diverse group of data on public contracts and other grants and funding. Matching data on other funds and contracts would allow the IGAE to cross-reference spending as well as develop additional risk dimensions. For example, it can help identify cross-subsidisation for the same activities, which should be considered a risk factor. Public procurement contracts received by a company can be scored using corruption risk indicators and then related to grants risks. For example, a company or agency (third party, grantor, grantee) participating in high-risk tenders might also be risky when it comes to grants. This group includes datasets from the Spanish Association of Foundations (*la Asociación Española de Fundaciones,* AEF), European Union Funds, and public procurement data.

Data for foundations

AEF's data provides information on foundations giving grants, including their types of activity, geographical areas, type of beneficiaries, date of constitution and origin of their administrative bodies. The list of the variables is presented in Table 3.7. The data is open access and can be easily downloaded in excel or PDF format. In total there are 15 840 foundations covered by the directory.

Table 3.7. List of variables of the Spanish Association of Foundations (AEF)

Variables	Description	Type of variable
Name	What is the name of the foundation	Text
Protectorate	Under which ministry/agency protectorate this foundation is	Text
Year	Year of constitution	Date
Contacts	What are the contact details of the foundation (email, phone)	Text
Address	Where the foundation operates	Text

Source: http://www.fundaciones.es/es/buscador-fundaciones

Matching this dataset to the BDNS requires several steps. First, all the observations should be filtered by type of beneficiary, using the online filtering, since the type of beneficiary is not a data field in the downloadable file. Second, the particular location should be matched to the locations of grantors or grantees. This will not provide the exact information as to whether the beneficiary received another grant from a certain foundation, but it indicates the presence of the foundation in the same location with the same types of beneficiaries.

The most relevant information for the IGAE to assess risks would be whether any of the beneficiaries were double granted for the same activities. To precisely track such risks requires checking the exact beneficiaries by their IDs, yet this source does not provide such detailed information. Hence, only aggregate information, which is much more imprecise, can be used from this source. The presence of a foundation supporting similar activities in the same locality (province) as grantor or grantee increases the probability of being double funded.

European Union (EU) Funds data

The next relevant dataset for the IGAE to consider matching to the BDNS data, as a medium priority, is data for EU Funds. The Spanish government and the European Commission provide the data, and they cover records from 2007 to 2020. The data are easily accessible and can be downloaded in Excel format. The list of relevant variables is presented in Table 3.8.

Table 3.8. List of variables (European Union aid)

Variables	Description	Type of variable
Budget references	The budget reference ID for this grant	Text
Subject of grant or contract	The purpose/subject of this grant	Text
Name of beneficiary	The name of beneficiary	Text
VAT number	The VAT number of beneficiary	Text
Contracted amount	The amount of money was contracted to beneficiary	Numeric
Number of budgetary commitments	The number of budgetary commitments the beneficiary has	Numeric
Programme name	The name of the programme under which the grant was allocated	Text
Responsible department	The department responsible for grant allocation	Text
Project start and end date	The start and end date of the project	Date

Source: https://ec.europa.eu/budget/financial-transparency-system/analysis.html

The data provides a VAT number as an ID for organisations, which can be transformed into a NIF number by removing the first two letters. Alternatively, names of organisations can be used for matching. Number of budgetary commitments, subject of grants or contracts, as well as project start and end dates are particularly relevant to identify whether the grantee received funding from the EU for the same project as its Spanish grant. Double funding is a fraudulent practice when the same project is funded more than one time by different donors, without providing information on contributions made. Therefore the project might be implemented, yet the extra public money disbursed is not used as intended.

Public procurement data

The last data source the IGAE could consider matching with its datasets is national public procurement data. The opentender.eu portal contains this data collected from two official government sources (*Ministerio de Hacienda y Función Pública* and *Plataforma de Contratación*), as well as Tenders Electronic Daily (TED), a European online public procurement portal. The data contains all the publicly available information on tenders, contracts, bidders, buyers and suppliers necessary for calculating the Corruption Risk Indicator (see Box 3.1). The list of relevant variables is presented in Table 3.9.

Table 3.9. List of variables (Public procurement data)

Variables	Description	Type of variable
Supplier ID	Unique ID of supplier	Text
Buyer ID	Unique ID of buyer	Text
Name of supplier	Name of supplier winning the contract	Text
Name of buyer	Name of buyer providing tender call	Text
Number of bids	How many bids were made per tender	Numeric
Procedure type	Is the procedure type open or restricted	Categorical
Public call	Was the call for tender available to public	Categorical
Length of bid submission	The length between start and end date of bid submission	Numeric
Length of decision period	The length between end date of bid submission and decision	Numeric
Connections	Are there recorded connections between supplier and procurement authority	Categorical

Source: Platforma de Contratacion https://contrataciondelestado.es/; Portal Institucional Del Ministerio De Hacienda y Funcion Pública: https://www.hacienda.gob.es; Tenders electronic daily: http://ted.europa.eu.

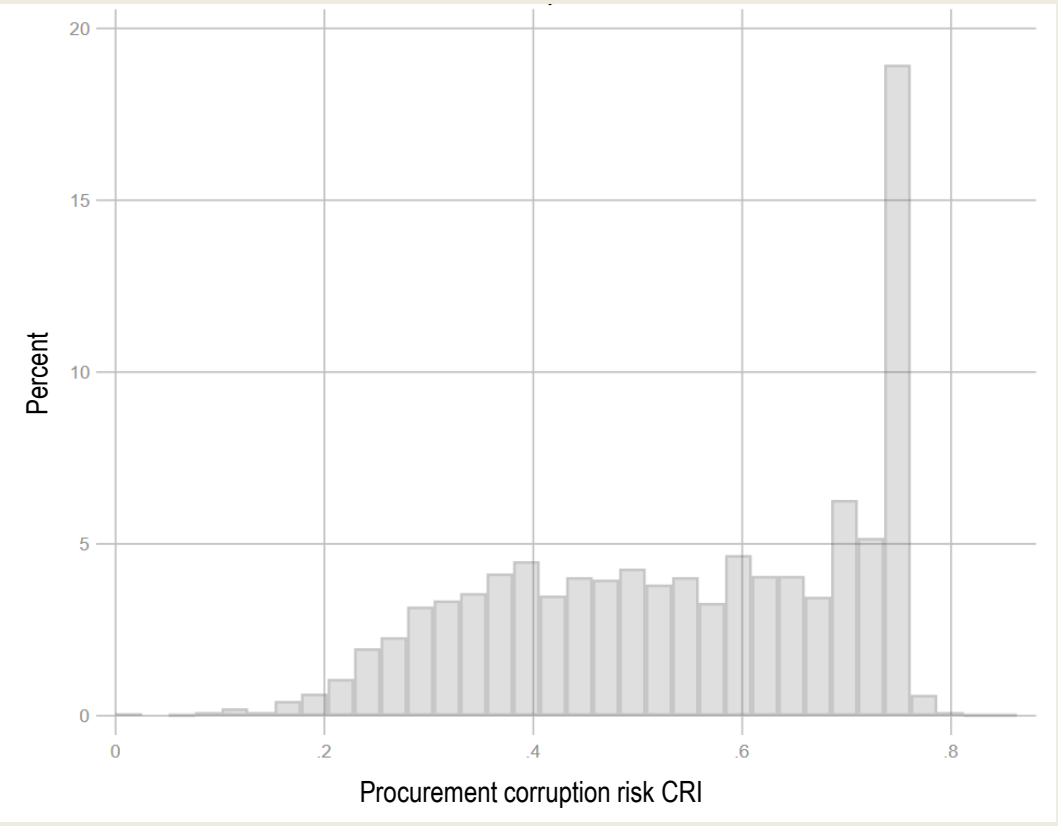
Suppliers IDs are the same as the grantees' NIFs, therefore this ID can be used for linking data. Alternatively, names of organisations as well as grantors names can be matched to the buyers or suppliers from procurement dataset. To assess if the procurement contracts won by bidding firms or tenders run by public sector grantors are prone to corruption, information on corruption proxies can be used. For example, single bidding on competitive markets, procedure type used, publication of the call for tenders, length of bid advertisement and decision period, as well as connections between supplier and procurement authority. Collating public procurement corruption risks in the procurement activities of grantees or grantors can shed additional light on grants fraud risks as it is expected that organisations that are risky in one domain will also be risky in a related domain. This logic of analysis is empirically demonstrated in Box 3.1.

Box 3.1. Matching IGAE Grants data with Public procurement data (opentender.eu dataset)

The Corruption Risk Indicator (CRI) proxies for the deliberate restriction of competition in public procurement tenders for the benefit of a connected bidding firm. The CRI methodology utilises administrative data to calculate corruption risk scores for each contract. Based on the methodology developed by (Fazekas and Kocsis, 2017[1]), the criterion for the selection of procurement risk indicators is the degree of association with unjustified restriction of competition, that is single bidding on competitive markets. It includes several corruption proxies in addition to single bidding such as procurement closed procedure type risk, lack of publicity of call for tenders, supplier tax haven registration, procurement authority dependence on supplier (i.e. agency capture), and the length of bid advertisement and decision periods.

The suppliers tax identification ID (NIF) was used to match the grants dataset to the cleaned public procurement dataset . After cleaning the NIF identification number from nonsensical entries, the grant fraud risk scores were aggregated for each supplier and matched directly to the procurement dataset. There were 103 872 contracts located by 6 408 suppliers that have received a grant. Figure 3.1 shows the aggregated CRI distribution for granted suppliers excluding suppliers with less than 3 contracts. There is an average CRI score of 0.55, considerably higher than the national average.

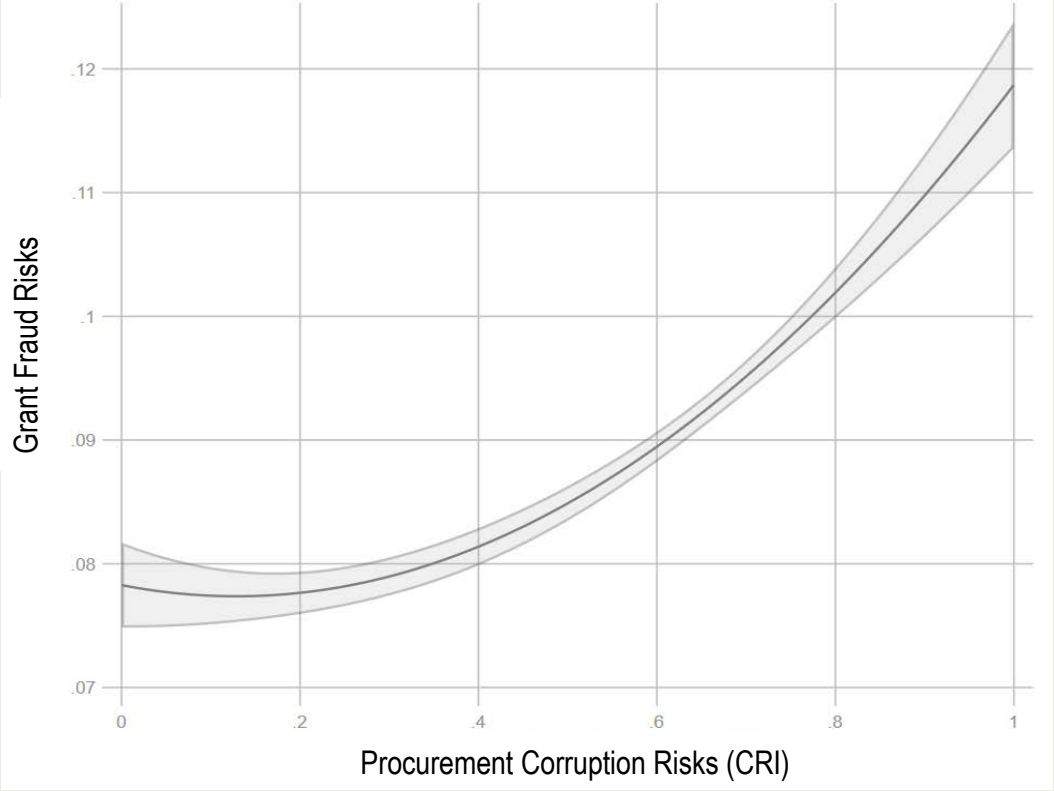
Figure 3.1. CRI distribution (Suppliers)



Source: Author

Matching the grant dataset to the public procurement dataset allows for deeper insights into the relationships between the risk scores. It is possible to run linear and non-linear regression analyses, including controls for buyer location, buyer type, type of market (CPV sectors), contract type and tender year. Both models in Table 3.10 show a positive correlation between the procurement corruption risk scores and the grant fraud risks. However, model 2 seems to fit better in capturing the non-linearity of this relationship. In Figure 3.2 we show the predictive margins from modelling the CRI in a quadratic relationship with the Grant Fraud Risk. These simple regression results assure us of the validity of both risk scores as they are aligned and convey a similar message, that higher corruption risk scores positively correlate with higher grant fraud risks. Moreover, the association is especially strong when public procurement corruption risks are above the sample average.

Figure 3.2. Correlation between CRI and Grant Fraud Risks (Predictive Margins)



Source: Author

Table 3.10. Correlation between CRI and Grant Fraud Risk

Dependent variable	Grant Fraud Risk	
Model	(1)	(2)
Sample	Granted	Granted
CRI	0.036***	-0.014
	(0.002)	(0.021)
CRI^2		0.054**
		(0.024)
Controls	✓	✓
Observations	103 151	103 151
R^2	0.1719	0.1721

Notes: Regression includes controls for contract values, contract type, buyer type, buyer location, market, contract type and tender year.
Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Source: Fazekas, M. and G. Kocsis (2017[1]), "Uncovering High-Level Corruption: Cross-National Objective Corruption Risk Indicators Using Public Procurement Data", *British Journal of Political Science*, Vol. 50/1, pp. 155-164, http://dx.doi.org/10.1017/s0007123417000461.

Benefits of drawing on multiple datasets

This chapter offered a detailed account of how and why different datasets can be linked to existing IGAE datasets with particular attention to promising fraud risk indicators enabled by the new data. These new indicators principally capture actor behaviour rather than simple background characteristics allowing for a far more precise risk assessment. However, data linking not only allows for calculating new indicators in one database and linking them to each other, but also for creating new indicators by drawing on multiple datasets. Such complex indicators offer additional insights on relevant risk dimensions. They also represent a more robust measure of actor behaviour, because multiple sources pointing at the same behaviour carry greater validity than a single dataset.

Drawing on multiple datasets is crucial for comprehensively mapping complex fraud behaviours, as well as for reducing the rate of false positives that are common in simple models (Fazekas, M., Ugale, G, & Zhao, A., 2019[2]). Combining multiple indicators stemming from different datasets is considered as good practice in risk measurement as it allows for measurement triangulation. In other words, it allows for increasing convergent validity. False positives are pervasive in simple risk assessments, as many indicators merely point at potential wrongdoing rather than actual bad deeds. Moreover, widely used indicators of conflicts of interest typically indicate the presence of a potential conflict rather than an actual conflict that represents abuse of a situation for undue personal gain. However, when conflicts of interest information is combined with data on outcomes, such as double-counting grants or anomalous financial performance, the combination of indicators provide greater validity to the measurement approach.

Matching datasets representing multiple dimensions of relationships can also power the use of advanced, multi-layer network analytics. Such multi-layered relationships can encompass connections between private companies and public grant making organisations through a range of contractual relationships, or links between companies' beneficial owners and politically exposed persons working in public sector bodies. Multiple network connections established through the use of large-scale, linked administrative datasets also allow for tracking temporal changes in connections across potentially risky entities and individuals, thereby increasing the analytical sophistication of risk modelling.

Conclusion

This section has reviewed a wide variety of potential useful additional datasets to the existing IGAE dataset. By doing so it set out a roadmap of data capture and matching maximizing analytical value for IGAE. Of the reviewed datasets, company information on registration, ownership and financials represents the highest potential for further refining the fraud risk assessment model. These datasets can be readily matched to IGAE's internal data using company registry IDs. Moreover, matching public procurement data to grants data, also demonstrated by analysing readily available datasets, can add great value as 2 sets of risk factors can be triangulated against each other producing more reliable risk assessment. Once these high priority datasets are brought into the IGAE data pipeline, further datasets can also be considered such as the bankruptcy register.

References

Fazekas, M., Ugale, G, & Zhao, A. (2019), *Analytics or Integrity: Data-Driven Decisions for Enhancing Corruption and Fraud Risk Assessments*, OECD Publishing, Paris, https://www.oecd.org/gov/ethics/analytics-for-integrity.pdf. [2]

Fazekas, M. and G. Kocsis (2017), "Uncovering High-Level Corruption: Cross-National Objective Corruption Risk Indicators Using Public Procurement Data", *British Journal of Political Science*, Vol. 50/1, pp. 155-164, http://dx.doi.org/10.1017/s0007123417000461. [1]

Notes

[1] In some cases, certain information is presumed to be present in the IGAE's datasets; however, confirmation of this was not possible because of anonymisation of most of the databases.

[2] The access to the dataset is restricted and requires paying a fee for each organisation and receiving a digital certificate. Free access is only allowed to the aggregated data per sector, year or business sector. The only company-level information available without additional restrictions is company status (i.e. operational or not). For the IGAE to use this data, it would need to gain full access to the complete and current dataset, either through paying the bulk access fee or setting up a special arrangement with the government data provider. Easy access, public alternatives also exist, for example, opencorporates.com, which is a private, social enterprise aiming to make all company data easily accessible around the world.

[3] See https://docs.librebor.me/python/.

[4] Due to a restricted access to the source, it is not clear if the information on beneficial owners is there. Yet, it is present in the company register; therefore, it is reasonable to expect that it also contains a variable in *LibreBOR*. In case it is not, the information can be obtained from the company register after receiving an electronic certificate.

Annex A. Descriptive statistics of variables in the cleaned dataset

The table shows the minimum, mean, median and maximum for numeric variables and the number of cases by the most frequent categories for categorical variables.

ABIERTO_CON420 S : 45701 N : 1004769	AUDAESTADO_CON490 N : 728696 S : 321774	FINALIDAD_CON540 12 : 408525 6 : 120412 16 : 84776 5 : 83064 11 : 81101 18 : 60650 (Other) : 211942	NOMINATIVA_CON610 N : 659757 S : 390713
PUBLICABLE_CON620 0 : 18819 1 : 1031651	IMPACTOGENERO_CON630 1 : 54816 2 : 836210 3 : 159301 4 : 143	FECHA_ACTUALIZACION Min. : 2000-03-31 Mean : 2019-04-12 Median : 2020-03-05 Max. : 2021-06-02	PAIS_TER100 ES :1048239 AR : 259 VE : 144 FR : 141 IE : 114 DE : 104 (Other) : 1469
ID_TER110 Length:1050470 Class :character	PAISDOM_TER250 ES : 1048238 AR : 261 VE : 146 FR : 137 IE : 114 DE : 104 (Other) : 1470	NATURALEZA_TER280 SOCIEDADES DE RESPONSABILIDA LIMITADA : 293159 ASOCIACIONES : 186563 CORPORACIONES LOCALES : 166903 SOCIEDADES CIVILES : 88221 (Other) : 315624	TIPOBEN_TER290 FSA : 5300 GRA : 6395 JSA : 512945 PFA : 435385
REGION_TER310 ES :179660 ES41 : 83013 ES51 : 65044 ES43 : 55058 ES511 : 44993 ES425 : 39884 (Other):582818	ID_COS Length :1050470 Class : character	DAT_COS_CSU210 Min. : 2018-01-01 Max. : 2020-12-31 Median : 2018-12-28 Mean : 2019-06-26	COSTE_ACT_CSU240 Min. : 0.000e+00 Max. : 3.120e+10 Median : 3.000e+03 Mean : 6.343e+04
IMPORTE_CONCE_CSU220 Min. : 0 Max. : 139112532 Median : 2296 Mean : 23886	AYUDA_EQUI_CSU250 Min. : 0 Max. : 139112532 Median : 2296 Mean : 23886	REGION_CSU260 ES : 179702 ES41 : 83266 ES51 : 69608 ES43 : 65568 ES511 : 45443 ES425 : 39871 (Other) : 567012	Year Min. : 2018 Median : 2018 Max. : 2020
FECHAPAGO_PAG210 Min. : 2017-07-25 Median : 2019-07-11	IMPORTE_PAG220 Min. : 0 Median : 2000	RETENCION_PAG230 N : 1048245 S : 2225	CON550 A : 179727 S : 77071

Mean : 2019-10-03 Max. : 2021-06-01	Mean : 15897 Max. : 105745000		J : 67460 (Other) : 688175
CON560 OTROS, SUBV : 200 PREST, SUBV : 1982 SUBV : 1045030 SUBV, OTROS : 109 SUBV, PREST : 3058 SUBV, VENTA : 91	CON580 PFA : 501971 JSA : 398807 FSA : 16830 PFA, GRA : 15287 GRA, PFA : 13359 JSA, PFA : 12484 (Other) : 91732	CON570 ES : 204438 ES42 : 135555 ES41 : 90127 ES51 : 67554 ES43 : 65829 ES61 : 39539 (Other) : 447428	SAN_dum 0 : 1049439 1 : 1031
Month_CSU210 12 :215890 9 :156651 6 :112147 10 :109546 11 : 98538 7 : 87721 (Other):269977	Nawards_TER_110 Min. : 1.00 Median : 9.00 Mean : 17.88 Max. :193.00	Amount_awards_TER_110 Min.: 1 Median : 95929 Mean: 85060 Max.:139915	NATIONAL_CSU260 0:870533 1:179937
REGIONAL_CSU260 0:1013951 1: 36519	MUNICIPAL_CSU260 0:216456 1:834014	NATIONAL_TER310 0:870551 1:179919	REGIONAL_TER310 0:1014840 1: 35630
MUNICIPAL_TER310 0:209850 1:840620	LOCAL_IMPL 0:100402 1:950068	SECTOR_CON550_AGR 0:589338 1:461132	SECTOR_CON550_MINING 0:995418 1: 55052
SECTOR_CON550_MANUF 0:968989 1: 81481	SECTOR_CON550_ELECTR 0:995312 1: 55158	SECTOR_CON550_WATER 0:991202 1: 59268	SECTOR_CON550_CONSTR 0:975388 1: 75082
SECTOR_CON550_RETAIL 0:971928 1: 78542	SECTOR_CON550_TRANSP 0:982215 1: 68255	SECTOR_CON550_ACCOM 0:965384 1: 85086	SECTOR_CON550_INFO 0:914043 1:136427
SECTOR_CON550_FIN 0:992458 1: 58012	SECTOR_CON550_RESTAT 0:996060 1: 54410	SECTOR_CON550_SCI 0:963038 1: 87432	SECTOR_CON550_ADMIN 0:946261 1:104209
SECTOR_CON550_SECUR 0:972335 1: 78135	SECTOR_CON550_EDUC 0:943212 1:107258	SECTOR_CON550_HEALTH 0:925630 1:124840	SECTOR_CON550_ART 0:887795 1:162675
SECTOR_CON550_OTHER 0:866250 1:184220	SECTOR_CON550_HOUSEHOL D 0:1029872 1: 20598	SECTOR_CON550_EXTRATER 0:1018305 1: 32165	

Annex B. Full list of variables in the uncleaned dataset

Variable	Short description	Additional variable description, if needed	Type
ADMINISTRACION_ANTE	Administration unit	What is the administration unit is providing the call	character
DEPARTAMENTO_ANT	Department	Which department is providing the call	character
ORGANO_ANTE	Body	Which official body is providing the call	character
ADMINISTRACION	Administration unit	Which administration unit is providing the call	character
DEPARTAMENTO	Department	Which department is providing the call	character
ORGANO	Body	Which official body is providing the call	character
DIR3_CON710	Granting body	Identification of one or more competent bodies to resolve concessions of the call.	character
DIR3_ANTE_CON705	Organising body	Identification of the body opening the call	character
CON100	Call ID	Unique identifier of each call, automatically assigned by the computer system when registering the call in the BDNS	character
REF_EXTERNA	Call manager	Who is the manager of the call	character
DESC_CONV	Description of the call in Spanish	What is the description of call in Spanish	character
BASEDESC_CON250	Description of the regulatory bases	Text of the title of the regulation that contains the regulatory bases for managing the call	character
BASEURL_CON260	Regulation URL	Link to the website that contains the full text in Spanish of the regulatory bases	character
ABIERTO_CON420	Open admission period	Indicates if the call keeps the application admission period open permanently	factor
INISOLICITUD_CON440	Application period start date	Starting date of the period enabled to admit requests	date
FINSOLICITUD_CON460	End date of the application period	End date of the period enabled to admit requests	date
AUDAESTADO_CON490	Condition of State Aid	Indicates if the aid of the call should be classified as ADE	factor
TIPOAYUDA_CON495	Type of ADE authorisation	Aid authorisation mechanism	factor
REGLAMENTO_CON502_50	Regulation of exemption by category of aid + Regulation of exemption by amount	EU Regulation exemption from the obligation of prior notification to the Commission by category of aid	factor
REFERENCIA_CON515	EU aid reference	Reference assigned by the EU as the aid identifier	character
FINALIDAD_CON540	Purpose	Public utility or social interest or promotion of a public purpose pursued with the granting of the subsidy	factor
FINJUSTIFICACION_CON600	Justification final date	Absolute end date of the deadline for submitting justifications for any concession	date
NOMINATIVA_CON610	Nominative grant	Nominative grant condition	factor
PUBLICABLE_CON620	Publication	Condition of publicity of the concessions	factor
IMPACTOGENERO_CON630	Gender impact	It rates the expected results in relation to the elimination of inequalities between women and men and the fulfilment of the equality policy objectives	factor

Variable	Short description	Additional variable description, if needed	Type
FECHA_ACTUALIZACION	Time at which a third party was updated in the database	What is the time at which a third party was updated in the database	date
PAIS_TER100	Third party country	Country that generates the identification of the third party	character
ID_TER110	Third party ID	Third party identifier	character
NOMBRE_TER210_240.x	Lastname + Business name	According to the information provided to the body obligated by the third party	character
PAISDOM_TER250	Country of third party	What is the country of the third party	factor
DOMICILIO_TER252	Address of the third party	What is the address of the third party	character
CODPOSTAL_TER254	Postal code of the third party	What is the postal code of the third party	character
PROVINCIA_TER258	Province of the third party	What is the province of the third party	character
MUNICIPIO_TER256	Municipality of the third party	What is the municipality of the third party	character
NATURALEZA_TER280	Legal nature of the third party	What is the legal nature of the third party	factor
TIPOBEN_TER290	Third party type	Cataloguing of third parties based on their legal nature and economic activity	factor
REGION_TER310	Region	Region in which the third party is established	character
ID_COS	Award ID	What is the unique identifier of award	character
TIPO_CONC_CSU204	Instrument	One of the legal or economic figures on the basis of which the grants and aid are awarded	factor
DISCRIMINADOR_CSU130	Discriminator	Grant award discriminator	character
OBJETIVOS_CSU205	Objective	Objective of the Aid Category Exemption Regulation	factor
DAT_COS_CSU210	Grant award date	Date of the resolution of the grant award	date
COSTE_ACT_CSU240	Costs	Amount of the fundable budget of the activity to which the grant award applies	numeric
IMPORTE_CONCE_CSU220	Amount awarded (grant)	Total amount committed in the grant award	numeric
AYUDA_EQUI_CSU250	Equivalent aid (grant)	Equivalent aid from grant award	numeric
REGION_CSU260	Region	Geographical location of the material application of the grant award	character
Year	Year of the award	Which year the award took place	date
CSU110	Announcement	Call identification	character
PAIS_CSU120	Beneficiary country	What is the country of beneficiary	factor
ID_CSU120	Beneficiary	Identification of the beneficiary	character
CSU130	Discriminator	Own reference of the granting body, free content, used to discriminate each grant award to the same beneficiary in the same call	character
DISCRIMINADOR_PAG110	Payment discriminator	Own reference of the granting body, free content, used to discriminate each payment of the same concession	character
FECHAPAGO_PAG210	Payment date (grant)	When the grant was paid	date
IMPORTE_PAG220	Amount paid (grant)	What is the amount of award	numeric
RETENCION_PAG230	Retention	Condition of tax withholding carried out	factor
PROYECTO_PRO130	Project ID	Project identification	character
DESCRIPCION_PRO210	Description	Project description	character
IMPORTEPROY220	Grant amount	Grant award amount allocated to the project	numeric
IMPORTEPROY230	Loan amount	Amount of the loan granted to the project	numeric

Variable	Short description	Additional variable description, if needed	Type
COSTE_PRO240	Project costs	Fundable cost of the project	numeric
AYUDA_PRO250	Equivalent aid (project)	Support equivalent to the project	numeric
REGION_PROY260	Region	Geographical location of the project	character
ANIO_EJE130	Year	Project execution year	date
EJE210	Grant amount	Amount of the grant award assigned to the executing agency in the year.	numeric
EJE220	Loan amount	Amount of the loan concession assigned to the executing agency in the year.	numeric
COSTE_EJE240	Costs	Aid equivalent to the project executor in the year	numeric
AYUDA_EJE250	Equivalent help (executor)	What is the amount of the help provided to executor	numeric
IDENTIFICADOR_EJE120	Executor ID	Identification of the executor	character
DISCRIMINADOR_REI110_1...8	Reimbursement discriminator	Own reference of the granting body, free content, used to discriminate each reimbursement procedure of the same beneficiary derived from the same grant	character
FECHADEV_REI210_1...8	Refund resolution date	Date of the resolution of the proceeding refund procedure	date
PRINCIPAL_REI230_1...8	Principal	Amount of refund	numeric
CAUSA_REI220_1...7	Causes	One or more of the causes that support the origin of the refund	factor
DISCRIMINADOR_DEV110	Return discriminator	Own reference of the granting body, free content, used to discriminate each voluntary return of one or more payments of the same concession	character
FECHADEV_DEV210	Return date	Date of the administrative resolution of acceptance of the return.	date
PRINCIPAL_DEV220	Amount of the principal of the return	Amount of the principal that the beneficiary returns without any resolution of repayment.	numeric
INTEERESES_DEV230	Amount of interest on the return	Amount of default interest calculated	numeric
CON550	Economic activities	One or more of the sectors of the economy foreseen in the call	factor
STRDESCRIPCION.x	Description of CON550	Description of variable CON550	character
DIARIO_CON310	Official journal of publication	Reference to the Official Gazette to which the extract of the call must be sent for publication	character
DESCRIPCION_CON335	Title in Spanish of the call	What is the Spanish title of the call	character
FECHA_CON351	Date of the signature of the call	When the call was signed	date
LOCALIDAD_CON352	Location of the signature footer of the call	Where the call was signed	character
PUBLI_CON390	Date of publication in the official gazette	When the call was published in the official gazette	date
URL_CON400	Reference in the official gazette of the extract in Spanish	What is the extract in official gazette in Spanish	character
CON560	Help instrument	One or more of the legal or economic figures on the basis of which the subsidies and aid are awarded	factor
STRDESCRIPCION.y	Description of CON560	The description of variable CON560	character
CON503	Objectives of the Aid Category Exemption Regulation	What are the objectives of the Aid Category Exemption Regulation	factor
STRDESCRIPCION_CON503	Description of CON503	Description of variable CON503	character
CON580	Types of beneficiary	One or more of the types of beneficiary foreseen in the call	factor
STRDESCRIPCION_CON580	Description of CON580	Description of variable CON580	character

Variable	Short description	Additional variable description, if needed	Type
CON570	Geographic regions	One or more geographical locations of the material application of the subsidy or aid provided for in the call	factor
STRDESCRIPCION_CON570	Description of CON570	Description of variable CON570	character
CON690	EU Fund financing amount	The amount of EU funds financial support	numeric
STRDESCRIPCION_CON690	Description of CON690	Description of variable CON560	character
STRVALOR	Type of EU financing institution	The EU financing institution	factor
DATSANC_1...3x	Date of sanction	When the sanction was imposed	date
STRDISCRIMINADOR_1...3x	Discriminator of the sanction	The organisation that is the discriminator of the sanction	character
MULTALEVE_SAN250_1...3x	Fine for minor infractions	The fine amount for minor infractions	numeric
MULTAGRAVE_SAN280_1...3x	Fine for serious infractions	The fine amount for serious infractions	numeric
MULTAMUYGRAVE_1...3x	Fine for very serious infractions	The fine amount for very serious infractions	numeric
PUBLICABLE_SAN440_1...3x	Condition of publicity of the sanction	Indicates whether the sanction should be public, according to art. 20.9 LGS	factor
LIMITE_SAN450_1...3x	Advertising deadline	What is the deadline for advertising	character
STRDESCRIPCION_SANC_1...3x	the description of STRVALOR value	Description of STRVALOR variable value	character
STRVALOR2_1...3x	Type of violation	Mild behaviours + Serious behaviours + Very serious behaviours	factor
ACTIVIDADES_TER320	Third party economic activities	What are the third party economic activity types	factor
DESC_ACTIVIDAD	Description of TER320	Description of values of TER320 variable	character

Annex C. List of variables used in the analysis

Variable	Short description	Additional variable description, if needed	Type
ABIERTO_CON420	Open admission period	Indicates if the call keeps the application admission period open permanently	factor
AUDAESTADO_CON490	Condition of State Aid	Indicates if the aid of the call should be classified as ADE	factor
FINALIDAD_CON540	Purpose	Public utility or social interest or promotion of a public purpose pursued with the granting of the subsidy	factor
NOMINATIVA_CON610	Nominative grant	Nominative grant condition	factor
PUBLICABLE_CON620	Publication	Condition of publicity of the concessions	factor
IMPACTOGENERO_CON630	Gender impact	It rates the expected results in relation to the elimination of inequalities between women and men and the fulfilment of the equality policy objectives	factor
PAIS_TER100	Third party country	Country that generates the identification of the third party	factor
PAISDOM_TER250	Country of domicile	Country where the third party is located	factor
NATURALEZA_TER280	Legal nature of the third party	The legal type of the party	factor
TIPOBEN_TER290	Third party type	Cataloguing of third parties based on their legal nature and economic activity	factor
COSTE_ACT_CSU240	Costs	Amount of the fundable budget of the activity to which the grant award applies	numeric
IMPORTE_PAG220	Amount paid (grant)		numeric
RETENCION_PAG230	Retention	Condition of tax withholding carried out	factor
CON560	Help instrument	One or more of the legal or economic figures on the basis of which the subsidies and aid are awarded	factor
CON580	Types of beneficiary	One or more of the types of beneficiary foreseen in the call	factor
SAN_dum	Sanctions	If the award was sanctioned	factor
Month_CSU210	Month of award	Month of the date when the grant was awarded	factor
Nawards_TER_110	Number of awards	Number of awards received by the same third party	numeric
Amount_awards_TER110	Amount of awards	Overall amount of awards received by the same third party	numeric
NATIONAL_CSU260 REGIONAL_CSU260 MUNICIAPAL_CSU260	Level of award	If the grant was awarded by national, regional or municipal body	factor
NATIONAL_TER310 REGIONAL_TER310 MUNICIAPAL_TER310	Level of third party location	If the third party is located at national, regional, municipal level	factor

LOCAL_IMPL	Local implementation	If the location of third party is the same as location of granting body	factor
SECTOR_CON550_AGR.. .EXTRATER	Sector of economy	Sectors of the economy foreseen in the call	factor

Glossary

Algorithm	Algorithms are exact sequential sets of commands that are performed over a designed input to generate an output in a clearly defined format. Algorithms can be represented in plain language, diagrams, computer codes and other languages.
Beneficiary/Grant recipient/Grantee	Any individual or organisation that receives grants to support their operations (also referred to as recipients, beneficiaries, or grantees)
Conflict of interest	A conflict of interest involves a conflict between the public duty and private interests of a public official, in which the public official has private-capacity interests which could improperly influence the performance of their official duties and responsibilities.
Control	Any action taken by management, the board, and other parties to manage risk and increase the likelihood that established objectives and goals will be achieved.[1]
Corruption	The active or passive misuse of the powers of Public officials (appointed or elected) for private financial or other benefits
Data analytics	A process of inspecting, cleaning, transforming, and modelling data with the goal of highlighting useful information, suggesting conclusions, and supporting decision-making.
Data architecture	Data architecture is composed of models, policies, rules or standards that govern which data is collected, and how it is stored, arranged, integrated, and put to use in data systems and in organisations.
Data cleaning	A set of procedures designed to identify and correct, when possible, any data errors, inconsistencies and unclear data features.
Data dictionary	A data catalogue that describes the contents of a database. Information is listed about each field in the attribute tables and about the format, definitions and structures of the attribute tables. A data dictionary is an essential component of metadata information.
Data Governance	Data Governance is a system of decision rights and accountabilities for information-related processes, executed according to agreed-upon models which describe who can take what actions with what information, and when, under what circumstances, using what methods.
Double funding	A scenario when identical activities and costs are funded twice through the use of public funds.
Ex-ante control	A control that that aims to reduce the possibility of an undesirable outcome.
Ex-post control	A control meant to identify errors after an event.
Fraud	Fraud is economic crime involving deceit, trickery or false pretences, by which someone gains unlawfully. An actual fraud is motivated by the desire to cause harm by deceiving someone else, while a constructive fraud is a profit made from a relation of trust.
Grant	Grants are transfers made in cash, goods or services for which no repayment is required.
Machine learning	A subset of artificial intelligence in which machines leverage statistical approaches to learn from historical data and make predictions in new situations.
Misappropriation	Acts involving the theft or misuse of an organisation's assets.
Network analysis	A set of integrated techniques to identify relations among actors and to analyse the social structures or patterns that emerge from the recurrence of these relations.
Positive Unlabelled/PU bagging	Positive unlabelled (PU) learning is a semi-supervised machine learning technique, which allows working with highly unbalanced data. PU learning could be used in cases when the majority of all available observations belongs to unlabelled cases

Random Forests	Random forest is a commonly-used machine learning algorithm which combines the output of multiple decision trees to reach a single result. It handles both classification and regression problems.
SHAP values	SHAP (SHapley Additive exPlanations) values express the average marginal contributions of all predictors to the predicted outcome.
Supervised (machine learning)	Supervised learning is a subcategory of machine learning and artificial intelligence. It is defined by its use of labelled datasets to train algorithms to classify data or predict outcomes accurately. As input data are fed into the model, it adjusts its weights until the model has been fitted appropriately.
Test dataset	A randomly selected sample of the dataset which is used to evaluate the quality (e.g. prediction accuracy) of the model estimated on the training dataset.
Training dataset	A randomly selected sample of the dataset which is used to estimate ('train') the machine learning model. The training and test datasets are mutually exclusive, that is each observations belongs to either the training or test datasets.

Note:
[1] Institute of Internal Auditors. (n.d.). *Governance, Risk & Control.* Retrieved from https://na.theiia.org/standards-guidance/topics/pages/governance-risk-and-control.aspx (accessed on 2 November 2021).

OECD Public Governance Reviews

Countering Public Grant Fraud in Spain

MACHINE LEARNING FOR ASSESSING RISKS AND TARGETING CONTROL ACTIVITIES

In the wake of the COVID-19 pandemic, governments face both old and new fraud risks, some at unprecedented levels, linked to spending on relief and recovery. Public grant programmes are a high-risk area, where any fraud ultimately diverts taxpayers' money away from essential support for individuals and businesses. This report identifies how Spain's General Comptroller of the State Administration (*Intervención General de la Administración del Estado,* IGAE) could better identify and control for grant fraud risks. It demonstrates how innovative machine learning techniques can support the IGAE in enhancing its assessment of fraud risks in grant data. It presents a working risk model, developed with datasets at the IGAE's disposal, and maps datasets it could use in the future. The report also considers the preconditions for advanced analytics and risk assessments, including ways for the IGAE to improve its data governance and data management.

Co-funded by the European Union

PRINT ISBN 978-92-64-77835-1
PDF ISBN 978-92-64-55436-8

9 789264 778351